IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 August 2022, accepted 23 August 2022, date of publication 31 August 2022, date of current version 14 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3203174

RESEARCH ARTICLE

Exploring Human Pose Estimation and the Usage of Synthetic Data for Elderly Fall Detection in Real-World Surveillance

SARDOR JURAEV[1], (Student Member, IEEE), AKASH GHIMIRE[2], JUMABEK ALIKHANOV[1], (Student Member, IEEE), VIJAY KAKANI[2], (Member, IEEE), AND HAKIL KIM[1], (Member, IEEE)

[1]Department of Electrical and Computer Engineering, Inha University, Incheon 402751, South Korea
[2]Department of Integrated System Engineering, School of Global Convergence Studies, Inha University, Incheon 402751, South Korea

Corresponding author: Hakil Kim (hikim@inha.ac.kr)

This research was supported by the BK21 Four Program funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea (NRF).

ABSTRACT The world's elderly population continues to grow at an unprecedented rate, creating a need to monitor the safety of an aging population. One of the current problems is accurately classifying elderly physical activities, especially falling down, and delivering prompt assistance to someone in need. Owing to the advancements in deep learning research, vision based solutions are employed for action recognition. One such popular approach is human pose estimation based action recognition or fall detection. Nevertheless, due to a lack of large-scale elderly fall datasets and the continuation of numerous challenges such as varying camera angles, illumination, and occlusion accurately classifying falls has been a problematic. To address these problems, this research first carried out a comprehensive study of the AI Hub dataset collected from real lives of elderly people in order to benchmark the performance of state-of-the-art human pose estimation methods. Secondly, owing to the limited number of real datasets, augmentation with synthetic data was applied and performance improvement was validated based on changes in the degree of accuracy. Third, this study shows that a Transformer network applied to elderly action recognition outperforms LSTM-based networks by a noticeable margin. Lastly, by observing the quantitative and qualitative performances of different networks, this paper proposes an efficient solution for elderly activity recognition and fall detection in the context of surveillance cameras.

INDEX TERMS Elderly care, fall detection, pose estimation, synthetic data, video surveillance.

I. INTRODUCTION

In today's world, one of the increasing challenges is caring for the elderly. By 2050 there will be 1.5 billion people 65 years or older, accounting for 16% of the world population [1]. As a result, monitoring the physical activities of older people, especially for fall detection and prevention, is critical to providing better elderly care and a longer life expectancy. Fortunately, with the advent of modern technologies assisted living has become more accessible to monitor the behavior of an aging population. In the last decade, researchers detected falls by wearable devices, environmental sensors, and cameras [2]. Sensors and wearable devices may offer a quick response time and better fall-detection accuracy. Nevertheless, these methods fail when a person forgets to wear the device or falls in an untracked area [3]. Thus, a more efficient approach is to take advantage of the increasing number of surveillance cameras which are easy to set up and that receive consistent data by tracking an entire area. However, the rapid increase of surveillance cameras raises serious concerns about privacy, and discriminatory bias against specific groups of people [4]. Among many methods, an effective one to avoid the

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S. Raval.

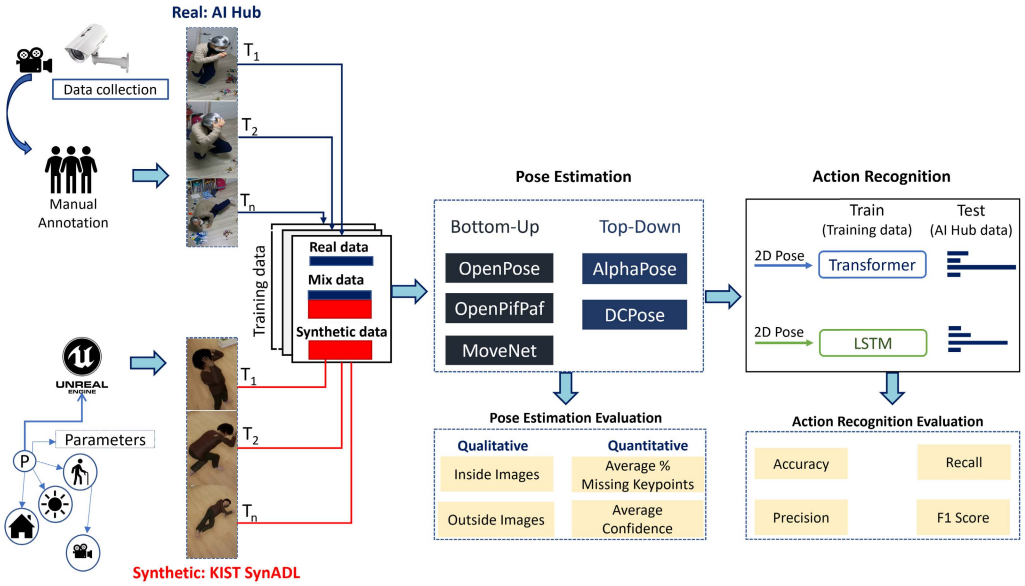


FIGURE 1. Overview of the Performance Evaluation Methodology. Using the AI Hub dataset, performance of bottom-up and top-down human pose estimation approaches underwent pose estimation and action recognition evaluations. Next, after mixing AI Hub and KIST SynADL datasets performance improvement was measured in terms of model accuracy. Here, *T* represents the number of frames.

previously mentioned problems is skeleton-based action recognition, which offers a high degree of anonymity [5]. Although human pose estimation achieves desirable results for scientific datasets, when encountering real-life challenges, such as occlusion, illumination, and steep camera angles the recognition accuracy may degrade leading to failure of action classifiers. However, the lack of real-world elderly activity datasets has limited current fall detection research focusing mainly on scientific datasets: [6], [7], [8], [9] and [10]. These datasets are collected from laboratory surroundings that differ from daily living environments. Furthermore, most of the fall detection datasets involve young people whose speed and actions differ from the elderly [11]. As a result, when an action recognition model is trained on such datasets and tested in real-world settings, these differences may result in unreliable action recognition [12]. Consequently, it is assumed that these methods do not exhibit generalization capabilities for elderly physical action recognition in real-world environments. To address these problems a qualitative and quantitative study was carried out on the AI Hub dataset [13], which covers 12 types of falls and daily life activities.

The major contributions of this paper are threefold: (i) In this research, the performance of state-of-the-art pose estimation methods is evaluated qualitatively and quantitatively by using the AI Hub dataset. (ii) This paper evaluates the degree of performance improvement from synthetic data when added to the limited amount of real-life data. (iii) After extensive experiments, accurate and efficient pose estimation and action recognition models are proposed for real-world fall detection and elderly physical activity recognition.

This paper is organized in the following order. Section II outlines a literature review of public fall detection datasets, pose-based human action recognition and fall detection in addition to synthetic data usage. Section III introduces the proposed methodology. Section IV presents implementation details, describing the qualitative and quantitative results. Section V discusses the findings of the research. Finally, section VI draws conclusions from this research. Code and video explanations are publicly available in https://sard0r.github.io/.

II. RELATED WORKS

A. VISION-BASED FALL DETECTION DATASETS

Because deep learning models require a large amount of data, it is crucial to select training and testing datasets to achieve high performance in real-world applications. For this reason, this research compared several vision-based public datasets and a dataset that fulfills the real-world requirements was chosen. Table 1 lists the well-known and public fall detection datasets comparing the year published, the types of fall, the number of subjects, daily activities included, occlusion, places recorded, participant's ages, dataset size, and number of locations.

1) Le2i

In 2013, being one of the first, Charfi *et al.* [6] presented the Le2i fall detection dataset, which was recorded by a single surveillance camera. In total, the dataset consists of 191 videos in which 143 contain falls and the other 48 show daily activities. To collect the dataset, nine subjects were involved in performing three types of fall-down actions and six different activities of daily life. Videos were captured

TABLE 1. Vision-based publicly available fall detection datasets.

Dataset	Year	Subjects	Number of fall types	Daily activities	Occlusion	Places	Participants' ages	Dataset size	Number of locations
Le2i FDD [6]	2013	9	3	5	Yes	Indoor	-	191 videos	4
URFD [7]	2014	1	2	4	No	Indoor	-	70 videos	1
HQFSD [8]	2016	10	24	13	Yes	Indoor	-	-	1
FDD [9]	2017	5	-	5	No	Indoor	19-50	22636 images	1
UP-Fall [10]	2019	2	5	6	No	Indoor	18-24	-	1
AI Hub [13]	2020	-	12	12	Yes	Indoor and Outdoor	60-80	2500 videos	10

in four different locations with a resolution of 320 x 240 at 25 fps.

2) UNIVERSITY OF RZESZOW

Later in 2014, Kwolek and Kepski [7] introduced the University of Rzeszow fall detection (URFD) dataset including both RGB and depth data types. It was recorded in a laboratory environment by a single subject using two Kinetic cameras. Overall, it contains 70 videos in which 2373 frames are falls, 7452 are non-falls, and the other 1719 are transition frames.

3) HIGH-QUALITY FALL SIMULATION DATASET

In order to bridge the gap between simulated datasets and realistic falls, Baldewijns *et al.* [8] presented the High-quality Fall Simulation Dataset (HQFSD). This dataset was captured using five web cameras using 640 x 480 resolution recorded at 12 fps in a nursing home room. Ten subjects were involved in 55 fall scenarios. Each of the five cameras recorded 2:25:54 hrs of fall data.D

4) FALL DETECTION DATASET

In 2017, Adhikari *et al.* [9] presented the Fall Detection Dataset (FDD) [12], which was captured using an uncalibrated Kinect sensor. Five subjects performed falling-down actions. In total, the dataset consists of 22,636 images recorded in five different rooms from eight viewing angles. Out of five participants, two of them are male (32 and 50 years of age) and three are female (19, 28, and 40).

5) UP-FALL

In 2019, the UP-Fall dataset [10] was presented. The dataset was captured using three modalities: wearable sensors, ambient sensors, and vision sensors. Seventeen subjects between ages 18 and 24 staged five falls and six daily life activities.

6) AI HUB DATASET

In 2020, the Korean government collected the AI Hub dataset for assisted living scenarios. The dataset was captured in both indoor and outdoor environments where elderly people aged 60 years and over performed 12 different actions. The dataset contains 2500 untrimmed videos recorded using CCTV at 3840 x 2160 resolution consisting of the following actions: falling down, leaning, sitting, bending, lying down, standing up, walking, crawling, picking something up, turning, drinking, and eating. Actions were recorded at home, in a hospital, at a community center, in a parking lot, a park, a market, a residential alley, a subway station, on a footbridge, and in front of apartment complexes.

This research was carried out using the AI Hub dataset for a number of reasons. First, because it was collected in 10 different environments there is great variance in distances and viewing angles from camera to subject. Second, unlike laboratory environments, real-world surveillance cameras stream continuously, day and night. Thus, as the videos were recorded both at night and during the daytime, it represents different levels of complexity. Third, it was collected in places where the elderly usually happen to be in need (for example, hospitals and community centers), enabling us to evaluate the performance of human pose estimation methods in real-world assisted living scenarios.

B. POSE-BASED HUMAN ACTION RECOGNITION AND FALL DETECTION

In recent years, pose-based human action recognition has attracted a lot of attention owing to big improvements in human pose estimation. Table 2 outlines related work. There have been many successful attempts to exploit human pose estimation for elderly action recognition. Yang *et al.* [14] proposed a pose refinement system in a combination of AlphaPose [15], OpenPose [16], and LCRNet++ [17] to extract accurate pose sequences from the Toyota Smarthome dataset [18], including cases of occlusion, truncation, and low resolution, and reporting a 4.4% accuracy increase compared to other multimodal methods. Later, Yang and colleagues [19] introduced a new framework called UNIK for real-world skeleton-based action recognition. Their proposed approach used the Toyota Smarthome dataset, achieving accuracies of 64.3% in cross-subject evaluations and 65% in cross-view evaluations. Moreover, Jang *et al.* [12] presented an RGBD dataset of daily activities by elderly people for care robots and proposed a four-stream adaptive CNN architecture for action recognition. Their work reported 90.10% accuracy by training the model on features obtained using OpenPose.

In terms of pose-based fall detection, Hasan *et al.* [20] built a robust human fall detection system using Long short-term memory (LSTM) and GRU networks based on the OpenPose pose estimator and reported 99% sensitivity using the Le2i, and URFD datasets. Lin *et al.* [21] approached the problem of fall detection with a similar method, introducing a fall detection framework utilizing OpenPose as a feature extractor and LSTM and GRU for classification, achieving 98.2% accuracy using the URFD dataset.

TABLE 2. Related works for pose estimation-based human action recognition and fall detection.

Reference	Pose Estimation	Dataset	Model	Result
Yang et al. [14]	AlphaPose, OpenPose, LCRNet++	Toyota Smarthome	Pose-Refinement System (SSTA-PRS)	Cross Subject 62.1%, Cross View 54%
Yang et al. [19]	LCRNet++	Toyota Smarthome	UNIK	Cross Subject 64.3%, Cross View 65%
Jang et al. [12]	OpenPose	ETRI-Activity3D	FSA-CNN	90.10%
Hasan et al. [20]	OpenPose	Le2i FDD and URFD	LSTM, GRU	99% sensitivity
Lin et al. [21]	OpenPose	URFD	LSTM, GRU	98.2% accuracy
Taufeeque et al. [22]	OpenPifPaf	UP-Fall dataset	LSTM	92.5% F1-score
Ramirez et al. [24]	AlphaPose	UP-Fall dataset	Random forest, SVM, KNN, MLP	99.34% accuracy
Yadav et al. [25]	OpenPose	UP-Fall dataset	ARFDNET	96.7% accuracy
Serpa et al. [26]	AlphaPose, OpenPose, PoseNet	URFD	MLP	94.5% sensitivity, 99.9% specifity

Taufeeque *et al.* [22] introduced real-time, multi-camera, multi-person fall detection by using skeleton features from OpenPifPaf [23] and LSTM for classification, reporting a 92.5% F1-score with the UP-Fall dataset. Furthermore, Ramirez *et al.* [24] exploited the AlphaPose pose estimation method and performed fall detection with the UP-Fall dataset, improving the average accuracy to 99.34% with a random forest classifier. More recently, Yadav *et al.* [25]presented an efficient activity recognition and fall detection system utilizing OpenPose and a combination of Conv1D and GRU networks. This system was trained on the UP-Fall dataset showing 96.7% accuracy. Similar to our research, Serpa *et al.* [26] evaluated human pose estimation methods such as AlphaPose, OpenPose, and PoseNet [27] as a solution to the fall detection problem. Their proposed work compared these methods in terms of average precision, frame accuracy, and keypoint accuracy, as well as action classification accuracy. To conduct a comparative study, they chose the URFD dataset and an MLP model for classification. The findings of the research were that AlphaPose outperformed the other two pose estimation networks at 94.5% for sensitivity and 99.9% for specificity. The shortcomings of this work are threefold. One is that the evaluation was performed on limited scientific data consisting of only 70 videos. Another shortcoming highlighted by the authors is that low light and occlusions were not considered but left for future work. Most importantly, performance evaluation was based solely on a quantitative comparison, excluding qualitative analysis.

C. SYNTHETIC DATA USAGE

While learning behaviors from synthetic data is an underresearched area, recently there has been a great deal of interest in usage of synthetic data for human action recognition and fall detection. Because deep learning models are extremely data hungry, to achieve better generalization a number of researchers have resorted to exploiting synthetic data to provide networks with an abundant amount. Wang *et al.* [11] generated a large-scale synthetic dataset covering 55 elderly activities using Unreal Engine platform. The study covered the exploitation of synthetic data as an augmentation for real data reporting up to 2.41% accuracy improvement in recognition of elderly physical activities. Similarly, Zherdev *et al.* [28] generated a synthetic dataset to overcome the data scarcity problem for elderly fall detection. The aim of the study was to evaluate the effectiveness of using only synthetic data in real fall scenarios. It reported a 97.6% fall detection accuracy when the network was trained with solely synthetic data and tested on URFD real dataset.

Throughout the literature, the generative adversarial network and computer graphics techniques have been popular for obtaining synthetic datasets. Khodabandeh *et al.* [29] and Wang *et al.* [30] generated synthetic data using a GAN, while other studies utilized computer graphics and game platforms to simulate human actions [11], [28], [31], [32]. To evaluate performance improvement from using synthetic data, KIST SynADL [11] was utilized for a number of reasons. Firstly, the body shapes and motions of the characters were captured from real actions by elderly people. Secondly, it was designed especially for augmenting realistic elderly datasets for action recognition by smart surveillance and care robots. Finally, it is a large synthetic dataset including 15 characters who performed 55 actions of elderly people in four different environments. Although the article by Wang *et al.* [11] is closely related to our research, it only considered synthetic data usage targeting recognition of daily activities from a care robot's view. Therefore, it utilized realistic datasets [33] and [12] which were captured in laboratory settings from a side view for application by care robots. By contrast, our study focuses on the usage of synthetic data for real-world surveillance applications which is explored for the first time.

III. METHODOLOGY

In this section, a performance evaluation methodology for human pose estimation methods is introduced as a solution to elderly action recognition and fall detection. The overall process of the proposed methodology is illustrated in Fig. 1. The main aim of this research is to explore the performance of human pose estimation methods for elderly action recognition and fall detection in real-world surveillance, and to illustrate the degree of improvement through synthetic data exploitation. To evaluate the potential of pose estimation for fall detection problems, five human pose estimation methods are considered with two prominent action recognition networks. In general, the raw video frames from the chosen real dataset are first fed into the selected pose estimation methods, which output sets of human poses. Next, qualitative and quantitative evaluations were performed for each pose estimation method utilizing the extracted human pose

TABLE 3. Comparision of the pose estimation methods. FPS values were computed a using Nvidia GTX 1080Ti.

	Model	Dataset	#KPs	mAP	FPS
DCPose	HRNet-W48	PoseTrack2017	17	79.2	8.09
AlphaPose	ResNet152	COCO	17	73.3	13.35
MoveNet	Thunder	COCO	17	72	87
OpenPifPaf	Shufflenetv2k30	COCO	17	71.8	15.67
OpenPose	Coco	COCO	18	61.8	13.5

features. After qualitative and quantitative analysis of the selected pose estimation models, the influence of synthetic data was observed when mixed with real data.

A. HUMAN POSE ESTIMATION AND DATA PREPROCESSING

Human pose estimation methods can be categorized into top-down and bottom-up approaches. The top-down approaches employ a human detector to find human candidates before doing single-person pose estimation. In contrast, the bottom-up approaches first predict the keypoints of a person, and later associate the detected keypoints to form full poses. There are advantages and disadvantages to both proposed approaches [34]. For example, bottom-up approaches are efficient in inference speed while predicting the keypoints irrelative to human candidates in a scene. Nevertheless, they may produce disconnected or error-associated parts in truncation or occlusion scenarios.

On the other hand, top-down approaches are dependent on person detector accuracy in addition to having inference speed in accordance with the person count in the image. However, the keypoint prediction performance of top-down approaches is highly accurate compared to the bottom-up approaches. Thus, to find the most accurate and efficient pipeline, two top-down and three bottom-up state-of-the-art human pose estimation methods were selected. From the top-down approaches, AlphaPose and DCPose [35] using YOLOv3 [36] as a detector were chosen for evaluation. From the bottom-up approaches, OpenPose, OpenPiPaf, and MoveNet [37] were selected. Table 3 shows the pose estimation methods with the chosen trained datasets and their respective mAPs and frames per second. All of the methods were trained on the COCO [38] dataset except DCPose, which was trained on the PoseTrack2017 [39] dataset. For a fair comparison, the most accurate models in terms of detection quality were chosen for every pose estimation method. First, 25 image sequences at 960 x 540 were fed into the selected human pose estimation models, which extracted 2D poses in a T x K shape, where K is the total number of keypoints and T is the number of frames. After applying human pose estimation, 18 and 17 keypoints were obtained from OpenPose and the other models, respectively. Each keypoint consists of three values: X and Y (representing human joint coordinates) in addition to C, measuring how correctly the keypoint was estimated. The keypoints obtained from pose estimation models were further normalized from 0 to 1 and fed into the action recognition models. For normalization, if X represents the extracted keypoint features:

$$X_{normalized} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

B. ACTION RECOGNITION MODELS

1) TRANSFORMER MODEL

Mazzia *et al.* [40] proposed the Action Transformer architecture for classifying human action recognition in short time steps. The proposed method outperformed many RNN-based networks achieving real-time performance on a CPU. Inspired by their success, the authors applied the Transformer-based architecture to elderly action recognition and fall detection for the first time. Fig. 2 illustrates the classification procedures by Transformer of elderly actions. In this proposed method, pose estimation is initially applied to the input video, after which estimated keypoints are preprocessed and projected linearly to the predefined transformer encoder. Following that, keypoint features from pose estimation models are added to class tokens and positional encoding, forming input embeddings by the Transformer encoder. Then, within the Transformer encoder, the input embeddings are projected onto the multi-head attention layer followed by fully connected layers. The output is then passed to the MLP head, which predicts the action classes.

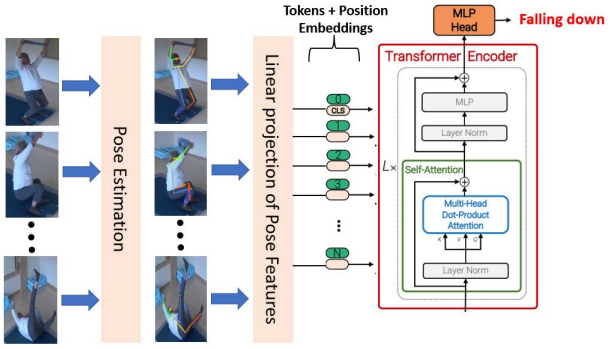


FIGURE 2. Architecture of the Transformer Model.

2) LSTM MODEL

Long short-term memory is an evolution of the classic Recurrent Neural Network. Unlike traditional RNNs, LSTM is designed to maintain the information of longer data sequences and to learn variations in time. In the literature, LSTM is popular in the application of action classification achieving state-of-the-art accuracy [20], [21], [22]. Thus, in this research for quantitative comparisons, LSTM was chosen for observation alongside Transformer. An overview of action recognition using LSTM can be seen in Fig. 3. For action classification, the extracted keypoint features of input videos from pose estimation models are passed as input to pre-defined LSTM layers. The output features of the last LSTM layer are then passed through dense layers followed by a softmax layer to predict the probabilities of each class.

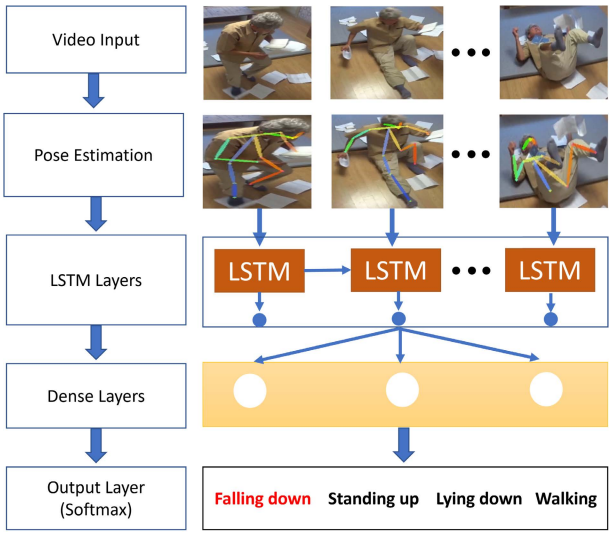


FIGURE 3. Architecture of the LSTM Model.

C. EVALUATION METRICS

1) POSE ESTIMATION METRICS

The confidence score of each detected keypoint represents how likely the pose estimation models have successfully detected human joints. Extracted keypoint features from an input video having higher average confidence rate produce preferable results when used to recognize human actions. To calculate the average confidence of keypoints in a video (ACV), first average confidence of keypoints in an image (IC), is calculated as seen below:

$$IC = \frac{\sum_{i=0}^{K-1} C_i}{K} \tag{2}$$

$$ACV = \frac{\sum_{i=0}^{T-1} IC_i}{T} \tag{3}$$

In equation (2), C is the confidence of each keypoint, K denotes the total number of keypoints, and T stands for the number of frames.

Owing to many factors, such as occlusion, dim light, and low-resolution pose estimation models fail to detect keypoints of the human body. The ability of pose estimation models to detect keypoints under the above-mentioned conditions directly impacts the performance of action recognition models. Because the number of keypoints for each pose estimation is not equal, in this paper for the evaluation of pose estimation models the average percentage of missing keypoints is calculated. To estimate the average percentage of missing keypoints (AMV) from a video, first the percentage of missing keypoints (IM) in an image is calculated as seen below:

$$IM = \frac{n}{K} \times 100\% \tag{4}$$

$$AMV = \frac{\sum_{i=0}^{T-1} IM_i}{T} \tag{5}$$

In equation (4), n represents the total number of missing keypoints from an image.

2) ACTION RECOGNITION METRICS

To evaluate the performance of action recognition models, four metrics of a confusion matrix [41], (accuracy, recall, precision, and F1-score) were chosen. Accuracy can be defined as the ratio of the total number of correct predictions to the total number of predictions. For a binary class classification, recall, precision, and F1-score can be defined as follows. Precision is a measure of correct predictions out of all the positive predictions. Recall (or Sensitivity) is a measure observing the accurately classified cases out of all positive cases. F1-score combines precision and recall into a single metric by taking their harmonic mean. However, because this study classifies four action classes, the evaluation method differs from the binary classification problem. For this study, precision, recall, and F1-score for each action class were calculated separately, after which the weighted average of each evaluation metric from all action classes was calculated.

IV. EXPERIMENTS

A. IMPLEMENTATION DETAILS

1) DATASETS

To conduct the experiments four action classes were chosen: falling down, standing up, lying down, and walking. Those four classes exist in both the AI Hub and the Kist SynADL datasets. KIST SynADL dataset was captured only from indoor environments. Therefore, only indoor video samples from the AI Hub dataset were considered for quantitative analysis. Fig. 4 depicts the distribution of real and synthetic videos exploited in our experiments. Sample instances from both datasets are visualized in Fig. 5. In total, 1296 videos were chosen from the AI Hub dataset. Classes such as falling down, standing up, lying down, and walking included 166, 280, 350, and 500 samples, respectively. For each action class, 1600 new synthetic video samples were added to the training dataset from KIST SynADL. The amount of available indoor data in realistic scenarios is extremely limited. Hence, one of the aims of this study was to evaluate the

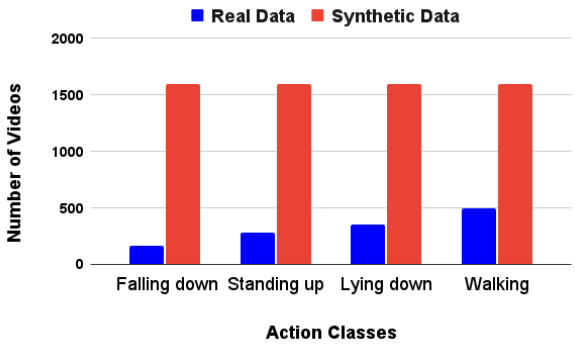


FIGURE 4. The number of videos utilized from the AI Hub and KIST SynADL datasets.

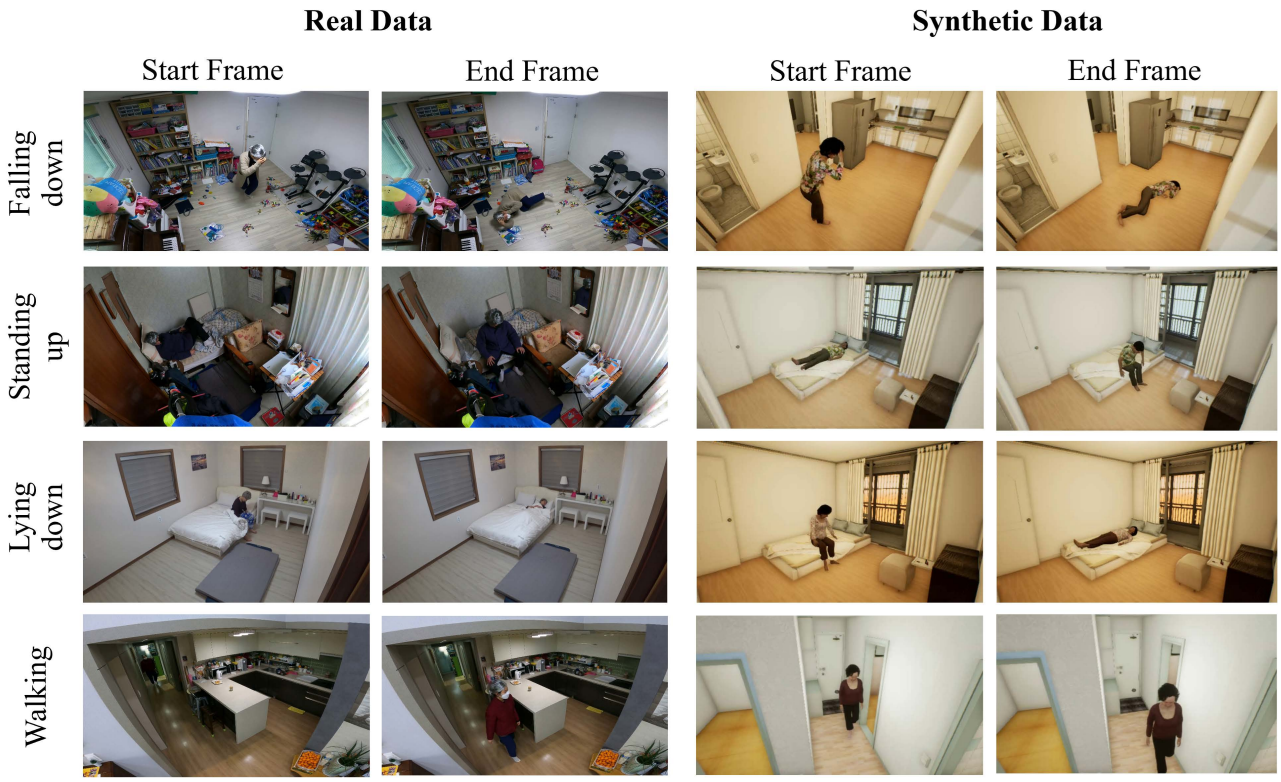


FIGURE 5. Examples of each Action classes from AI Hub and KIST SynADL datasets.

performance gains obtained from exploiting synthetic data to combat data limitations. In other words, synthetic data were mixed with real video samples for training the deep learning models. For fair evaluation and reproducibility, the AI Hub dataset was divided into training and testing sets at a 3:1 ratio. Multiple experiments were performed where synthetic data used in the training set was increased by 1600, with each class given 400 new synthetic samples. The testing accuracy of action recognition models such as Transformer and LSTM were calculated on the real test dataset.

2) TRAINING SETTINGS

Table 4 summarizes the hyperparameters obtained from progressively training the two action recognition models. In order to obtain optimal hyperparameters for each network, the Optuna [42] framework was utilized with Hyperband [43] algorithms. Hyperparameters were tuned for the training set, excluding the test set. Thus optimal parameters were selected for training the Transformer and LSTM models. For a fair evaluation, both models were trained for 200 epochs on a personal computer with 16 GB RAM, an Intel I9-11900K CPU, and the Nvidia GTX 1080Ti GPU. For the LSTM model, three LSTM layers were designed with hidden units of 32, 64, and 32, after which there were four dense layers with dense units of 128, 64, 32, and 16. The Transformer model was composed of four Transformer encoder layers each consisting of one multi-head self-attention layer. In both

TABLE 4. Training details of Transformer and LSTM.

Transformer		LSTM	
Training		**Training**	
Epochs	200	Epochs	200
Batch size	128	Batch size	128
Transformer encoder layers	4	LSTM layers	3
Multi-Head attention layers	1	Dense layers	4
Optimizer	Adam	Optimizer	Adam
Regularization		**Regularization**	
Learning rate	0.001	Learning rate	0.001
Dropout	0.3	Dropout	0.3

of the action recognition models, the Adam [44] optimizer was used with a learning rate of 0.001. Lastly, a dropout rate of 0.3 was applied to both models for regularization.

B. EXPERIMENTAL RESULTS

1) QUALITATIVE ANALYSIS

Human pose estimation is considered a challenging task in real-world applications. These challenges arise under many conditions, such as camera angle, lighting condition, occlusions, and camera-to-subject distance [45]. We evaluated the performance of the above-mentioned pose estimation models for elderly physical activity recognition in real-life surveillance scenarios. In this regard, a qualitative comparison was

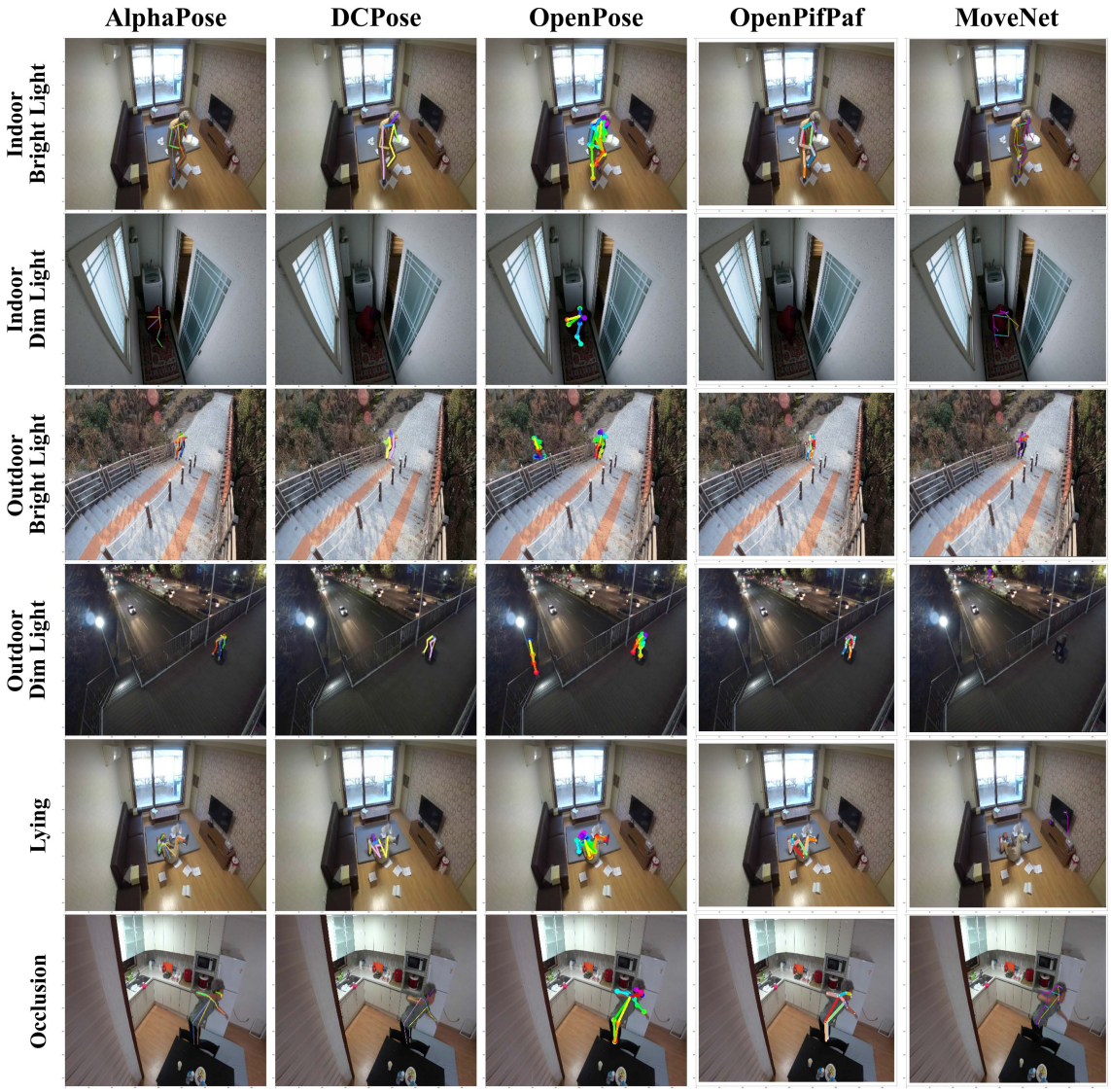


FIGURE 6. Qualitative analysis of state-of-the-art pose estimation models.

conducted on indoor and outdoor image samples from the AI Hub dataset. Fig. 6 shows qualitative analysis where each row depicts different scenario complications, such as indoor bright and dim light, outdoor bright and dim light, occlusions, and lying. Each column illustrates the keypoint detection result of a given pose estimator. Among pose estimation models, AlphaPose achieved the most robust results in all cases. One exception was observed in indoor dim light scenarios where it failed to detect occluded parts of the person. OpenPifPaf and DCPose demonstrated similar results, failing only under indoor dim light. Although OpenPose predicted the keypoints of a person in every scenario, it suffered from false positives in outdoor environments. MoveNet performed the worst, failing to detect keypoints accurately with respect to body joints in most scenarios. We observed that OpenPose and MoveNet detected false positive keypoints when a person was far from the camera and when there were other non-human objects similar to the shape of a person.

2) QUANTITATIVE ANALYSIS

a: PERFORMANCE FROM SELECTED PIPELINES

To evaluate the robustness of the chosen pose estimation methods quantitative experiments were conducted using the AI Hub dataset. First, performance from the pose-estimation-based action recognition pipelines was compared. Results for accuracy, precision, recall, and F1-score in both action recognition pipelines are shown in Table 5. Both Transformer- and LSTM-based pipelines achieved good performance when coupled with AlphaPose. The worst performance was observed when MoveNet was used as a pose feature extractor in the action recognition pipeline. The best pipelines were AlphaPose for pose features and Transformer for action classification. Pipeline results for accuracy, recall,

TABLE 5. Transformer and LSTM pipeline results for each human pose estimation model with the AI Hub dataset.

Pose Estimation	Accuracy	Precision	Recall	F1-score
LSTM pipeline				
AlphaPose	88.29	88	88	87
DCPose	88.13	88	87	87
OpenPifPaf	82.94	82	79	80
OpenPose	81.07	81	81	81
MoveNet	68.92	69	69	69
Transformer pipeline				
AlphaPose	**89.22**	**90**	**89**	**89**
DCPose	88.73	**90**	88	**89**
OpenPifPaf	84.11	84	84	84
OpenPose	83.41	83	83	83
MoveNet	71.49	66	66	66

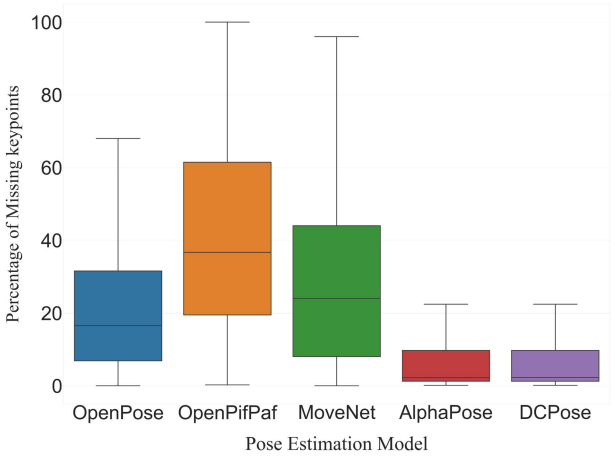


FIGURE 7. Comparisons for average percentage of missing keypoints with the AI Hub dataset.

precision, and F1-score were 89.22%, 90%, 89%, and 89%, respectively. As an action classification model, Transformer showed superior performance compared to LSTM when paired with any pose estimation method. Performance differences ranged up to 2.57%.

b: POSE ESTIMATION EVALUATION

Secondly, for quantitative analysis of pose estimation models, the average percentage of missing keypoints and the average confidence for keypoints in T frames were compared. The reason was that the AI Hub dataset did not provide ground truth keypoints for body joints. Therefore, the mAP of the detected keypoints could not be calculated. To solve this issue, two metrics were proposed in Sec. III: average percentage of missing keypoints and average confidence of keypoints. The average percentage of missing keypoints for all methods is illustrated using boxplots in Fig. 7. Average confidence of keypoints for all methods is explained in Fig. 8. We can see that AlphaPose and DCPose showed similar percentages of missing keypoints whereas the former achieved the best results with mean and median of 9.66% and 2.26%, respectively. OpenPifPaf demonstrated the highest

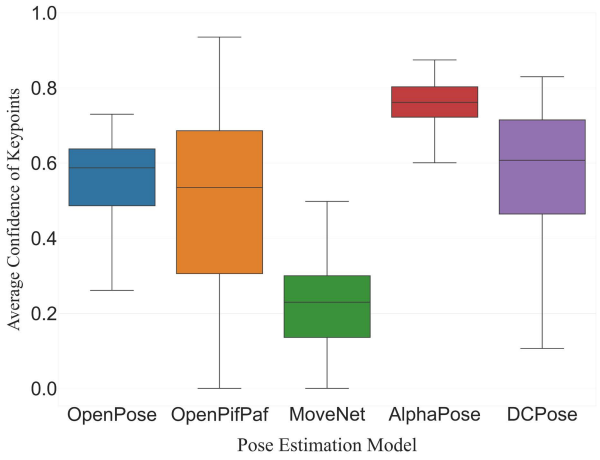


FIGURE 8. Comparisons for average confidence of pose estimation models with the AI Hub dataset.

percentage of missing keypoints with mean and median of 42.12% and 36.51%, respectively. Similarly, from studying the average confidence in keypoints, AlphaPose performed the best with mean and median of 75.29% and 76.15%, respectively, while MoveNet showed the worst result with mean and median of 21.55% and 22.92%, respectively.

c: ACTION RECOGNITION MODEL COMPARISON

Lastly, for fair evaluation of action recognition in the models' performance, the important characteristics of Transformer and LSTM were calculated when trained and tested on the AI Hub data. Table 6 shows the number of parameters, MFLOPs, GPU memory requirements, and inference time of the Transformer and LSTM models. The table shows that both models had comparable inference speeds. It is also worth mentioning that the number of parameters was 3.6 times higher in Transformer.

TABLE 6. Model Parameters, MFLOPS, GPU Memory Requirements, and Inference Time.

Models	# Parameters	MFLOPs	Memory requirement (GPU)	Inference time
Transformer	0.221M	0.0004	232.6KB	0.9671ms
LSTM	0.061M	0.0003	78.6KB	0.9342ms

3) IMPACT OF THE SYNTHETIC DATA

a: SYNTH+REAL TRAINING

This section reports the impact from increasing the training data by using synthetic video samples. First, the performance from action recognition models trained with real-only (Real), synthetic-only (Synthetic), and mixed synthetic and real (Real+Synthetic) data were compared. Corresponding data are shown in the second row of Table 7 for both Transformer and LSTM networks. It is interesting to see the degree of generalization when the model was trained on completely synthetic data and evaluated on real data. The table shows that training the action classification models with

TABLE 7. Training jointly on synthetic and real data.

Transformer					
Training data	**AlphpaPose**	**DCPose**	**OpenPifPaf**	**OpenPose**	**MoveNet**
Real	**89.22**	88.73	84.11	83.41	71.49
Synthetic	**70.19**	68.96	60.63	59.11	53.21
Real + Synthetic	**94.35**	**93.66**	**87.14**	**86.69**	**74.87**
LSTM					
Training data	**AlphpaPose**	**DCPose**	**OpenPifPaf**	**OpenPose**	**MoveNet**
Real	88.29	88.13	82.94	81.07	68.92
Synthetic	67.95	66.27	59.73	58.37	52.74
Real + Synthetic	**93.20**	**92.95**	**85.40**	**84.32**	**72.30**

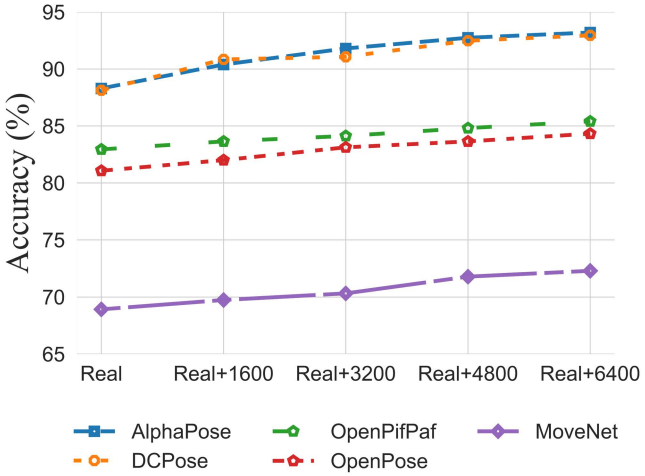


FIGURE 9. Impact of synthetic data on accuracy: LSTM.

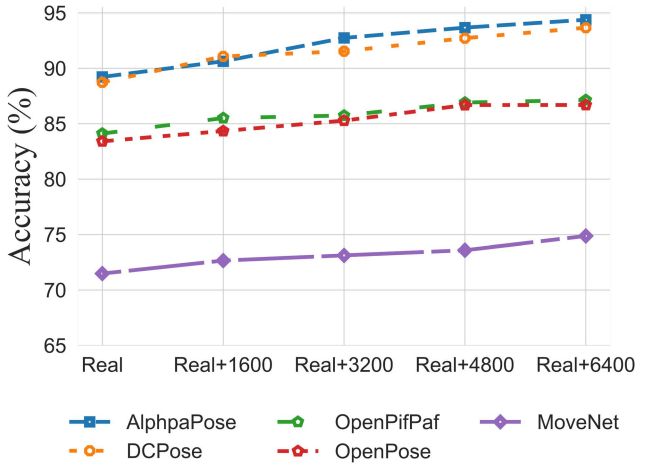


FIGURE 10. Impact of synthetic data on Accuracy: Transformer.

only synthetic data can achieve up to 70.19% accuracy with a Transformer model that uses AlphaPose keypoints. This already demonstrates a promising generalization capability with synthetic data. Up to 89.22% accuracy was obtained using only real data for training. When adding synthetic data to real training data, accuracy increased to 94.35%. Similar behavior was observed for other keypoint extractor methods.

b: IMPACT OF SYNTHETIC SAMPLES PER CLASS

In the above, we saw an increase in performance when adding synthetic data. Next, we examined the trend in performance improvement by adding synthetic data incrementally. Graphs in Fig. 9 and Fig. 10 capture the increased trends when adding more synthetic data to the training set of the action classification model.

For each action classification model (*i.e.,* Transformer and LSTM), five pipelines were analyzed with each pipeline corresponding to different keypoint extractors. Analysis showed that the amount of synthetic data had a direct impact on action recognition performance. To capture the trend, we added 1600 synthetic data in each step (400 samples for each class). With each addition of synthetic video samples to the real data there was a consistent increase in accuracy from both the Transformer and LSTM models. The accuracy improvement was noticeable in the early steps, which tended to decrease marginally with more synthetic data. The highest increase can be attributed to AlphaPose with the Transformer pipeline at a 5.16% improvement, where accuracy went from 89.22% to 94.35%. The lowest increase was obtained from OpenPifPaf and LSTM pipelines (an increase in accuracy of 2.46%). The mean increase in accuracy from Transformer and LSTM action recognition models when coupled with all pose estimation methods was 3.96% and 3.77%, respectively. It is worth noting that the trend in the performance increase generalized similarly with both Transformer- and LSTM-based action classification models.

c: INFLUENCE OF SYNTHETIC DATA ON FALL DETECTION

To visualize the performance improvement after adding synthetic data, a Transformer pipeline was used. Keypoints were extracted from AlphaPose and given as input to the Transformer model. Three scenarios were tested with two pipelines where, in all cases, test data were the same three samples of falling down from the AI Hub dataset. In the first pipeline, the Transformer-based action recognition model was trained only on real (AI Hub) data. Next, 6400 synthetic video samples were added to the real data to create a larger training set. Then, on this larger set, the second Transformer-based pipeline was trained. Performance improvements after adding synthetic data are shown in Table 8 for the three cases shown in Fig. 11 where each row represents a single video clip failure with different snapshots. When the pipeline was trained only using real data, all three cases failed with high probabilities of identifying the wrong category. The pipeline that was trained with the additional synthetic dataset correctly predicted the correct class with high probability. One possible explanation is that the nature of falling down is diverse. There is not enough falling down data in the AI Hub dataset to capture such diversity. However, the synthetic Kist SynADL dataset can be used to cover some missing action variations.

FIGURE 11. Improved failure cases after including synthetic data in training. All three scenarios were failing when model is trained only using real data. Adding synthetic data then retraining the model successfully corrected failure cases.

TABLE 8. Performance Comparison of the Transformer pipeline trained with and without synthetic data.

	Training Data: Real		Training Data: Real + Synthetic	
#Case	Predicted class	Probability	Predicted class	Probability
1	Lying down	0.961	Falling down	0.999
2	Lying down	0.562	Falling down	0.931
3	Walking	0.552	Falling down	0.985

V. DISCUSSION

It is noteworthy that top-down approaches performed better than bottom-up approaches. AlphaPose was the most robust pose estimator in both qualitative and quantitative analyses. In terms of robustness, DCPose was similar to AlphaPose but suffered the highest latency. The pose estimation model with the lowest latency was MoveNet, but it was the least robust in both qualitative and quantitative studies. Overall, for top-down pose estimation, we noted that the metrics defined in Sec III correlate with action classification model robustness. However, this conclusion cannot be drawn from bottom-up approaches. A possible explanation is that bottom-up pose estimation demonstrated false positive results. In other words, false positives where non-human objects appear to be a person to the pose estimator decreased accuracy.

Also, one of the findings was that the addition of synthetic data to real data improved the accuracy of the action recognition models. This observation holds for all the pose estimation models. In order to build highly robust elderly care applications, it is crucial, yet challenging, to obtain large-scale elderly behavior datasets. A better solution can be building synthetic datasets that can be obtained at less cost without manual annotation. The current quality of synthetic datasets does not show high generalization when used alone to build models. However, they can be used to augment a real dataset, boosting performance from action classification models.

Another finding of this work is that the Transformer model showed inference speed comparable to the LSTM model. Thus, in real-world applications using AlphaPose in combination with a Transformer model can be assumed to demonstrate high accuracy with lower latency. This approach can be optimized to target real-time applications, especially in monitoring elderly people where the problem should be solved instantly to prevent sudden accidents.

VI. CONCLUSION

Exploiting state-of-the-art human pose estimation methods for pose-based action recognition and fall detection was the main emphasis of this study. Specifically, this paper explored action classification for elderly-care-monitoring applications that include fall detection. As a pose estimation model, we used five methods: AlphaPose, DCPose, OpenPose, OpenPifPaf, and MoveNet. LSTM and Transformer were explored as potential methods to model action sequences. Perhaps most importantly, this study examined the benefits from using synthetic data for pose-based action recognition and fall detection due to the limited amount of real-world data. AlphaPose was found to be the most accurate human pose estimator in surveillance scenarios. Transformer outperformed LSTM in accuracy, precision, recall, and F1-score. Results show that exploitation of synthetic data improved action recognition performance significantly. A limitation of this research is that quantitative evaluations were performed only on indoor data. As future research, one can extend the present work by considering observations from outdoor elderly human behavior datasets as well.

ACKNOWLEDGMENT

(Sardor Juraev and Akash Ghimire contributed equally to this work.)

REFERENCES

[1] Y. Kamiya, N. M. S. Lai, and K. Schmid. (Jan. 2021). *World Population Ageing 2020*. [Online]. Available: https://bit.ly/3aPJb9A

[2] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, pp. 144–152, Jan. 2013.

[3] M. N. H. Mohd, Y. Nizam, S. Suhaila, and M. M. A. Jamil, "An optimized low computational algorithm for human fall detection from depth images based on support vector machine classification," in *Proc. IEEE Int. Conf. Signal Image Process. Appl. (ICSIPA)*, Sep. 2017, pp. 407–412.

[4] T. Mukhiddin, H. R. Arousha, A. Ubaydullo, L. Wookey, and S. Lee, "Privacy-preserving of human identification in CCTV data using a novel deep learning-based method," in *Proc. IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, Jan. 2022, pp. 211–214.

[5] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 2969–2978.

[6] 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. Imag.*, vol. 22, no. 4, Jul. 2013, Art. no. 041106.

[7] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Comput. Methods Programs Biomed.*, vol. 117, no. 3, pp. 489–501, Dec. 2014.

[8] G. Baldewijns, G. Debard, G. Mertes, B. Vanrumste, and T. Croonenb, "Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms," *Healthcare Technol. Lett.*, vol. 3, no. 1, pp. 6–11, Mar. 2016.

[9] K. Adhikari, H. Bouchachia, and H. Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network," in *Proc. 15th IAPR Int. Conf. Mach. Vis. Appl. (MVA)*, May 2017, pp. 81–84.

[10] L. Martínez-Villaseñor, H. Ponce, J. Brieva, E. Moya-Albor, J. Núñez-Martínez, and C. Peñafort-Asturiano, "UP-fall detection dataset: A multimodal approach," *Sensors*, vol. 19, no. 9, p. 1988, Apr. 2019.

[11] H. Hwang, C. Jang, G. Park, J. Cho, and I.-J. Kim, "ElderSim: A synthetic data generation platform for human action recognition in eldercare applications," 2020, *arXiv:2010.14742*.

[12] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "ETRI-activity3D: A large-scale RGB-D dataset for robots to recognize daily activities of the elderly," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 10990–10997.

[13] (2020). *AI Hub Dataset*. [Online]. Available: https://bit.ly/3ob99qT

[14] D. Yang, R. Dai, Y. Wang, R. Mallick, L. Minciullo, G. Francesca, and F. Bremond, "Selective spatio-temporal aggregation based pose refinement system: Towards understanding human activities in real-world videos," 2020, *arXiv:2011.05358*.

[15] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2334–2343.

[16] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[17] G. Rogez, P. Weinzaepfel, and C. Schmid, "LCR-Net++: Multi-person 2D and 3D pose detection in natural images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1146–1161, May 2020.

[18] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, "Toyota smarthome: Real-world activities of daily living," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 833–842.

[19] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca, and F. Bremond, "UNIK: A unified framework for real-world skeleton-based action recognition," 2021, *arXiv:2107.08580*.

[20] M. M. Hasan, M. S. Islam, and S. Abdullah, "Robust pose-based human fall detection using recurrent neural network," in *Proc. IEEE Int. Conf. Robot., Autom., Artif.-Intell. Internet Things (RAAICON)*, Nov. 2019, pp. 48–51.

[21] C.-B. Lin, Z. Dong, W.-K. Kuan, and Y.-F. Huang, "A framework for fall detection based on OpenPose skeleton and LSTM/GRU models," *Appl. Sci.*, vol. 11, no. 1, p. 329, Dec. 2020.

[22] M. Taufeeque, S. Koita, N. Spicher, and T. M. Deserno, "Multi-camera, multi-person, and real-time fall detection using long short term memory," *Proc. SPIE*, vol. 11601, pp. 35–42, Feb. 2021.

[23] S. Kreiss, L. Bertoni, and A. Alahi, "OpenPifPaf: Composite fields for semantic keypoint detection and spatio-temporal association," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13498–13511, Aug. 2022.

[24] H. Ramirez, S. A. Velastin, I. Meza, E. Fabregas, D. Makris, and G. Farias, "Fall detection and activity recognition using human skeleton features," *IEEE Access*, vol. 9, pp. 33532–33542, 2021.

[25] S. K. Yadav, A. Luthra, K. Tiwari, H. M. Pandey, and S. A. Akbar, "ARFDNet: An efficient activity recognition & fall detection system using latent feature pooling," *Knowl.-Based Syst.*, vol. 239, Mar. 2022, Art. no. 107948.

[26] Y. R. Serpa, M. B. Nogueira, P. P. M. Neto, and M. A. F. Rodrigues, "Evaluating pose estimation as a solution to the fall detection problem," in *Proc. IEEE 8th Int. Conf. Serious Games Appl. Health (SeGAH)*, Aug. 2020, pp. 1–7.

[27] Google. (2019). *Tensorflowjs Posenet*. [Online]. Available: https://bit.ly/2seAmfa

[28] D. Zherdev, L. Zherdeva, S. Agapov, A. Sapozhnikov, A. Nikonorov, and S. Chaplygin, "Producing synthetic dataset for human fall detection in AR/VR environments," *Appl. Sci.*, vol. 11, no. 24, p. 11938, Dec. 2021.

[29] M. Khodabandeh, H. R. V. Joze, I. Zharkov, and V. Pradeep, "DIY human action dataset generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 1448–1458.

[30] L. Wang, Z. Ding, Z. Tao, Y. Liu, and Y. Fu, "Generative multi-view human action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6212–6221.

[31] A. Roitberg, D. Schneider, A. Djamal, C. Seibold, S. Reiß, and R. Stiefelhagen, "Let's play for action: Recognizing activities of daily living by learning from life simulation video games," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep./Oct. 2021, pp. 8563–8569.

[32] D. Ludl, T. Gulde, and C. Curio, "Enhancing data-driven algorithms for human pose estimation and action recognition through simulation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3990–3999, Sep. 2020.

[33] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+ D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[34] M. Li, Z. Zhou, J. Li, and X. Liu, "Bottom-up pose estimation of multiple person with bounding box constraint," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 115–120.

[35] Z. Liu, H. Chen, R. Feng, S. Wu, S. Ji, B. Yang, and X. Wang, "Deep dual consecutive network for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 525–534.

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

[37] Google. (2021). *Tensorflowjs MoveNet*. [Online]. Available: https: https://bit.ly/3Pho1jG

[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Zürich, Switzerland: Springer, 2014, pp. 740–755.

[39] U. Iqbal, A. Milan, and J. Gall, "PoseTrack: Joint multi-person pose estimation and tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2011–2020.

[40] V. Mazzia, S. Angarano, F. Salvetti, F. Angelini, and M. Chiaberge, "Action transformer: A self-attention model for short-time pose-based human action recognition," *Pattern Recognit.*, vol. 124, Apr. 2022, Art. no. 108487.

[41] *Sklearn Confusion Matrix*. Accessed: Jul. 31, 2022. [Online]. Available: https://bit.ly/3v1qFC5

[42] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 2623–2631.

[43] J. S. Tham, Y. C. Chang, and M. F. A. Fauzi, "Automatic identification of drinking activities at home using depth data from RGB-D camera," in *Proc. Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, Dec. 2014, pp. 153–158.

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

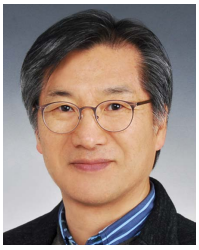
[45] M. Cormier, A. Clepe, A. Specker, and J. Beyerer, "Where are we with human pose estimation in real-world surveillance?" in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 591–601.

SARDOR JURAEV (Student Member, IEEE) received the B.S. degree from Inha University, in 2021, where he is currently pursuing the M.E. degree with the Department of Electrical and Computer Engineering. His research interests include deep learning and its applications to computer vision, and intelligent video surveillance.

AKASH GHIMIRE is currently pursuing the B.E. degree with the Department of Integrated System Engineering, School of Global Convergence Studies, Inha University. His research interests include deep learning and its applications to computer vision, human pose estimation, and intelligent video surveillance.

JUMABEK ALIKHANOV (Student Member, IEEE) received the B.S. degree from the Tashkent University of Information Technology, in 2014, and the M.E. degree from Inha University, in 2017, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering. His research interests include machine learning and its applications to computer vision, sensor data science, and natural language processing.

VIJAY KAKANI (Member, IEEE) received the B.Tech. degree in electronics and communication engineering from Jawaharlal Nehru Technological University, Kakinada, India, in 2012, the M.S. degree in computers and communication systems from the University of Limerick, Ireland, in 2014, and the Ph.D. degree in information and communication engineering (major) and future vehicle engineering (minor) from Inha University, South Korea, in 2020. He is currently an Assistant Professor with the Department of Integrated System Engineering, School of Global Convergence Studies, Inha University. His research interests include autonomous vehicles, sensor signal processing, applied computer vision, deep learning, systems engineering, and machine vision applications.



HAKIL KIM (Member, IEEE) received the M.Sc. and the Ph.D. degrees in electrical and computer engineering from Purdue University, in 1985 and 1990, respectively. In 1990, he joined the College of Engineering, Inha University, Incheon, South Korea, where he is a Full Professor with the Department of Information and Communication Engineering. In order to retain the balance between academic research and commercial development, he founded Vision Inc., in 2014, and is currently the CEO of the company. His research interests include biometrics, intelligent video surveillance, and embedded vision for autonomous vehicles. Since 2003, he has been actively involved as the Project Editor in the International Standardization of Biometrics at ISO/IEC JTC1/SC37.