IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 February 2024, accepted 6 March 2024, date of publication 18 March 2024, date of current version 22 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3375767

RESEARCH ARTICLE

Fall Detection in Low-Illumination Environments From Far-Infrared Images Using Pose Detection and Dynamic Descriptors

JESÚS GUTIÉRREZ[1], SERGIO MARTIN[2], (Senior Member, IEEE), VICTOR H. RODRIGUEZ[3], SERGIO ALBIOL[3,4], INMACULADA PLAZA[3,4], (Senior Member, IEEE), CARLOS MEDRANO[3,4], (Senior Member, IEEE), AND JAVIER MARTINEZ[3]

[1]Escuela Internacional de Doctorado, Programa Tecnologías Industriales, UNED, 28015 Madrid, Spain
[2]Electrical and Computer Engineering Department, UNED, 28040 Madrid, Spain
[3]EduQTech, E.U. Politécnica de Zaragoza, 44003 Teruel, Spain
[4]IIS Aragón, Universidad de Zaragoza, 50009 Zaragoza, Spain

Corresponding author: Sergio Martin (smartin@ieec.uned.es)

This work was supported in part by the In4Labs (Open platform to facilitate the development of Industry 4.0 remote laboratories) Project funded by Ministerio de Ciencia, Innovación y Universidades (MCIU)/AEI/10.13039/501100011033 under Grant TED2021-131535BI00; in part by European Union (EU) NextGenerationEU/PRTR; in part by Gobierno de Aragón under Grant T49_23R; and in part by the Education—Quality—Technology (EduQTech) Group, Centro Politécnico de Teruel, Universidad de Zaragoza.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by UNED's Ethics Committee.

ABSTRACT In an increasingly aging world, the effort to automate tasks associated with the care of elderly dependent individuals becomes more and more relevant if quality care provision at sustainable costs is desired. One of the tasks susceptible to automation in this field is the automatic detection of falls. The research effort undertaken to develop automatic fall detection systems has been quite substantial and has resulted in reliable fall detection systems. However, individuals who could benefit from these systems only consider their use in certain scenarios. Among them, a relevant scenario is the one associated to semi-supervised patients during the night who wake up and get out of bed, usually disoriented, feeling an urgent need to go to the toilet. Under these circumstances, usually, the person is not supervised, and a fall could go unnoticed until the next morning, delaying the arrival of urgently needed assistance. In this scenario, associated with nighttime rest, the patient prioritizes comfort, and in this situation, body-worn sensors typical of wearable systems are not a good option. Environmental systems, particularly visual-based ones with cameras deployed in the patient's environment, could be the ideal option for this scenario. However, it is necessary to work with far-infrared (FIR) images in the low-light conditions of this environment. This work develops and implements, for the first time, a fall detection system that works with FIR imagery. The system integrates the output of a human pose estimation neural network with a detection methodology which uses the relative movement of the body's most important joints in order to determine whether a fall has taken place. The pose estimation neural networks used represent the most relevant architectures in this field and have been trained using the first large public labeled FIR dataset. Thus, we have developed the first vision-based fall detection system working on FIR imagery able to operate in conditions of absolute darkness whose performance indexes are equivalent to the ones of equivalent systems working on conventional RGB images.

INDEX TERMS Computer vision, convolutional neural, fall detection, infrared imaging.

I. INTRODUCTION

The United Nations report on population aging [1] suggests that the number of people aged over 60 has doubled since

The associate editor coordinating the review of this manuscript and approving it for publication was Shovan Barma.

1980 and is expected to double again around 2050, reaching the figure of 2 billion. At that time, the number of individuals over 60 will exceed that of individuals under 24.

While this demographic aging phenomenon is more pronounced in developed countries, it actually affects developing nations as well, where over two-thirds of the world's population over 60 reside. Consequently, we can anticipate an exponential increase in resources dedicated to elderly care in the coming years, and in the near future, this sector could potentially become one of the most economically significant. This projected increase in resources has led to a substantial rise in research interest in areas related to elderly care.

The elderly care sector, at present, has automated a limited number of processes. However, considering the described expectations, if societies aim to provide high-quality care to this community at a reasonable cost, the sector will need to automate as many tasks as reasonably possible.

One of the areas that could accept automation is the fall detection one. It is a very relevant activity, as for the aforementioned community over 30% of falls have important consequences, ranging from hip fractures to concussions and, a good number of them, end up by causing death [2].

In this paper a novel fall detection system able to work on far-infrared (FIR) imagery is proposed. It is based on three key elements: i) the use of FIR images; ii) A human pose detection system able to infer the position of the most important body joints; iii) A fall detection system fed by the position of the main body joints provided by step ii able to determine whether a fall has taken place. In section II-C it is argued that the proposed system overcomes several limitations of current system. Firstly, it protects privacy, a major concern of potential vision-based fall detection system's users [3], in scenarios where they accept the use of this kind of systems, i.e., older individuals getting up at night with absent human supervision. Secondly, it addresses the insufficient amount of real world data [4] by employing dynamic descriptors that hold physical meanings related to human balance. Therefore, the goal of this paper is the implementation and performance evaluation of a visual-based fall detection system able to operate in low-illumination environments, a scenario not properly covered by the present systems.

Thus, the main contributions of this paper to the state of the art are:

- The implementation of the first fall detection system working on FIR imagery and, therefore, the only high resolution image-based system able to operate under low-illumination conditions ever developed. This system integrates the output provided by a pose estimation neural network and the fall detection system proposed by the authors in [4] with excellent results. This way, developing this system requires human pose neural networks properly trained using FIR imagery.
- The development of FIR-human, the first major labeled FIR dataset developed to train human pose estimation neural networks. It contains video-clips of a number of volunteers engaged in different activities and falls. All major joints of volunteers are labeled in each of the images of these video-clips, so the dataset can be used with training and validation purposes in the field of human pose estimation. The dataset, which contains over 250000 labeled images, has been made public for research and can be found at https://ieee-dataport.org/documents/fir-human.
- A comparative study of the performances of the eight most relevant neural network architectures used in the field of two-dimensional human pose estimation after they have been trained using FIR imagery. These architectures include both the convolutional and transformer families and cover both the direct regression and heat-map approaches and their outputs are used to feed the fall detection system.

Regarding the outlined contributions, the first one is considered the most relevant one, with the others serving as necessary elements to achieve it.

The rest of this paper is structured as follows. Section II presents a discussion on the evolution of human pose estimation systems, as well as the most relevant neural network architectures used in this field to determine the position of the major human joints from an image. Additionally, the most used datasets used in this field to train and compare the presented neural networks are introduced. Finally, the techniques currently employed for automatic fall detection are described and how they are not suitable for addressing the problem of fall detection in nighttime conditions for unsupervised elderly individuals.

In Section III, FIR-Human is described, the first publicly available database containing FIR video clips along with annotations of the volunteers' joint positions. This database allows training the networks presented in Section II from scratch, an option not explored until now, as all previous works used visible spectrum images to train such networks. Additionally, this section describes the methodology for training neural networks using FIR-Human and the performance indexes used to test the human pose estimation networks. Finally, the performance indexes employed to evaluate the fall detection system as a whole are also presented.

Section IV compiles the obtained results, both regarding the accuracy in determining joint positions and the system's overall ability to determine if a fall occurs. Subsequently, these results are discussed, evaluating the reasons for imprecise position estimation under certain conditions and the causes of errors in fall detection.

Finally, in Section V, final conclusions are drawn, and future lines of work are proposed.

II. BACKGROUND AND RELATED WORK

A. HUMAN POSE ESTIMATION

Human pose estimation has traditionally been one of the most challenging fields of study in computer vision, as determining the position of body key-points has proven to be an elusive task.

This process involves determining the pose of the image in either two or three dimensions, and a number of different approaches have been proposed in the literature to solve it. All these approaches can be classified into two different groups [5]; the generative and the discriminative ones.

Generative human pose estimation methods involve a geometric projection of a volumetric human body model onto the image plane, aligning it with the observed image. These methods focus on solving the intrinsic problem of pose estimation by assessing the probability of an observation given a pose of the model. In this way, the process, which aims to find an absolute minimum, requires a complex search over the model state space to achieve it. A good number of systems have been developed following these methods [6], [7], [8], [9], and, as expected, they are susceptible to errors related to local minima, necessitating accurate initial pose estimations to mitigate this issue. The most common methodologies to obtain the sought minimum are local optimization [10], [11], [12] and stochastic search [13], [14].

Generative methods deliver good results under optimal conditions. However, under poor lighting or occlusion conditions, their performance is significantly degraded.

Unlike generative methods, discriminative ones are capable of establishing a direct relation between the array of features collected from images and a set of different poses. As a result, multi-dimensional boundaries separating classes associated with poses can be determined for the array of features.

The determination of boundaries requires system training based on real data, which takes time and demands processing power. However, once the boundaries have been established, the amount of processing power and time required for pose determination is much lower than that required by generative models. This is because generative models need to go through an optimization process in a high-order state space every time they estimate a pose.

The most popular discriminative methods to estimate human pose are support vector machines [15], [16], [17], Relevance Vector Machines [18], [19], mixture of experts [20], [21], manifold learning methods [22], [23], [24], [25], [26], pose embedding methods [27], [28], locality-constraint linear code like [29], [30], bag-of-words [20], random forest [31], [32], [33] and deep learning methods.

In global terms, discriminative methods, although showing higher resistance than generative ones to performance degradation due to occlusion and poor lighting conditions, still present important restrictions under those circumstances.

Deep learning methods, while computationally more expensive than other ones, have proven to be not only more accurate than the rest in optimal conditions, but also more resilient to the adverse impact of occlusion and poor lighting. These characteristics have made them gain high relevance over the last few years.

In broad terms, human pose estimation based on deep learning models consists of two basic steps. During the first step, the model focuses on joint recognition (e.g., shoulder, knee, ankle), while the second phase is centered on joint grouping, so that the array of joints configures a valid human pose configuration.

Two common approaches have been used for pose estimation of individuals in images. The first one, known as top-down, identifies the number of individuals in the image and isolate them, usually by creating a bounding box. Once individuals have been isolated, the system focuses on joint identification and pose determination. This first philosophy is followed by a number of different artificial neural networks (ANN) architectures [34], [35], [36]. The second approach [37], [38], [39], bottom-up, follows a reverse logic and start by identifying joints to group them together afterwards in a coherent entity representing a person.

The bottom-up philosophy presents several advantages over the top-down option, as it is better able to overcome early commitment problems associated with a faulty detection of individuals in the image. Furthermore, although the computational cost of top-down approaches is lower than that of the bottom-up one when the number of individuals is low, it becomes higher as the number grows.

While ANN architectures able to estimate human pose is large, all of them deliver one of the two following outcomes. They can either directly regress the coordinates of a person's joints, or they can generate a probability map, called a heat map, which represents the likelihood that an area of an image contains a specific joint.

Traditionally, the backbone architectures used in human pose estimation are based on convolutional neural networks. DeepPose [40] is the first relevant network in this area presented in a research paper. It uses a classical convolutional network as a backbone, Alexnet. Since then, a number of alternative convolutional architectures capable of delivering heat-maps or regressing joint positions have been proposed.

The introduction of ViT [41] meant the introduction of an alternative option to the use of convolutions in the field of artificial vision. This alternative is based on the use of transformers, an architecture developed for the field of natural language processing which identifies how relevant an item of the input vector is for the rest of elements.

Visual transformers have been very recently introduced in the world of human pose estimation and the number of proposed networks for this purpose based on them is still limited. This new architecture can rival state of art convolutional networks and, in certain occasions, where relative positions become relevant, it can outperform them. However, although the computational cost for equivalent results tends to be lower, the transformers architectures require far more training information than convolutional networks to reach equivalent performances [42].

Following the criteria stablished in [43], which thoroughly revises the most important two-dimensional human

estimation neural networks belonging to both groups, convolutional and transformers, the most relevant architectures are selected.

These architectures, which are extensively presented in the reference papers that introduce them, are briefly described in the following paragraphs. They can be reproduced from the papers that introduce them, but unlike what is done in their original design, where they are trained using conventional RGB images, we have trained them using FIR images to assess their performance with this type of imagery.

1) CONVOLUTIONAL ARCHITECTURES

a: DeepPose

DeepPose is a convolutional architecture presented in [40] whose backbone is Alexnet [44]. The network uses a cascade of regresssors to refine joint position determination. It crops the image around the joint coordinates estimated by the previous stage, so further stages can improve joint position determination, as these new-cropped images have higher resolution levels.

DeepPose has three stages that operate in cascade. All stages make the input image go through 5 convolutional layers reducing horizontal complexity to gain depth information before injecting the extracted features in a block of two fully connected layers that regresses joints positions. Then, the image is cropped around that point and it is passed to the next stage of the cascade for a more precise joint regression.

b: ConvNet POSE

ConvNet POSE is a convolutional architecture presented in [45]. It represents the first approach to heat-map generation leaving the previous regression philosophy used by DeepPose.

The architecture integrates three modules. The first one produces a coarse heat-map generated by convolutional and pooling layers. A second module crops the image around the predicted position of every joint. Finally, the third module is used for heat-map fine-tuning.

c: CONVOLUTIONAL POSE MACHINES

Convolutional pose machines, presented in [34], is an architecture capable of producing an array of 2D heat maps that represent the probability distribution in space for the location of each key-point. The architecture is multi-stage and end-to-end trainable. In the initial stage, the input data is the original image, which, after being processed by a standard Visual Geometry Group structure, generates heat maps for every joint. Subsequent stages employ a similar strategy, although the input data is an aggregation of the heat maps produced by the previous stage and the original image.

This architecture uses large receptive fields to learn spatial relationships that, in conjunction with the combined input of the original image and the heat maps generated by the previous stage, enhance the accuracy of the network output.

d: STACKED HOURGLASS NETWORKS FOR HUMAN POSE ESTIMATION

The stacked hourglass architecture presented in [35] takes its name from its appearance, which resembles an hourglass. It combines the bottom-up and top-down approaches, as the initial layers of each stage are convolutional and reduce horizontal complexity while gaining depth and the final layers are deconvolutional and execute the reverse operation. This structure captures local information contained in the image at different scales, which allows the network to learn different relationships, such as body position, limb movements, and the relationship between joints.

The architecture stacks several hourglass stages in order to get optimal performances and the down-sampling effect is obtained using max pooling techniques while the up-sampling one uses nearest-neighbor interpolation.

e: HUMAN POSE ESTIMATION WITH ITERATIVE ERROR FEEDBACK

Iterative error feedback, presented in [46], is an architecture capable of identifying what is wrong in the network's forecast and correcting it in an iterative way. This approach incorporates error predictions into the initial solution to iteratively correct and optimize joint position determination. Unlike the previous method, which directly identifies key-point positions, this approach progressively corrects an initial forecast to optimize it.

This multi-stage process uses the fusion of the original image and the heat-map produced by the previous stage as the initial stage data. With this information, the errors in the predictions from the previous stage are forecasted, and joint positions are updated accordingly. These updated joint positions are then used to generate updated heat-maps, which will serve as input, along with the initial image, for the next stage.

f: CASCADE FEATURE AGGREGATION FOR HUMAN POSE ESTIMATION

This architecture, presented in [36], is based on a cascade of hourglass stages that aggregate predictions from previous stages with the original output of the initial backbone, aiming to better capture the local information contained in an image.

This approach enables feature aggregation through image inspection at different levels. Human joints are located through low-level inspection, while in complex environments with poor lighting or occlusion conditions, high-level inspection helps to refine their position.

2) TRANSFORMERS ARCHITECTURES

a: TFPose

TFPose, which is presented in [47], is an architecture based on transformers that directly regresses key-point positions. Its backbone extracts multilevel feature maps by processing the input image through a series of convolutional layers with

increasing strides. These maps are then flattened and concatenated together to feed a transformer-encoder block, following a Deformable DERT [48] design. This encoder block consists of six consecutive encoder layers, taking as input the output of the previous one. Finally, a decoder block is used to regress the coordinates of all joints from the encoder block output.

The novelty of methods based on transformers is the attention mechanism they implement in the encoder block. This way, input images or their feature maps are divided into spatial segments, and the encoder block determines the level of importance of every segment in relation to the rest, allowing the network to learn spatial relations among joints in this case.

b: ViTPose

This architecture, presented in [49], is based on transformers and can produce key-point heat-maps. Unlike the previous network, ViTPose is purely based on transformers and does not use convolutions to extract multilevel feature maps. In this case, the input image goes through an embedding block, which fragments it into tokens that are flattened and concatenated into a single tensor. This tensor then feeds the encoder block, which consists of a series of transformer sub-blocks, with each one feeding the following one.

Similar to the previous network, the objective of the encoder block is to learn the relative spatial relationships among body key-elements to work effectively in cluttered environments with low illumination or occlusion conditions.

Finally, a decoder block, fed by the encoder block's output, produces an array of heat-maps associated with each joint.

B. DATASETS USED FOR HUMAN POSE ESTIMATION SYSTEM TRAINING

This work introduces FIR-Human, the first public dataset of its kind, including far infrared video clips of people engaged in daily life activities and falls, providing both two and three-dimensional coordinates of their major joints. The quantity of images within FIR-Human is sufficient to train human pose estimation neural network capable of inferring the position of human joints in both two and three dimensions from scratch. In this work that training will be limited to two dimensional networks, as the fall detection system we implement works with that input. Prior to the introduction of FIR-Human, previous works implementing pose estimation networks able to infer joint's position used RGB images for network training and small sets of labeled FIR images for performance testing. This is the case of the systems proposed in [50] and [51], where the network is evaluated using reduced FIR datasets specifically designed for that purpose.

Despite the absence of FIR datasets used for pose estimation system training, there is a substantial number of datasets containing RGB images for the training of human pose estimation neural networks. The most important ones are FLIC [52], LSP [53], MPII Human Pose [54] and COCO [55], which are systematically used to train and validate pose estimation architectures. Additionally, Human 3.6M [56], MPI-INF-D-HP [57], NTU RGB+D [58] and Deepcap [59] have also been used with this purpose.

AUTOMATIC FALL DETECTION

The research effort to develop automatic fall detection systems has been quite substantial. Tasnim et al. state in [60] that the number of research papers found on Google Scholar in 2022 exceeds 4000. This endeavor has produced positive results, creating systems capable of reliably performing automatic fall detection. Traditionally, these systems have been categorized into three types: wearable, ambient, and vision-based [61].

All systems, irrespective of their category, follow a common approach to fall determination or gait analysis. They all process signals related to the person's movement and, in one way or another, define that movement. The signal is typically pre-processed to minimize noise, and then it is analysed to infer movement descriptors. Finally, these descriptors that define movement are classified using various techniques to ascertain whether a fall has occurred or if a specific gait aligns with the requirements associated with a high fall probability.

Wearable systems employ sensors carried by the monitored person to assess their movements. The predominant sensors in these systems are either accelerometers or gyroscopes, although other types, such as microphones, pressure sensors, electrocardiography, and electroencephalography, are also utilized. Despite the drawbacks related to limited connectivity capabilities and restricted edge processing power, these devices have attained a high level of maturity.

Ambient fall detection systems rely on contact, passive infrared, acoustic, radar, or Wi-Fi technologies. Their primary distinction, compared to wearable devices, lies in sensor placement. While in wearable systems, sensors are carried by the monitored person, in the case of ambient ones, sensors are positioned around them.

While these systems offer a significant advantage over wearable ones, as they are not dependent on batteries, their level of maturity is still limited. As a result, most commercial fall detection systems are currently wearable.

Vision-based systems operate with visible spectrum, near-infrared, or depth video inputs. In tandem with the advancement of artificial vision technologies, primarily driven by the use of artificial neural networks, a substantial amount of effort has been invested in recent years to develop these types of devices. This significant research endeavor, as evidenced by the number of published research papers in the area [62], has enabled vision-based fall detection systems to attain a noteworthy degree of maturity.

Despite the generally limited acceptance of these systems [3], the majority of commercial automatic fall detection systems use wearable technologies, particularly the inertial ones, while a good number of the most recent ones are based on vision-based technologies. This is likely a reflection of the technology's maturity, which, in turn, may indicate how

well-suited a certain technology is for addressing the issue of fall determination.

Nevertheless, as indicated in [3], the utilization of these systems may be perceived as acceptable by the elderly community and their caregivers in specific situations, provided that they are truly tailored to the user's needs. These situations are primarily linked to times when human supervision is minimal or absent [63]. A common scenario of this nature, described by numerous interviewees in [3], is associated with semi-supervised patients getting up at nighttime. In this scenario, the individual is often disoriented, heightening the risk of a fall. Furthermore, in the event of a fall, it might go unnoticed until the next day, significantly delaying potentially necessary medical intervention.

In these situations, the use of inertial-based systems is a suboptimal solution, as the requirement to wear sensors attached to the body disqualifies its use in situations where patients wear light outfits to maximize comfort during sleep. Ambient systems might appear to be the optimal choice in these conditions, but the current low maturity of these technologies discourages their use. Vision-based systems combine good technological maturity with optimal operational conditions under the described circumstances. However, the low illumination conditions associated with these scenarios disqualify the use of visual or near-infrared cameras.

In contrast, FIR sensors and their images are perfectly suitable for this situation, as the images they provide are not dependent on light. Moreover, their use contributes to privacy protection, and several groups related to elderly care have expressed a preference for them. Finally, the introduction of low-cost, high-resolution FIR cameras allows the development of a system with these characteristics at a very low price. Consequently, automatic fall detection based on FIR imagery could be the optimal approach to address the safety problem posed by semi-supervised patients getting up at night-time.

Besides, a key aspect of practical fall detection is the generalization problem associated to a lack of real data is extensively studied in [4]. This way, the vision-based automatic fall detection system revision made in [62] concludes that, although the systems' performances are very satisfactory in laboratory environments, the significant differences between simulated and real falls, and between falls of elderly and young people, documented in [64] and [65], as well as the difficulty to access real-world data as a consequence of privacy protection, yield reasonable doubts about their performances in real circumstances.

These doubts are a direct consequence of the use of kinematic descriptors [62] to evaluate whether a fall has taken place. These descriptors are features inferred from the falls contained in the datasets used in system training. This way, if the video datasets do not contain real falls of elderly people, the obtained descriptors could be inaccurate or incorrect and the system performances in the real-world could be poorer than expected. The addressed problem is, therefore, a problem of generalization or, in other words, how to assess whether an elderly person has experienced a fall based on information inferred from simulated falls performed by young people who fall in a different way.

Generalization problems are addressed through a number of methods based on the correction of the modeling errors associated to a training phase based on non-comprehensive information. This family of approaches, considered in [66] for the field of data-driven fault diagnosis and in [67] and [68] for the field of automatic control, implies access to, at least, a limited amount of real data after initial system training in order to correct the modeling errors caused by non-comprehensive training information. Unfortunately, the absence of any video database containing real falls of elderly people makes the implementation of this approach impossible in the field of video based automatic fall detection.

In [4], this generalization problem is approached from a different perspective than the conventionally adopted approaches based on the use of kinematic descriptors. In classical approaches, descriptors that characterize a fall are inferred from fall images contained in datasets developed for research using different techniques. However, since these images have been recorded by actors or volunteers who are significantly younger than the elderly community, it can be expected that kinematic descriptors, inferred for a young population and generalized to the rest, are not suitable for a monitoring system devoted to elderly individuals.

This way, given the absence of real data, alternative approaches need to be implemented to address this generalization problem. In [4], the use of a neural network called CoGNet is proposed to determine the center of gravity position of a person projected onto the ground, as well as the support polygon defined by their feet, based on the position of their joints. This position is determined by bi-dimensional human pose estimation networks.

This approach, based on the use of dynamic descriptors, approximates the human body in terms of stability and balance. This way, the described generalization problems can be overcome, since the differences between real and simulated falls lose relevance, as all falls are a direct result of failures in the continuous effort of the body to maintain balance, regardless of other considerations.

This work implements, for the very first time, a fall detection system that operates with high-resolution FIR imagery. All previous works in the field of fall detection using infrared sensors are limited to the use of very low-resolution Passive Infrared Sensors (PIR). The studies conducted in this regard are described by Ben-Sadoun et al. [69], which compiles research published by IEEE Xplore Digital Library, MEDLINE (PubMed), MDPI, SpringerLink, and ScienceDirect. In total, it identifies 15 systems based on these technologies, utilizing various types of sensors with resolutions ranging from 8×8 pixels to 16×16 pixels, figures significantly distant from the 480×640 pixels used by our system.

The datasets used to evaluate the performance of these systems were specifically developed for each of them, making it impossible to establish comparisons between them or with other systems. In any case, with the exception of two systems whose performances were lower, their accuracy, precision, sensitivity, or specificity fell within the range of 85-90%.

In a subsequent review, Elagovan et al. [63] studied the night fall detection systems, concluding that the number of systems using infrared sensors is very small and that none of them use FIR imagery, as all identified systems use very low-resolution PIR sensors.

C. SUMMARY

To sum up this section, it has been argued that practical and widespread use of fall detection systems faces problems related to user acceptance and generalization abilities of the automatic detection system. Getting up at nighttime is a scenario in which user are willing to accept an automatic supervision system. The use of FIR images preserves users' privacy, overcoming users' concerns and avoiding wearing uncomfortable sensors. Moreover, using dynamic descriptors allows overcoming the generalization problem. Thus, in this paper we propose to implement a fall detection system from FIR images. The system includes two stages: a pose detection system, and then a fall detection system based on dynamic descriptors obtained from the pose.

The implementation of an automatic fall detection system that works with FIR images and incorporates the concept of dynamic descriptors requires bi-dimensional human position estimation methods properly trained on FIR imagery. This way, it requires a dataset specifically developed for this purpose, as there is none with these characteristics. Thus, this works introduces FIR-Human, as a novel dataset that fulfils these requirements.

III. MATERIALS AND METHODS

A. FIR-HUMAN

Our new dataset, called FIR-Human, is the only one of its kind to the best of our knowledge. It includes video clips recorded by five volunteers engaged in various activities. The dataset contains over 250,000 far infrared images (FIR) of the volunteers, along with the 3-dimensional and 2-dimensional annotations of their 19 main body joint positions.

The group of five volunteers is composed of four males and one female, with body mass indexes (BMI) ranging from 16 to 24, ages spanning from 25 to 56 and heights that vary within the interval of 161 cm to 180 cm. This ensures variability in body types and movements. Additionally, the volunteers wore a variety of clothes, and the thermal conditions of the laboratory where the recording was conducted were highly diverse. In this way, the dataset obtained is rich and varied. All volunteers signed an informed consent to participate in the research.

The potential uses of this dataset include training systems for FIR human pose estimation in both 2 and 3 dimensions, human action recognition based on FIR imagery, surveillance, healthcare, and potentially, autonomous driving.

FIR-Human is publicly available for download for academic and research use under the conditions established in the license agreement at https://ieee-dataport.org/documents/fir-human.

1) DATA MODALITIES

A Seek FIR camera is used to record our dataset, which is synchronized with a Qualisys MoCap (Motion Capture) system. The MoCap system captures the 3-dimensional position of markers placed on the main body joints. Through this process and after appropriate processing, a sequence of 3-dimensional positions for each joint in each video clip is collected.

The FIR video clips are recorded at 23.98 frames per second, and each frame has a resolution of 480×640 pixels. The joint information consists of 3-dimensional positions of 19 major body joints (Head top, forehead, neck base, shoulders, elbows, wrists, hips, knees, ankles, heels and toes tips), defined with an error of less than 5 millimetres for each recorded frame. Additionally, the 2-dimensional projection of those coordinates onto the recording plane is also provided.

2) ACTION CLASSES

The dataset contains a total of 27 action classes. Twenty-six of them represent daily life activities, while the remaining one includes different types of falls. The various actions are repeated by the volunteers in four different positions, allowing for frontal, rear, and side views of the same actions to be recorded.

The dataset is divided into three blocks. The first block, which includes the motions of four volunteers, is used for system training, and in this group, all volunteers are recorded executing 13 daily life activities. The second block includes a single person who performs a different set of actions for validation purposes. Finally, the third block includes four volunteers who are recorded from different perspectives falling forward, falling backward, and side falling. The falls start from static or dynamic situations, and a number of them are low-energy slow falls, a common type of fall in the elderly community [70]. All the activities executed in the different blocks are described in Table 1.

A few examples of labeled images belonging to video-clips of volunteers performing different activities and fallings can be seen in figure 1.

B. TWO-DIMENSIONAL POSE ESTIMATION

Eight different state-of-the-art neural networks, which represent the most relevant architectures used in the field of human pose estimation have been used to determine the position of the main human joints. These architectures, introduced in section II-A, include DeepPose [40], ConvNet POSE [45], Convolutional pose machines [34], Stacked hourglass [35], Iterative Error feedback [46], Cascade feature aggregation [36], TFPose [47] and ViTPos [47]. No changes have been introduced regarding the structure

TABLE 1. Description of activities.

	Block 1	Block 2	Block 3
Activities	Giving directions	Brushing teeth	Forward, backward and lateral falls
	Discussing	Encouraging your team	
	Eating	Toasting	
	Taking photos	Taking a selfie	
	Exercising on the ground	Crouching for meditation	
	Running in place	Walking a dog	
	Walking	Throwing a stone	
	Sitting and standing up	Talking on the phone	
	Coughing	Stretching yourself	
	Exercising	Hopping	
	Playing basketball	Kicking a ball	
	Picking up objects	Tying shoelaces	
	Limping	Rotating trunk	

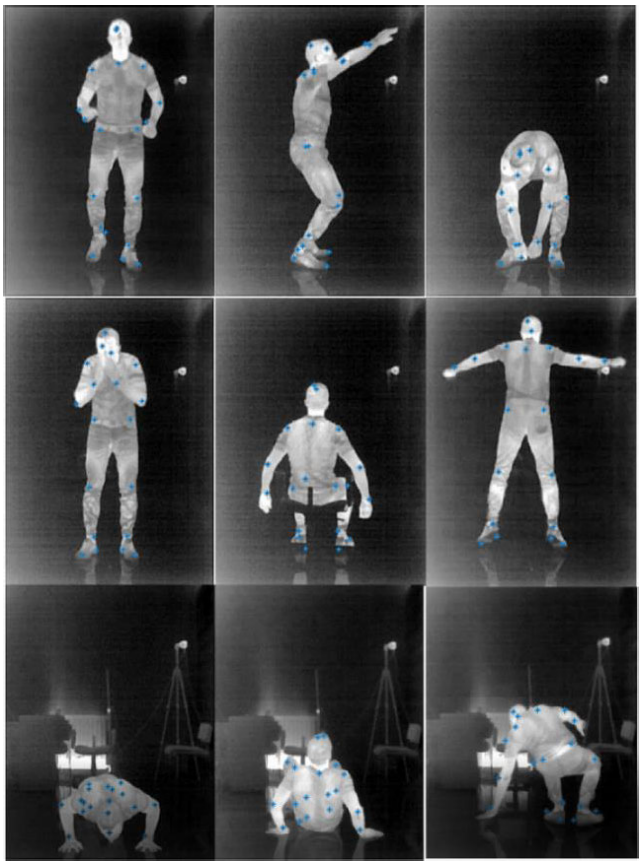


FIGURE 1. Volunteer (a) running, (b) playing basketball, (c) picking up an object, (d) coughing, (e) sitting, (f) exercising, (g) falling forward, (h) falling backwards, (i) side falling.

or hyper-parameters presented in the reference papers that introduce them, as the FIR imagery is fed into the network using the standard thee-channel input philosophy used with RGB images. Finally, it is worth noting that, unlike what is done in the original reference papers, the networks are trained using FIR images instead of conventional RGB imagery.

1) NETWORK TRAINING

All selected networks are trained using the first block of the FIR-Human dataset. This section of the dataset contains 149635 annotated images of four volunteers engaged in 13 different daily life activities. The diversity in sample amplitude, body types (ample range of BMI's and heights as well as sex as explained in III.A) and clothing, along with the various thermal conditions of the laboratory (whose temperature, depending on the clip, range from 16 to 31 degrees), provide optimal conditions for network training.

The block one of the FIR-Human dataset was used for network training while the block two was employed for testing and network comparison. To enrich both blocks as much as possible the data augmentation strategy proposed in [71] was adopted. It includes random rotations (45°, -45°), random scaling (0.65, 1.35), flipping and half body data augmentation. This way, the total number of images was multiplied by four.

All proposed networks are implemented using PyTorch with an Adam function used as optimizer. The chosen batch size was 32 images and all networks were trained for 220 epochs, a number high enough for all of them to show a stable behavior. The initial learning rate was 10-3 and it was dropped to 10-4 and 10-5 at the 160 and 200 epochs respectively, following the same rationale explained in [72].

All networks were trained on an NVIDIA RTX-3080 GPU.

2) LOSS FUNCTION

The used loss function in all cases is an L2 function, also known as Mean Squared Error (MSE), which is calculated as the average of the sum of all squares of the differences between true and predicted values.

$$L_2 = \frac{\sum_{i=1}^{i=n} (y_i - f(x_i))^2}{n} \quad (1)$$

where y_i is the real value and $f(x_i)$ is the network's forecasted one.

In spite of its sensitivity to outliers this function is usually preferred over L1, as it allows an easier gradient determination, favoring this way network training.

3) HUMAN POSE ESTIMATION EVALUATION METRICS

A number of evaluation metrics allow network performance evaluation and comparison over a common dataset.

The most important ones are:

- PCP (Percentage of Correct Parts), which measures the correct detection rate of limbs, considering the detection as correct when the distance between the two predicted joint locations and the true ones is less than half the limb length [40].
- PDJ (Percentage of Detected Joints). This metric regards the detection of a joint as correct when the distance between the forecasted and real joint positions is below a percentage of the distance between right hip and left shoulder.
- PCK (Percentage of Correct Key-points). A metric similar to PDJ, although in this metric, the reference distance is the maximum side length of the external rectangle of ground truth body joints [73]. PCKh is a variation of PCK, whose reference distance is defined as 50% of the ground-truth head segment length [54]. PCKh@0.5, by far the most used evaluation metric in the field of human pose estimation, considers the joint correctly detected when the error in forecasting is below 50% of the PCKh reference distance.

Due to the generalized use of PCKh, a common metric used in all the papers that present and evaluate the architectures described in the previous paragraphs, PCKh will be used in this work as the common metric for comparison.

C. FALL DETECTION EVALUATION METRICS

The system integrates a 2D human pose estimation network, CoGNet, and the fall detection algorithm described in [4] to address the problem of fall detection in poorly illuminated environments.

For 2D human pose estimation, the networks previously considered are utilized, as they represent most of the state-of-the-art human pose recognition networks ever developed. The output of these networks is a matrix containing the 2D positions of the main body joints, which is then passed to CoGNet. CoGNet is responsible for assessing, by using the algorithm proposed in [4], whether a fall has taken place.

It is important highlighting that CoGNet is already trained using datasets different from FIR-human [4] and, therefore, the blocks 2 and 3 of FIR-Human are used solely for system testing purposes. The datasets used to train CoGNet are Human 3.6M [56], MPI-INF-D-HP [57], NTU RGB+D [58] and Deepcap [59].

The dataset used for testing consists of the 72 falls from the block 3 of FIR-Human dataset. Additionally, the video-clips of block 2 from that dataset are split into 8-second clips, resulting in 195 videos that show a person executing 13 different daily life activities. This way, 72 video-clips containing falls (FIR-Human block 3) and 195 video-clips containing daily life activities different from falling (FIR-Human block 2) are used for system testing purposes.

TABLE 2. PCkh@0.5 for the different human joints.

MODEL	PCKh@0.5			
	Head	Shoulder	Elbow	Wrist
DeepPose (ResNet - 101) [40]	87.9	79.3	76.8	75.2
ConvNet Pose [45]	96.7	91.2	83.7	78
CPM [34]	98.6	91.3	86.8	87.2
Stacked hourglass (3 Stages) [35]	98.8	95.6	91	87.3
HPE IF [46]	96.3	90.9	81.3	72.6
Cascade (ResNet-101 Cascaded with 2 ResNet-50) [36]	96.5	94.7	90.8	87.1
TFPose (Resnet -50; Nd=6) [47]	98.6	95.2	90.8	86.2
ViTPose (ViTAE-G) [49]	98.6	96.9	94.3	92.1
	Knee	Ankle	Foot	Hip
DeepPose (ResNet - 101) [40]	70.7	49.8	47.2	71.1
ConvNet Pose [45]	80.6	64.6	61.5	81
CPM [34]	88.7	78.1	74	89.1
Stacked hourglass (3 Stages) [35]	89.8	83.4	78.6	90.2
HPE IF [46]	82.4	66.2	63.9	82.8
Cascade (ResNet-101 Cascaded with 2 ResNet-50) [36]	89.5	83.7	82.9	89.9
TFPose (Resnet -50; Nd=6) [47]	89.5	82.4	80.4	89.9
ViTPose (ViTAE-G) [49]	92.6	90	89.1	93

The metrics used to evaluate the fall detection system's performance, which are the most common ones in this area [62], are the following ones:

$$Sensitivity\_SE = \frac{TP}{TP + FN} x100 \tag{2}$$

$$Specificit\_SP = \frac{TN}{TN + FP} x100 \tag{3}$$

$$Accuracy\_AC = \frac{TP + TN}{TP + TN + FN + FP} x100 \tag{4}$$

where TP, TN, FP and FN respectively stand for true positives, true negatives, false positives and false negatives.

IV. RESULTS AND DISCUSSION

A. TWO-DIMENSIONAL POSE ESTIMATION EVALUATION

Performance comparison is summarized in figure 2, where PCK is represented as a function of the normalized head segment length.

Table 2 collects the performances of the different networks as a function of joint.

Table 3 presents the computational cost required by the different models while figures 3 and 4 compare the relation between performance (total PCKh@0.5 in the first case and foot PCKh@0.5, the most difficult joint to place, in the

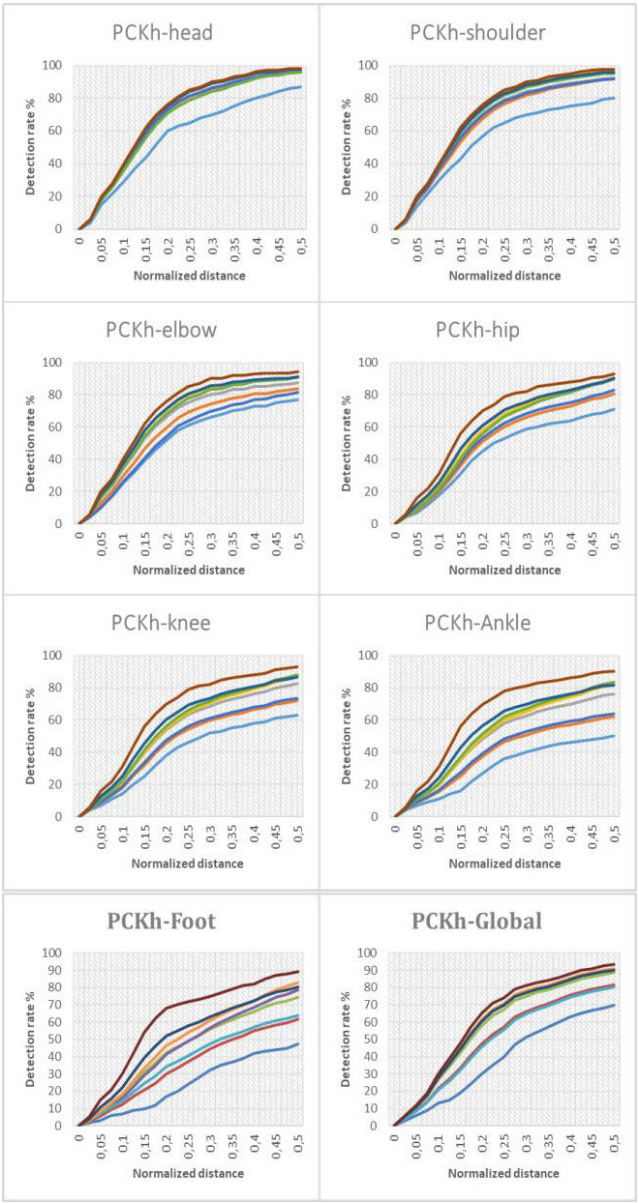


FIGURE 2. System performance comparison.

TABLE 3. Computational cost.

MODEL	Output	Input image	GFLOPs
DeepPose (ResNet - 101) [40]	Regression	(3, 192, 256)	7.69
ConvNet Pose [45]	Heat-map	(3, 192, 256)	28.56
CPM [34]	Heat-map	(3, 288, 384)	63.57
Stacked hourglass (3 Stages) [35]	Heat-map	(3, 384, 384)	64.5
HPE IF [46]	Heat-map	(3, 192, 256)	36.58
Cascade (ResNet-101 Cascaded with 2 ResNet-50) [36]	Heat-map	(3, 288, 384)	61.3
TFPose (Resnet - 50; Nd=6) [47]	Regression	(3, 288, 384)	20.4
ViTPose (ViTAE-G) [49]	Heat-map	(3, 432, 576)	76.59

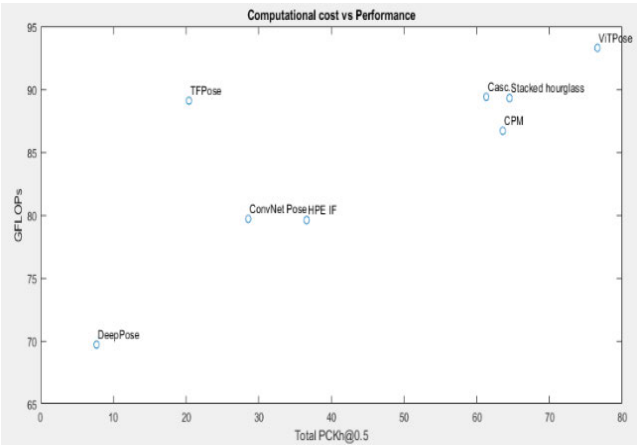


FIGURE 3. Total PCKh@0.5 (Global performance) vs computational cost.

second) vs computational cost (GFLOP's) for the different considered networks.

Figure 5 illustrates the ground truth heat-maps of a FIR image and the predictions made by the different architectures.

As expected, and in line with the results obtained in the different papers that present each considered system, table 2 shows a significant difference in performances between the networks that directly regress joint coordinates and the ones which output heat-maps.

Although the introduction of transformers in the field of artificial vision is very recent, and the number of models applied to human pose estimation is still limited, the models based on transformers used in this work offer better performances than the ones based on the classic use of convolutional neural networks.

All models demonstrate exceptional performance at identifying the head, as can be easily inferred from figure 2, where the obtained PCKh is quite similar for most of them. A similar result is observed in the case of the shoulders, the closest joint to the head. However, as the joints get farther from the head, the model's ability to determine joint position degrades significantly, especially in the case of the ankle and the foot. Additionally, the performances of the different systems, which are similar for the less challenging key-points, vary widely for the most challenging ones.

The computational cost of models based on transformers is lower than that of the systems based on neural networks for the same input resolution, and, with slight divergences, better image input resolutions lead to better outcomes, especially for the most challenging joints, albeit at a higher computational cost.

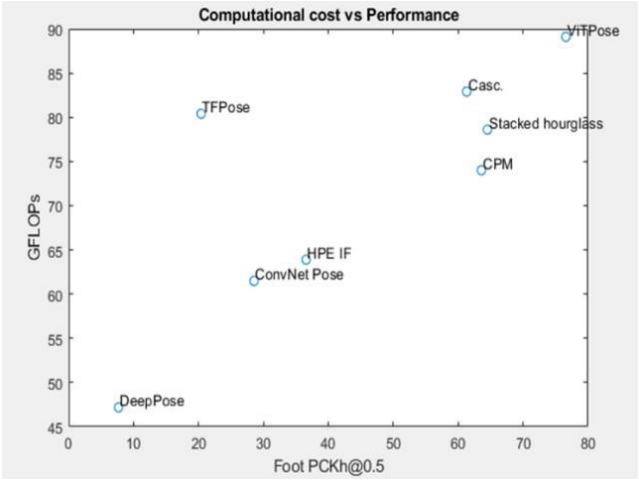


FIGURE 4. Foot PCKh@0.5 (Network performance to place the most challenging joint) vs computational cost.

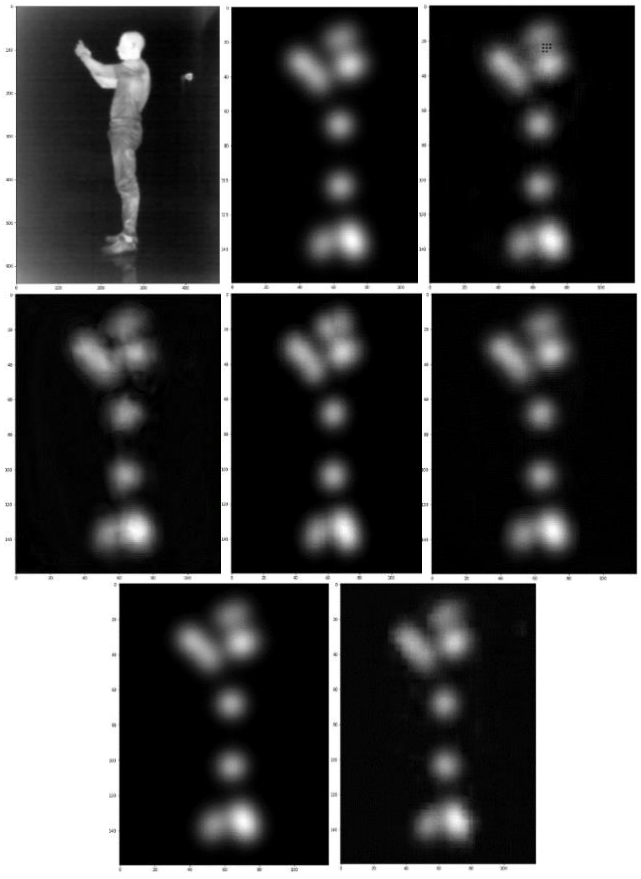


FIGURE 5. *(a) Base image, (b) Ground truth heat-map, (c) ConvNet Pose prediction, (d) CPM prediction, (e) Stacked hourglass prediction, (f) HPE IF prediction, (g) Cascade prediction, (h) ViTPose prediction.*

Finally, the computational power required to run these networks in the limits of realistic motion perception (24 frames per second), places the borders of the needed processing power between 184.56 GFLOP's/sec (7.69 GFLOP's per frame according to table 3 × 24 frames per second) for the case of DeepPose and 1838.16 GFLOP's/sec (76.59 × 24) for ViTPose. These figures are out of boundaries for some processors, especially the ones mounted on mobile devices (table 4), which are often less capable than the hardware used by desktop computers. Furthermore, these requirements will be increased when other threads need to be run in parallel.

TABLE 4. Processing power of chipsets mounted on modern mobile devices.

Chipset	Processing power (GFLOPS)
A13 Bionic	786
Exynos 2100	1530
Snapdragon 888	1720
Google Tensor	2171

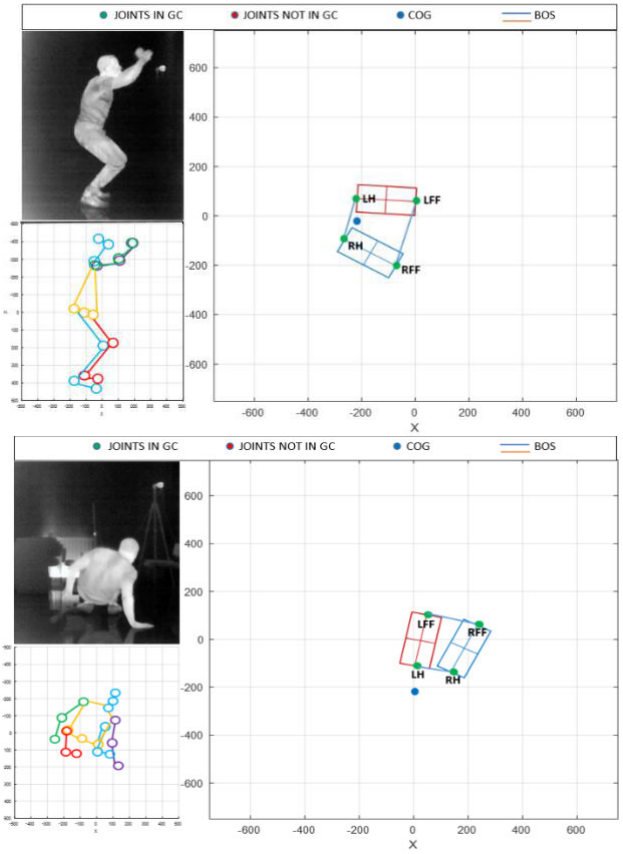


FIGURE 6. Presentations of (a) volunteer playing basketball (b) volunteer falling backwards from the FIR-Human dataset. Joints position, Center of Gravity (COG) and Base of Support (BoS) are presented.

B. FALL DETECTION SYSTEM EVALUATION

Figure 6 illustrates the two-dimensional network's outcome once it has been processed through the fall detection system proposed in [4].

This system takes the outcome of two-dimensional pose estimation networks and evaluates, by analyzing the relative

TABLE 5. System accuracy comparison on FIR-Human dataset by type of 2D network employed.

Model	SE	SP	AC
DeepPose (ResNet - 101) [40]	62.50%	61.54%	61.80%
ConvNet Pose [45]	72.22%	70.77%	71.16%
CPM [34]	77.78%	78.97%	78.65%
Stacked hourglass (3 Stages) [35]	84.72%	81.03%	82.02%
HPE IF [46]	75.00%	72.31%	73.03%
Cascade (ResNet-101 Cascaded with 2 ResNet-50) [36]	83.33%	87.69%	86.52%
TFPose (Resnet -50; Nd=6) [47]	80.56%	84.10%	83.15%
ViTPose (ViTAE-G) [49]	95.83%	96.92%	96.63%

movement of the major human joints, whether the person keeps balance or falls down. The great advantage of this concept is that it can be trained using RGB video-clips and, therefore, an extensive dataset of falls recorded using FIR video cameras is not necessary. Additionally, this system approaches fall in terms of body balance and stability making, this way, irrelevant the differences between simulated and real falls, as all falls are a direct result of fails in the continuous effort of the body to keep balance, regardless of other considerations.

Table 5 presents the performance evaluation indexes of the system which integrates the outcome of the considered two-dimensional pose estimation networks presented in 3.1 with the fall detection system described in [4]. Results vary widely depending on the performances of the used 2D network. Additionally, table 6 reflects the confusion matrices for all the considered cases.

Although comparison with other fall detection systems working on FIR imagery is not possible as, to the best of our knowledge, there is no other FIR image-based fall detection system, the results obtained by the most performant network shown in table 5 are in the ranges of specificity, sensitivity and accuracy of the RGB fall detection systems considered in [4] and presented in tables 7 and 8 and, in some cases, over exceed them.

Although comparison with other fall detection systems working on FIR imagery is not possible as, to the best of our knowledge, there is no other FIR image-based fall detection system, the results obtained by the most performant network shown in table 5 are in the ranges of specificity, sensitivity and accuracy of the RGB fall detection systems considered in [4] and presented in tables 7 and 8.

Our system's overall fall detection performance is especially dependent on the precision of the network to position the lower joints. This way, the most performant networks determining the position of the lower joints yield the best results, as tables 2 and 5 clearly prove. These results are

TABLE 6. Confusion matrixes.

Model	Real	Forecasted	
		Fall	Not a fall
DeepPose (ResNet - 101) [40]	Fall	45	27
	Not a fall	75	120
ConvNet Pose [45]	Fall	52	20
	Not a fall	57	138
CPM [34]	Fall	56	16
	Not a fall	41	154
Stacked hourglass (3 Stages) [35]	Fall	61	11
	Not a fall	37	158
HPE IF [46]	Fall	54	18
	Not a fall	54	141
Cascade (ResNet-101 Cascaded with 2 ResNet-50) [36]	Fall	60	12
	Not a fall	24	171
TFPose (Resnet -50; Nd=6) [47]	Fall	58	14
	Not a fall	31	164
ViTPose (ViTAE-G) [49]	Fall	69	3
	Not a fall	6	189

TABLE 7. Accuracy comparison of different methods on RGB UR fall dataset.

Method	Data	AC
Qingzhen Xu et al. [74]	RGB UR fall dataset	91.7%
X. Cai et al. [75]		96.2%
X. Wang et al. [76]		97%
S. Kalita et al. [77]		94.28%
C. Menacho et al. [78]		88.55%
D. Kumar et al. [79]		98.1%
S. Kasturi et al. [80]		96.34%

TABLE 8. Sensitivity and Specificity comparison of different methods on UR fall dataset.

Method	Data	Sensitivity	Specificity
B. Dai et al. [81]	RGB UR fall dataset	95%	96.7%
P. Soni et al. [82]		98.15%	97.1%
S. Kalita et al. [77]		93.33%	95%
A Carlier et al. [83]		95.5%	93.2%
Q. Feng et al. [84]		91.4%	-
S Bhandari et al. [85]		96.67%	95%

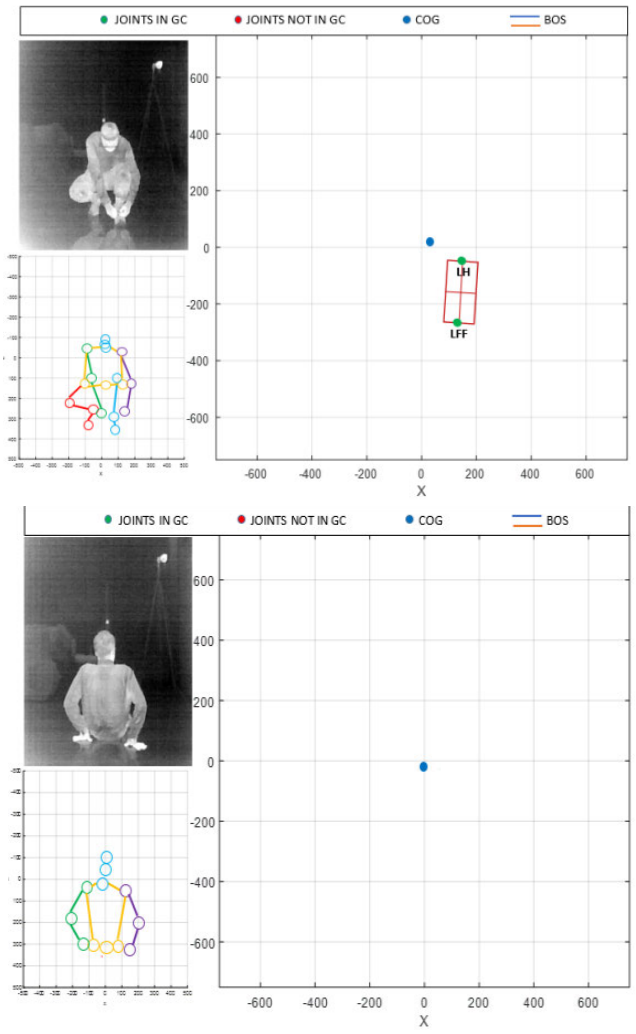


FIGURE 7. Presentations of system errors (a) false fall positive as a consequence of fading of the right foot using Deep-Pose (b) false fall negative as a result of the impossibility to determine the base of support due to occlusion using CPM. Joints position, Center of Gravity (COG) and Base of Support (BoS) are presented.

well aligned with the findings shown in table 10 of the paper presenting CoGNet [4]. This table illustrates the diminishing ability of CoGNet to properly determine the base of support (BoS) and center of gravity (COG) of a person as the number of joints passed to the network diminishes. The effect is particularly noticeable when the lower joints' positions are not accurately provided, as these positions are the key ones to determine whether the person will keep his balance and, therefore, imprecisions in them will sharply diminish the overall performance of the fall detection system.

The fading of the lower joints is a direct consequence of their distance from the head, which is particularly noticeable in the first networks of table 2. As the networks become more performant, they yield better results as we can observe in the lower rows of the table.

Moreover, although neural networks used in the field of artificial vision are more resilient to errors induced by noise due to occlusion than previous pose estimation systems they are still sensitive to this phenomenon [62].

The combined effect of these phenomena leads to inaccurate joint coordinates being passed to CoGNet, especially in the case of the most distant joints from the head; such as knees, ankles, and feet. These inaccuracies result in incorrect BoS definitions, causing the system to misjudge situations where the COG approaches the real limits of the BoS. However, these effects tend to diminish as the network's outcomes affecting lower joints become more precise.

This way, after a careful review of the false negatives presented in table 6, all of them are related to falls recorded from angles that induce occlusion effects during the last phases of the fall in the case of the most performant 2D networks, while in the case of less performant networks false negatives are also linked to fading of lower joints phenomena. In either case CoGNet is unable to properly place the COG in relation to the BoS. Figure 7.b illustrates an example of false fall negative as a consequence of occlusion phenomena.

On the other hand, all false positives are linked to situations where the movements place the COG very close to the limits of the BoS but still within it. As in the previous case, misjudgments by CoGNet regarding the COG's position in relation to the BoS are the result of occlusion phenomena in the case of the most performant 2D networks, while the less performant ones are also sensible to fading phenomena. Figure 7.a is an example of a false fall positive as a result of a fading phenomenon which affects the right foot of a volunteer while is bent over trying his shoe.

V. CONCLUSION AND FUTURE WORK

The study carried out in [63] identifies a set of situations associated with semi-supervised patients in which human supervision is not provided, and in which these patients would accept the use of automatic fall detection systems. One of these situations is when these patients are sleeping. Under these conditions, the use of a fall detection system based on FIR (Far Infrared) could be the most suitable solution.

This work implements the first system of this kind by integrating the output of a neural network for position estimation with the fall detection methodology proposed in [4]. In this way, the system uses the relative positions of the joints to determine whether the person is maintaining balance or experiencing a fall. To do this, these positions are used as input for a neural network called CoGNet, which calculates the position of the center of gravity projected on the ground and the body's support base. Finally, a fall detection algorithm uses all this information to determine if a fall has occurred.

In order to evaluate the system's capabilities based on the output of the neural network for position estimation, a wide range of these networks has been selected, which, according to the criteria established in [43], represent the essential architectures developed in the field of human position estimation. These networks have been trained using FIR-Human, the first large publicly available FIR database in which the

fundamental joints of volunteers are labeled in both two and three dimensions.

The performances of these networks are in line with those obtained with RGB databases. These performances show that architectures based on Transformers yield better results than those using convolutions for the same computational demand.

In general terms, as table 2 and figure 2 present, direct regression offers worse performances than heat-map generation techniques, and regardless of the approach, joints closer to the head are less challenging for all models compared to those that are further away. Moreover, while the system performances for the easier key-points are similar, they vary widely for the most difficult ones.

Finally, the results clearly show that higher resolution images allow the models to generate better results at the cost of increased computational demands.

As for the overall performance of the fall detection system, it depends primarily on the accuracy with which the two-dimensional network establishes the position of the main joints, especially those closer to the feet, as illustrated in table 5.

Although it is not possible to compare this system with others, as it is the first of its kind, the accuracy indexes presented in table 4 for the most performant networks fall within the range offered by classic systems that work with RGB images, as can be deduced from the comparison of table 4 with tables 7 and 8. This demonstrates that the proposed automatic fall detection system, which operates on FIR imagery, is a valid solution for working in poorly or non-illuminated environments. Additionally, the use of FIR images contributes to privacy protection, addressing concerns raised by different communities related to the elderly care sector.

The study of both false positives and false negatives in Table 6 allows us to deduce that inaccuracies in joint determination, which in turn are the source of failures in the fall detection system, are primarily the result of two phenomena: occlusion and fading of joints distant from the head. In both situations, CoGNet experiences limitations in calculating the position of the Center of Gravity (COG) in relation to the Base of Support (BoS). In the case of more capable 2D networks, errors in fall determination are primarily the result of occlusion phenomena, while in the case of less capable networks, fading of joints further from the head compounds these phenomena.

On the other hand, the amount of processing power required to run the networks considered in this work within the constraints of realistic motion perception exceeds the capabilities of some processors, especially when parallel threads need to be executed, as evident from Tables 2 and 3.

Finally, all the network architectures considered in this study were originally designed to operate with RGB imagery, which means they are designed to accept three-channel inputs, one for each primary color. However, FIR imagery represents temperature using a two-color palette, typically assigning one color, usually white, to the highest temperature detected in the image and the other color, typically black, to the lowest temperature. Therefore, potential redesigns of these architectures to work with two-channel inputs may not only improve their performance but also reduce the computational requirements needed to run them. Future research in this direction may significantly enhance both the performance of pose estimation networks when working with FIR imagery and the overall performance of the fall detection system while reducing the processing power required to operate it.

REFERENCES

[1] *World Population Ageing 2017: Highlights*, Department of Economic and Social Affairs, United Nations, New York, NY, USA, 2017.
[2] D. A. Sterling, J. A. O'Connor, and J. Bonadies, "Geriatric falls: Injury severity is high and disproportionate to mechanism," *J. Trauma, Injury, Infection, Crit. Care*, vol. 50, no. 1, pp. 116–119, Jan. 2001.
[3] J. Gutiérrez, V. Rodriguez, and S. Martin, "Fall detection system based on far infrared images," in *Proc. Congreso de Tecnología, Aprendizaje y Enseñanza de la Electrónica*, Jun. 2022, pp. 1–7.
[4] J. Gutierrez, S. Martin, and V. Rodriguez, "Human stability assessment and fall detection based on dynamic descriptors," *IET Image Process.*, vol. 17, no. 11, pp. 3177–3195, 2023.
[5] W. Gong, X. Zhang, J. Gonzàlez, A. Sobral, T. Bouwmans, C. Tu, and E.-H. Zahzah, "Human pose estimation from monocular images: A comprehensive survey," *Sensors*, vol. 16, no. 12, p. 1966, Nov. 2016.
[6] W. Zhang, L. Shang, and A. B. Chan, "A robust likelihood function for 3D human pose tracking," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5374–5389, Dec. 2014.
[7] U. Güdükbay, I. Demir, and Y. Dedeoglu, "Motion capture and human pose reconstruction from a single-view video sequence," *Digit. Signal Process.*, vol. 23, no. 5, pp. 1441–1450, Sep. 2013.
[8] G. Pons-Moll and B. Rosenhahn, "Model-based pose estimation," in *Visual Analysis of Humans: Looking at People*. London, U.K.: Springer, 2011, pp. 139–170.
[9] V. Parameswaran and R. Chellappa, "View independent human body pose estimation from a single perspective image," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2004, pp. 1–7.
[10] G. Pons-Moll, L. Leal-Taixe, T. Truong, and B. Rosenhahn, "Efficient and robust shape matching for model based human motion capture," in *Proc. Joint Pattern Recognit. Symp.*, 2011, pp. 416–425.
[11] V. Ganapathi, C. Plagemann, D. Koller, and S. Thrun, "Real-time human pose tracking from range data," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, 2012, pp. 738–751.
[12] M. A. Brubaker, D. J. Fleet, and A. Hertzmann, "Physics-based person tracking using the anthropomorphic Walker," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 140–155, Mar. 2010.
[13] J. Gall, B. Rosenhahn, T. Brox, and H.-P. Seidel, "Optimization and filtering for human motion capture: A multi-layer framework," *Int. J. Comput. Vis.*, vol. 87, nos. 1–2, pp. 75–92, Mar. 2010.
[14] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *Int. J. Comput. Vis.*, vol. 61, no. 2, pp. 185–205, Feb. 2005.
[15] R. Okada and S. Soatto, "Relevant feature selection for human pose estimation and localization in cluttered images," in *Proc. ECCV*, vol. 2, 2008, pp. 434–445.
[16] R. Ronfard, C. Schmid, and B. Triggs, "Learning to parse pictures of people," in *Proc. Eur. Conf. Comput. Vis.*, Copenhagen, Denmark, 2002, pp. 700–714.
[17] W. Zhang, J. Shen, G. Liu, and Y. Yu, "A latent clothing attribute approach for human pose estimation," in *Proc. Asian Conf. Comput. Vis.*, 2015, pp. 146–161.
[18] S. Sedai, M. Bennamoun, and D. Q. Huynh, "Evaluating shape and appearance descriptors for 3D human pose estimation," in *Proc. 6th IEEE Conf. Ind. Electron. Appl.*, Jun. 2011, pp. 293–298.
[19] S. Li and A. B. Chan, "3D human pose estimation from monocular images with deep convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, 2015, pp. 332–347.

[20] H. Ning, W. Xu, Y. Gong, and T. Huang, "Discriminative learning of visual words for 3D human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[21] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," *Neural Comput.*, vol. 6, no. 2, pp. 181–214, Mar. 1994.

[22] O. Freifeld and M. J. Black, "Lie bodies: A manifold representation of 3D human shape," in *Proc. ECCV*, vol. 7572, 2012, pp. 1–14.

[23] A. Baak et al., "A data-driven approach for real-time full body pose reconstruction from a depth camera," in *Consumer Depth Cameras for Computer Vision: Research Topics and Applications*. London, U.K.: Springer, 2013, pp. 71–98.

[24] C. M. Christoudias and T. Darrell, "On modelling nonlinear shape-and-texture appearance manifolds," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 1067–1074.

[25] J. Gall, A. Yao, and L. Van Gool, "2D action recognition serves 3D human pose estimation," in *Proc. ECCV*, vol. 3, 2010, pp. 425–438.

[26] V. I. Morariu and O. I. Camps, "Modeling correspondences for multi-camera tracking using nonlinear manifold learning and target dynamics," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 545–552.

[27] G. Mori, C. Pantofaru, N. Kothari, T. Leung, G. Toderici, A. Toshev, and W. Yang, "Pose embeddings: A deep architecture for learning to match human poses," 2015, *arXiv:1507.00302*.

[28] A. Gupta, T. Chen, F. Chen, D. Kimber, and L. S. Davis, "Context and observation driven latent variable model for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[29] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[30] L. Sun, M. Song, D. Tao, J. Bu, and C. Chen, "Motionlet LLC coding for discriminative human pose estimation," *Multimedia Tools Appl.*, vol. 73, no. 1, pp. 327–344, Nov. 2014.

[31] V. Belagiannis, C. Amann, N. Navab, and S. Ilic, "Holistic human pose estimation with regression forests," in *Proc. Int. Conf. Articulated Motion Deformable Objects*, Palma De Mallorca, Spain, 2014, pp. 20–30.

[32] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[33] J. Y. Chang and S. W. Nam, "Fast random-forest-based human pose estimation using a multi-scale and cascade approach," *ETRI J.*, vol. 35, no. 6, pp. 949–959, Dec. 2013.

[34] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.

[35] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. 14th Eur. Conf. Comput. Vis. (ECCV)*, Amsterdam, The Netherlands. London, U.K.: Springer, Oct. 2016.

[36] Z. Su, M. Ye, G. Zhang, L. Dai, and J. Sheng, "Cascade feature aggregation for human pose estimation," 2019, *arXiv:1902.07837*.

[37] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1302–1310.

[38] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "DeeperCut: A deeper, stronger, and faster multi-person pose estimation model," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, 2016, pp. 34–50.

[39] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.

[40] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1653–1660.

[41] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[42] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM Comput. Surv.*, vol. 54, no. 10, pp. 1–41, Jan. 2022.

[43] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, "The progress of human pose estimation: A survey and taxonomy of models applied in 2D human pose estimation," *IEEE Access*, vol. 8, pp. 133330–133348, 2020, doi: 10.1109/ACCESS.2020.3010248.

[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1–9.

[45] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, and C. Bregler, "Efficient object localization using convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 648–656.

[46] J. Carreira, P. Agrawal, K. Fragkiadaki, and J. Malik, "Human pose estimation with iterative error feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4733–4742.

[47] W. Mao, Y. Ge, C. Shen, Z. Tian, X. Wang, and Z. Wang, "TFPose: Direct human pose estimation with transformers," 2021, *arXiv:2103.15320*.

[48] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[49] Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "ViTPose: Simple vision transformer baselines for human pose estimation," 2022, *arXiv:2204.12484*.

[50] Y. Zang, C. Fan, Z. Zheng, and D. Yang, "Pose estimation at night in infrared images using a lightweight multi-stage attention network," *Signal, Image Video Process.*, vol. 15, no. 8, pp. 1757–1765, Nov. 2021.

[51] I.-C. Chen, C.-J. Wang, C.-K. Wen, and S.-J. Tzou, "Multi-person pose estimation using thermal images," *IEEE Access*, vol. 8, pp. 174964–174971, 2020, doi: 10.1109/ACCESS.2020.3025413.

[52] B. Sapp and B. Taskar, "MODEC: Multimodal decomposable models for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3674–3681.

[53] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," in *Proc. BMVC*, 2010, p. 5.

[54] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2D human pose estimation: New benchmark and state of the art analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014.

[55] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, Zurich, Switzerland, 2014, pp. 740–755.

[56] 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.

[57] D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko, W. Xu, and C. Theobalt, "Monocular 3D human pose estimation in the wild using improved CNN supervision," in *Proc. Int. Conf. 3D Vis. (3DV)*, Oct. 2017, pp. 506–516.

[58] 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.

[59] M. Habermann, W. Xu, M. Zollhöfer, G. Pons-Moll, and C. Theobalt, "DeepCap: Monocular human performance capture using weak supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5051–5062.

[60] N. T. Newaz and E. Hanada, "The methods of fall detection: A literature review," *Sensors*, vol. 23, no. 11, p. 5212, May 2023.

[61] P. Vallabh and R. Malekian, "Fall detection monitoring systems: A comprehensive review," *J. Ambient Intell. Humanized Comput.*, vol. 9, no. 6, pp. 1809–1833, Nov. 2018.

[62] J. Gutiérrez, V. Rodríguez, and S. Martin, "Comprehensive review of vision-based fall detection systems," *Sensors*, vol. 21, no. 3, p. 947, Feb. 2021.

[63] R. Elagovan, T. Perumal, and S. Krishnan, "Fall detection systems at night," *Computer*, vol. 56, no. 6, pp. 44–51, Jun. 2023, doi: 10.1109/MC.2022.3200404.

[64] M. Kangas, I. Vikman, L. Nyberg, R. Korpelainen, J. Lindblom, and T. Jämsä, "Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects," *Gait Posture*, vol. 35, no. 3, pp. 500–505, Mar. 2012.

[65] J. Klenk, C. Becker, F. Lieken, S. Nicolai, W. Maetzler, W. Alt, W. Zijlstra, J. M. Hausdorff, R. C. van Lummel, L. Chiari, and U. Lindemann, "Comparison of acceleration signals of simulated and real-world backward falls," *Med. Eng. Phys.*, vol. 33, no. 3, pp. 368–373, Apr. 2011.

[66] H. Tao, L. Cheng, J. Qiu, and V. Stojanovic, "Few shot cross equipment fault diagnosis method based on parameter optimization and feature mertic," *Meas. Sci. Technol.*, vol. 33, no. 11, Nov. 2022, Art. no. 115005.

[67] X. Song, P. Sun, S. Song, and V. Stojanovic, "Event-driven NN adaptive fixed-time control for nonlinear systems with guaranteed performance," *J. Franklin Inst.*, vol. 359, no. 9, pp. 4138–4159, Jun. 2022.

[68] Z. Zhuang, H. Tao, Y. Chen, V. Stojanovic, and W. Paszke, "Iterative learning control for repetitive tasks with randomly varying trial lengths using successive projection," *Int. J. Adapt. Control Signal Process.*, vol. 36, no. 5, pp. 1196–1215, May 2022.

[69] G. Ben-Sadoun, E. Michel, C. Annweiler, and G. Sacco, "Human fall detection using passive infrared sensors with low resolution: A systematic review," *Clin. Interventions Aging*, vol. 17, pp. 35–53, Jan. 2022.

[70] H. Lee, K. J. Bein, R. Ivers, and M. M. Dinh, "Changing patterns of injury associated with low-energy falls in the elderly: A 10-year analysis at an Australian major trauma centre," *ANZ J. Surg.*, vol. 85, no. 4, pp. 230–234, Apr. 2015.

[71] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, Oct. 2021.

[72] X. Liu, P. Li, D. Ni, Y. Wang, and H. Xue, "LightPose: A lightweight and efficient model with transformer for human pose estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2674–2678.

[73] G. Ning, Z. Zhang, and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1246–1259, May 2018.

[74] Q. Xu, G. Huang, M. Yu, and Y. Guo, "Fall prediction based on key points of human bones," *Phys. A, Stat. Mech. Appl.*, vol. 540, Feb. 2020, Art. no. 123205, doi: 10.1016/j.physa.2019.123205. [Online]. Available: http://www.sciencedirect.com.ezproxy.uned.es/science/article/pii/S0378437119318011

[75] X. Cai, S. Li, X. Liu, and G. Han, "Vision-based fall detection with multi-task hourglass convolutional auto-encoder," *IEEE Access*, vol. 8, pp. 44493–44502, 2020, doi: 10.1109/ACCESS.2020.2978249.

[76] X. Wang and K. Jia, "Human fall detection algorithm based on YOLOv3," in *Proc. IEEE 5th Int. Conf. Image, Vis. Comput. (ICIVC)*, Jul. 2020, pp. 50–54, doi: 10.1109/ICIVC50857.2020.9177447.

[77] S. Kalita, A. Karmakar, and S. M. Hazarika, "Human fall detection during activities of daily living using extended CORE9," in *Proc. 2nd Int. Conf. Adv. Comput. Commun. Paradigms (ICACCP)*, Feb. 2019, pp. 1–6, doi: 10.1109/ICACCP.2019.8882928.

[78] C. Menacho and J. Ordoñez, "Fall detection based on CNN models implemented on a mobile robot," in *Proc. 17th Int. Conf. Ubiquitous Robots (UR)*, Jun. 2020, pp. 284–289, doi: 10.1109/UR49135.2020.9144836.

[79] D. Kumar et al., "Elderly health monitoring system with fall detection using multi-feature based person tracking," in *ITU Kaleidoscope: ICT for Health: Networks, Standards and Innovation (ITU K)*. Atlanta, GA, USA: IEEE, 2019.

[80] Y. M. Galvão, V. A. Albuquerque, B. J. T. Fernandes, and M. J. S. Valença, "Anomaly detection in smart houses: Monitoring elderly daily behavior for fall detecting," in *Proc. IEEE Latin Amer. Conf. Comput. Intell. (LA-CCI)*, Nov. 2017, pp. 1–6, doi: 10.1109/LA-CCI.2017.8285701.

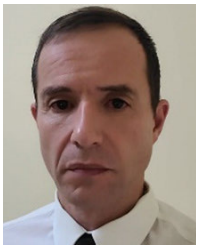
[81] B. Dai, D. Yang, L. Ai, and P. Zhang, "A novel video-surveillance-based algorithm of fall detection," in *Proc. 11th Int. Congr. Image Signal Process., BioMedical Eng. Inform. (CISP-BMEI)*, 2018, pp. 1–6, doi: 10.1109/CISP-BMEI.2018.8633160.

[82] P. K. Soni and A. Choudhary, "Automated fall detection from a camera using support vector machine," in *Proc. 2nd Int. Conf. Adv. Comput. Commun. Paradigms (ICACCP)*, Feb. 2019, pp. 1–6, doi: 10.1109/ICACCP.2019.8882966.

[83] A. Carlier, P. Peyramaure, K. Favre, and M. Pressigout, "Fall detector adapted to nursing home needs through an optical-flow based CNN," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 5741–5744, doi: 10.1109/EMBC44109.2020.9175844.

[84] Q. Feng, C. Gao, L. Wang, Y. Zhao, T. Song, and Q. Li, "Spatio-temporal fall event detection in complex scenes using attention guided LSTM," *Pattern Recognit. Lett.*, vol. 130, pp. 242–249, Feb. 2020, doi: 10.1016/j.patrec.2018.08.031. [Online]. Available: http://www.sciencedirect.com.ezproxy.uned.es/science/article/pii/S016786551830504X

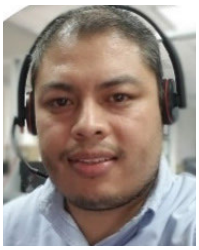
[85] S. Bhandari, N. Babar, P. Gupta, N. Shah, and S. Pujari, "A novel approach for fall detection in home environment," in *Proc. IEEE 6th Global Conf. Consum. Electron. (GCCE)*, Oct. 2017, pp. 1–5.



JESÚS GUTIÉRREZ received the B.S. degree in industrial engineering and the M.S. (by Research) degree in electrical, electronic and industrial control engineering from Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain, in 2012 and 2019, respectively, where he is currently pursuing the Ph.D. degree in industrial engineering technologies.

His professional career has developed in the Air Force, since 1998, up to today in the areas of air operations, logistic support, and future systems programs. His research interest includes different areas within the field of elderly care tasks automation.

SERGIO MARTIN (Senior Member, IEEE) is currently pursuing the Ph.D. degree with the Electrical and Computer Engineering Department, Industrial Engineering School, National University for Distance Education (UNED), Spain. He is also an Associate Professor with UNED. He is also a Computer Engineer in distributed applications and systems with the Carlos III University of Madrid. He teaches subjects related to microelectronics and digital electronics with the Industrial Engineering School, UNED, since 2007. Since 2002, he has been participating in national and international research projects related to mobile devices, ambient intelligence, location-based technologies and in projects related to "e-learning," virtual and remote laboratories, and new technologies applied to distance education. He has published more than 200 papers both in international journals and conferences.

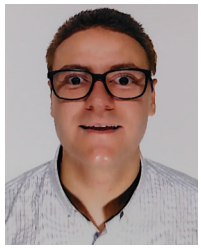


VICTOR H. RODRIGUEZ received the B.S. and M.S. degrees in electronic engineering from Instituto Tecnológico de Durango, Mexico, in 2011 and 2014, respectively, and the Ph.D. degree in electronic engineering from the EduQTech Research Group, University of Zaragoza, Spain, in 2019. He is currently working in the E-mobility field in the automotive industry with Preh Inc., MI, USA, and collaborating with EduQTech in research activities. His current interest includes fall detection using wearable sensors.

SERGIO ALBIOL received the Graduate, Diploma of Advanced Studies (D.E.A.), and Ph.D. degrees in computer science from Universitat Politécnica of Valencia, in 1999, 2010, and 2014, respectively. He is also an Assistant Professor with the Department of Computer Science and Systems Engineering, Universidad de Zaragoza, Teruel, Spain, where he has been teaching with the Computer Science and Systems Engineering Department, since 2001. He has organized workshop sessions regarding virtual rehabilitation theories and applications. His research interests include patients with serious injuries and illnesses by using virtual rehabilitation techniques in the area of virtual motor rehabilitation and systems based on interaction for the recovery of mental disorders.

INMACULADA PLAZA (Senior Member, IEEE) received the degree in physics, the D.E.A. degree in manufacturing engineering, and the Ph.D. degree from the Department of Electronic Engineering and Communications, University of Zaragoza, Spain. She is currently the Founder of the IEEE Education Society Spanish Chapter, from which she was the Vice-Chairperson (2008–2010) and the Chairperson (2010–2012). She is also with the University of Zaragoza as a Full Professor in electronic technology. She founded the research group EduQTech. She has been the Leader of 11 collaboration agreements and the main researcher of 27 projects (regional, national, and international) and the author of more than 100 works. She was the Dean of Escuela Universitaria Politécnica de Teruel, Spain. She is also the Director of the COGITIAR Chair. She also directs Instituto de Estudios Turolenses. She has received numerous awards in research and management, some of them at international levels, for instance: ''IEEE Education Society Chapter Achievement Award,'' 2011—Chairperson of the Chapter, and ''2011 Best Large Chapter Award—IEEE Region 8'' (Chairperson). She was awarded with the TAEE Association Award to the Professional Career (2016) and the ''Enterprising Woman 2018'' from the University of Zaragoza. Finally, in 2023, she received the ''IEEE WIE Inspiring Member of the Year Award.''

CARLOS MEDRANO (Senior Member, IEEE) received the joint Ph.D. degree in physics from the University of Zaragoza, Spain, and Joseph Fourier, Grenoble, France, in 1997, and the Ph.D. degree from the Department of Electrical, Electronic and Control Engineering, Spanish University for Distance Education, in 2015. He has been a Tenured Faculty Member with the Department of Electronic Engineering and Communications, University of Zaragoza, since 1998. He has a broad research career, including topics, such as magnetism, data acquisition and control systems, and computer vision. He is the author of more than 40 articles in peer-reviewed journals. His current interests include ambient and wearable sensors and pattern recognition for health and wellbeing applications.



JAVIER MARTINEZ received the B.S. degree in electronic and automatic engineering and the M.S. degree in industrial engineering from the University of Zaragoza, Spain, in 2016 and 2019, respectively, and the Ph.D. degree in electronic engineering from the EduQTech Research Group, Department of Electronic Engineering and Communications, University of Zaragoza, in 2023. He is the author of more than 40 articles in peer-reviewed journals. His research interests include flexible pressure sensors and wearable electronic systems for health applications.