Fall detection with a non-intrusive and first-person vision approach

Wang, Xueyi; Talavera Martínez, Estefanía; Karastoyanova, Dimka; Azzopardi, George

Published in:
IEEE Sensors Journal

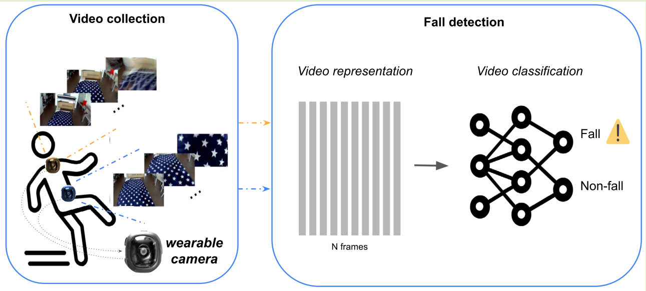
DOI:
10.1109/JSEN.2023.3314828

Publication date:
2023

Fall Detection With a Nonintrusive and First-Person Vision Approach

Xueyi Wang, Estefanía Talavera, Dimka Karastoyanova, *Member, IEEE*, and George Azzopardi

***Abstract*—Falls have been widely recognized as one of the most dangerous incidents for the elderly and other people with mobility limitations. This problem has attracted wide scientific interest, which has led to several investigations based on nonvision wearable sensors and static cameras. We investigate the challenge of fall detection and recognition using egocentric wearable cameras, which, besides portability and affordability, capture visual information that can be further leveraged for a broad set of lifelogging applications. In this work, five volunteers were equipped with two cameras each, one attached to the neck and the other to the waist. They were asked to simulate four kinds of falls and nine types of nonfalls. The newly collected dataset consists of 5858 short video clips, which we make available online. The proposed approach is a late fusion methodology that combines spatial and motion descriptors along with deep features extracted by a pretrained convolutional neural network. For the spatial and deep features, we consider the similarity of such features between frames in regular intervals of a given time window. In this way, it is the transition between such frames that are encoded in our approach, while the actual scene content does not play a role. We design several experiments to investigate the best camera location and performance for indoor and outdoor activities and employ leave-one-subject-out cross-validation to test the generalization ability of our approach. For the fall detection (i.e., two-class) problem, our approach achieves 91.8% accuracy, 93.6% sensitivity, and 89.2% specificity.**



***Index Terms*—Decision fusion, deep learning, fall detection, machine learning, wearable cameras.**

I. Introduction

FALLS have been widely considered to be one of the most dangerous and fatal accidents for elderly people due to less control over their body, weaker bones, longer time to recovery [1], and derived neurological diseases [2], among others. The burden of disease due to fall-related injuries in the adult population has increased by 54% since 1990 [3]. In 2017, around 14% of the adult population in Western countries sought medical treatment for fall-related injuries [3].

A fall can have several consequences. Primarily, it can cause severe damage to the elderly concerned, both physically and mentally, which may lead to chronic impairment or even fatality [3], [4]. Secondarily, the implications of a fall may have an effect on the care support system, which tends to include family members and healthcare professionals [5]. As a result, there will also be financial repercussions across the board, which are difficult to quantify. Even in cases of full physical recovery, the psychological trauma experienced by the affected individual may impact their lifestyle due to fear of falling, for example, less walking alone outside, which can cause other medical and psychological issues due to lack of mobility and socialization [3]. This motivates the need for our work where we develop a fall detection approach with a wearable camera. It contributes to providing more freedom of movement to the elderly and people with mobility impairments.

The overarching goal is to develop a full IoT-based system that includes wearable cameras connected to a Cloud infrastructure and in contact with primary care agencies. We envisage that such a system would have three main building blocks: 1) fall detection and recognition based on visual data; 2) alarms and management of data; and 3) validation of alarms and online learning. The scope of our work includes the first building block, namely the use case of fall detection, which we consider the core component. In this regard, we propose a methodology that relies on the analysis of the visual data gathered by a wearable camera. In practice, this component can be operated by a sliding window of a fixed length and interval. Although it may be reasonable to contemplate the

Manuscript received 9 June 2023; revised 2 September 2023; accepted 4 September 2023. Date of publication 19 September 2023; date of current version 14 November 2023. The associate editor coordinating the review of this article and approving it for publication was Dr. Xiaojin Zhao. *(Corresponding author: Xueyi Wang.)*

Xueyi Wang, Dimka Karastoyanova, and George Azzopardi are with the Faculty of Science and Engineering, Bernoulli Institute of Mathematics, Computer Science and Artificial Intelligence, Information System Group, University of Groningen, 9747 AG Groningen, The Netherlands (e-mail: xueyi.wang@rug.nl; d.karastoyanova@rug.nl; g.azzopardi@rug.nl).

Estefanía Talavera is with the Faculty of Electrical Engineering, Department of Mathematics and Computer Science (EEMCS), Data Management and Biometrics (DMB) Group, University of Twente, 7522 NB Enschede, The Netherlands (e-mail: e.talaveramartinez@utwente.nl).

Digital Object Identifier 10.1109/JSEN.2023.3314828

integration of multiple sensors to address the given problem, our deliberate choice is to concentrate solely on egocentric visual data. By doing so, we aim to establish a solid foundation and gain a comprehensive understanding of the capabilities and potential of this specific data type. The subsequent two components, which are beyond the scope of this work, would be triggered whenever our algorithm detects a fall. In that case, a signal will be sent to the corresponding primary care agency and the respective data (e.g., a video clip of the last 10 s and the location) will be stored on the Cloud. The verified falls would then be used to fine-tune the proposed fall detection models.

Wearable cameras provide an affordable and portable solution that can be used anywhere and anytime with a visible scene. This is in contrast to fixed surveillance cameras, which are considered intrusive and can only capture a predefined field of view [6]. Furthermore, such cameras provide a first-person perspective, capturing valuable contextual information and enabling a comprehensive understanding of the surroundings and detection of environmental factors contributing to falls. In fact, wearable cameras have already attracted versatile applications in various fields, including tasks such as activity recognition [7], segmentation [8], egocentric 3-D body pose estimation [9], sentiment recognition [10], and anticipation [11], showcasing their potential impact. Ongoing research and development efforts focus on algorithms, privacy concerns, and computational efficiency, reinforcing the consensus on wearable cameras as an innovative solution for fall detection. Additionally, a secondary benefit of the egocentric view is its potential for accurately determining the precise location, particularly when combined with GPS technology. This capability is especially useful in complex indoor buildings. The ability to pinpoint the location of an event can be vital for primary care personnel in effectively responding to and locating the incident. Notable is the fact that data collected by egocentric wearable cameras can be used beyond fall-related problems. For instance, it can be used to quantify the extent of socialization [12] which when analyzed over time may potentially be used to detect early signs of psychological problems, such as depression and cognitive decline [2]. As a matter of fact, wearable cameras have also been evaluated for the elderly in memory assistance [13], supporting independent living [14], and emergency response [15]. They serve as memory aids by recording daily activities, interactions, and reminders, aiding individuals with cognitive decline. These cameras also offer real-time assistance, such as navigation prompts and medication reminders, promoting independent living. In emergencies, wearable cameras notify designated contacts or emergency services, providing valuable live video feeds for faster and more accurate assistance.

The benefits of wearable cameras highlighted above motivate our decision to investigate a wearable egocentric camera-based solution.

Our contributions are threefold.

1) Due to the lack of publicly available datasets, we collect a new 13-class dataset, which we call RUG-EGO-FALL, and make it publicly available.[1] It is composed of 5858 video clips collected from five volunteers who were asked to simulate four kinds of falls and nine types of nonfalls. RUG-EGO-FALL is the largest dataset compared to 330 clips in [16] and 237 in [17] in the field of fall detection by egocentric cameras.
2) We propose a classification model that takes input from spatial and temporal features. The work includes a comparative analysis assessing the generalization ability based on deep and handcrafted features, together with different combinations of them with feature and decision fusion.
3) We investigate the effects of the camera location (neck or waist) and the scene location (indoor or outdoor) for the detection and recognition of the type of fall.

The rest of this article is organized as follows. We give an account of the state-of-art in Section II. In Section III, we describe our design for data collection and the gathered dataset. Then, we describe the proposed approach in Section IV. Section V depicts the experimental framework and reports the obtained results, which is then followed by a discussion in Section VI. Finally, we draw our conclusions in Section VII.

II. Literature Review

This section provides an overview of prior studies on fall detection, which is divided into two parts. First, we present a summary of the sensors examined in previous research. Second, we describe the major algorithms that have been proposed to date.

A. Sensors for Fall Detection

Abundant studies have explored different types of sensors and approaches to detect falls. The state-of-the-art review in [5] provides an overview of different sensors on which various research lines have relied for fall detection. The review divided sensors used in fall detection into four categories, namely fixed vision-based cameras [18], [19], [20], [21], wearable sensors including wearable cameras [22], [23], [24], [25], ambient based on radar and RGB-D sensors [26], [27], and sensor fusion [18], [19], [20], [21]. Below, we elaborate on the former two categories, which are the most relevant to our work.

1) Fixed Vision-Based: These are among the well-known sensors in this field. Previous studies have demonstrated the usage of vision-based methods, including infrared [28], RGB [29], and RGB-D depth cameras [30], [31], [32]. Because of the reliability in image framing during recording and the advantage of cable connection, a fixed RGB surveillance camera is the most widely spread visual sensor and has been investigated in many research studies. The study in [33] investigated a sensor network to collect data in a retirement residential home. Fixed cameras were deployed to monitor the daily life of the elderly. In other studies [18], [20], [21], stunts/volunteers were hired to simulate falls and nonfalls, which were captured by fixed cameras. The performance of systems that rely upon fixed RGB cameras is, however, limited to predefined locations (e.g., indoor only) and faces challenges related to occlusions. The performance suffers when the subject of interest is (partially) obstructed. Moreover,

[1] https://github.com/Xueyi-Wang/EGOFALLS/tree/main

because fixed cameras are restricted to specific locations, they cannot be used in portable systems.

2) Wearable Sensors: This category of sensors has the advantage of being portable, which allows the recording of data in both indoor and outdoor environments. Additionally, wearable sensors have been shown to provide a promising source for fall detection. They leverage the physiological variations of the human body by embedded sensors such as accelerometers [20], [34], gyroscopes, electrocardiography (ECG), electroencephalography (EEG), electromyography (EOG) [35], or a fusion of wearable sensors, such as the accelerometer and gyroscope [36], EOG and Plantar pressure [37], or the accelerometer, gyroscope, and magnetometer [38]. However, although these wearable nonvisual sensors are portable, they are unable to differentiate between the types of falls and to determine the precise location within an indoor environment, both of which are critical pieces of information for paramedics. As an alternative, wearable vision devices are a plausible technology for the detection, recognition, and localization of falls in the wild.

3) Wearable Cameras: In recent years, the egocentric vision has gained popularity and has been explored in various studies across different disciplines. This approach provides a unique and immersive viewpoint, capturing visual information from the wearer's perspective. It has been used to gain insights into daily activities and socialization patterns as we mentioned above. We consider fall detection as another important application for wearable cameras. Unlike fixed surveillance cameras, wearable cameras offer the advantage of being discreet and providing full user control. Users have the ability to turn off the camera at any time, ensuring privacy protection. Ongoing advancements are focused on developing lightweight and comfortable camera systems to enhance ergonomics. To the best of our knowledge, the research conducted by Casares et al. [16] is the first that uses wearable cameras for the detection of falls. They deployed a wearable smart camera attached to the waist with memory and a microprocessor. They recruited three participants who collected a small dataset of 330 videos with an equal number of falls, sitting, and lying down video events. They achieved 91% accuracy for the fall (binary) detection problem in terms of internal cross-validation.

The research in [17] involved a lab experiment where participants were equipped with a wired web camera attached to their waists. In particular, they used the Microsoft LifeCam camera with a microprocessor, 64-MB SDRAM, 16-MB NOR FLASH, and a wireless transmission module. The experiments were designed with three kinds of activities, namely falling, sitting, and lying down. In total, the three participants collected 237 clips, 150 of which were falls, 43 lying down, and 44 sitting down. The authors proposed to process data on the edge and only in the case of a detected fall, an alarm message must be relayed through a central server. They achieved an accuracy rate between 84% and 86% for the three-class problem. The Microsoft LifeCam camera was also investigated for another fall detection system [39]. A modified histogram of oriented gradients was employed as a feature descriptor and yielded 93.78% and 89.8% accuracy for indoor and outdoor fall detection, respectively.

The studies in [40] and [41] investigated the potential of the cameras in smartphones for the problem at hand. They also implemented a solution with a Raspberry PI that immediately processes the videos captured by the camera in real-time. Their experimental results showed that falls can be classified with a 95% sensitivity for binary fall classification. They compared two methods: one that involves only a histogram of gradient (HOG) features and the other a combination of HOG and optical flow to form their descriptor, with the latter approach yielding superior results. The smartphone, however, is not as practical as a wearable camera because of its intrusive character.

We postulate that the hardware solutions in the above-mentioned studies are neither convenient nor practical in real applications and most of them rely on nonportable approaches. In this study, we choose a small portable wearable camera to collect data in daily life in a nonintrusive manner, anywhere and anytime with a visible scene.

B. Classification Models for Fall Detection Using Wearable Cameras

1) Methods With Handcrafted Features and Learning-Free Models: The research by Casares et al. [16] conducted in a lab with wearable but wired cameras employed traditional HOG features for the detection of falls, which was inspired by the method proposed in [42]. In particular, it employed separate gradient orientation and strength histograms as the descriptor to detect anomalous activity, such as falls. The method involved the computation of the horizontal and vertical gradients of every pixel for each individual local region to calculate the gradient orientation and gradient strength. They achieved 91% accuracy with an 11.36% false-positive rate for binary fall classification.

Using the dissimilarity between consecutive frames in a video was the subject of investigation in [17] and [43]. They calculated such dissimilarity and applied a hierarchical classification model consisting of three layers with two thresholds. In the first layer, a threshold was applied to detect whether one of the three events of interest (falling, sitting, or lying down) had occurred. The second threshold was implemented to classify the event as fall or nonfall. The third layer was used to classify different kinds of nonfall activities by computing the average optical flow for horizontal and vertical directions over consecutive frames. They obtained a true positive rate of 82.69% for lying down, 86.79% for sitting down, 92.15% for falls from standing, and 78.84% for falls from sitting.

A fusion-based system that relies on a wearable camera and an accelerometer was investigated in [6]. The gradient local binary patterns (GLBPs) descriptor was used instead of the edge strength that was explored in their previous works [17], [43]. GLBP is computed by considering the eight neighboring pixels for each location. A value of 1 is assigned to a neighboring pixel if its intensity value is greater than the location under investigation, and 0 otherwise. They also applied the hierarchical classification model mentioned above but used the data from the accelerometer to classify a fall or nonfall by a predefined threshold. The study achieved a sensitivity of 96.36% and a specificity of 92.45% by the sensor fusion of the wearable camera and accelerometer on the same dataset in [43].

The work reported in [44] experimented with a combination of motion compensation, trajectory selection, and Fisher encoding to form a feature descriptor coupled with a support vector machine (SVM) classification model. Motion compensation, object features over foreground regions, and the use of an attention point to guide feature extraction turned out to be the key characteristics of their work. Nine public datasets, namely KTH, YouTube, Hollywood2, UCF sports, IXMAS, UIUC, Olympic Sports, UCF50, and HMDB51 were tested and their approach outperformed state-of-the-art approaches.

Ozcan et al. [39] extended their work by proposing an adaptive way to determine thresholds based on a relative-entropy-based procedure. Their algorithm relies on different Ali-Silvey distance measures. Using that approach and by means of fivefold cross-validation, the authors obtained a mean sensitivity of 93.77% and a mean specificity of 92.44% for the two-class problem of fall detection.

The Trajectory detector is another popular feature descriptor, which employs the motion information of trajectories, something that may also be relevant for the application of fall detection. It is obtained by either tracking techniques based on the Kanade-Lucas-Tomasi (KLT) tracker or using a scale-invariant feature transform (SIFT) descriptor between consecutive frames or a combination of both approaches [45]. This approach has demonstrated its effectiveness on nine public datasets of activity classification.

2) Methods With End-to-End Learning Models: Deep-learning-based approaches are notable for their ability to learn the most effective features coupled with a classification model without human intervention. This great benefit has led to a paradigm shift in developing solutions for vision-based applications. Activity classification has been widely addressed in the literature before relying on machine learning and deep learning. However, to our best knowledge, only one study of fall detection by wearable cameras has evaluated the deep-learning paradigm. In [46], a GoPro was mounted to the neck of three subjects to collect their own dataset in three different locations. A 2-D CNN unsupervised model was built to classify different activities, including squat down, stumble, stagger, fall down, and collision. They achieved an accuracy of 0.916 with internal cross-validation and 0.949 with leave-one subject-out cross-validation.

In this work, we evaluate the robustness of a fusion approach that relies on handcrafted (spatial and motion) and deep features for the detection and recognition of falls from videos gathered by wearable cameras.

C. Impact of Data: Young Versus Elderly; Real Versus Simulated

Previous research on fall detection has primarily relied on data collected from simulated falls and other activities performed by young subjects, as the rarity of genuine falls poses challenges for data collection [18], [19], [20], [21], [22], [23], [24]. While real falls are more diverse as highlighted in [47], simulated falls offer control, reproducibility, and ethical considerations, allowing researchers to assess system performance under consistent conditions and prioritize participant safety. In this sense, simulated falls serve as a benchmark for system validation. Limited studies have collected falls from elderly individuals, however, their datasets were not made publicly available due to privacy concerns [48], [49], [50], [51], [52]. In one study, Aziz et al. [48] collected ten genuine falls from 400 h of recording using triaxial accelerometers worn by both young adults and older individuals and found that while simulated data may not fully represent real falls, their trained model was able to reliably identify falls in real-world scenarios. Another study [51] collected 29 real falls during 21 000 h of monitoring via surveillance cameras in the homes of elderly individuals and determined that occlusion caused by movable furniture, doors, and curtains was a major difference between real-world and simulated data, resulting in a deterioration of the algorithm's performance. Notably, none of the above-mentioned studies utilized egocentric cameras, which have the potential to address issues of occlusion.

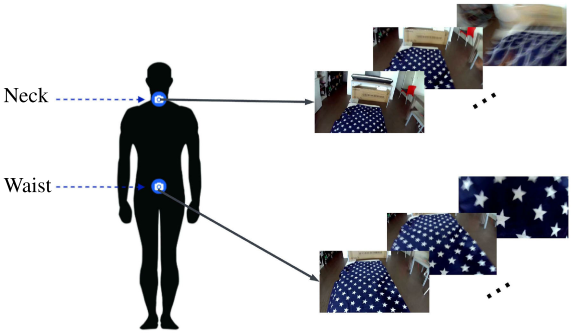


Fig. 1. Illustration of the two body locations (neck and waist) that are used in the data collection, with examples of frame sequences captured from neck and waist locations.

III. Materials

We started our study by collecting a dataset, which serves as one of the contributions of our work. We are not aware of other public datasets that can be used for such purposes. The dataset is composed of 5858 video clips and is the largest in comparison with other datasets collected from egocentric cameras; 330 clips in [16] and 237 in [17]. The dataset is designed to classify falls and nonfalls in daily activities, but also for the recognition of the actual type of fall or other human actions in daily life.

A. Equipment

The collection of events is recorded with OnReal G1 wearable cameras. OnReal G1 is a lightweight portable mini action camera, weighing 25 g and with small dimensions (420 × 420 × 200 mm), that captures videos with a resolution of up to 1080 pixels and 30 frames/s. The camera can be attached to different locations on the human body, such as the waist and neck, which we used for our data collection, as shown in Fig. 1. In practice, the cameras were clipped with the shirt collars and with the belts. Videos can be recorded with a time-lapse mode of different durations (1/5/10/30/60 s). We chose the 30-frames/s mode to collect our data.

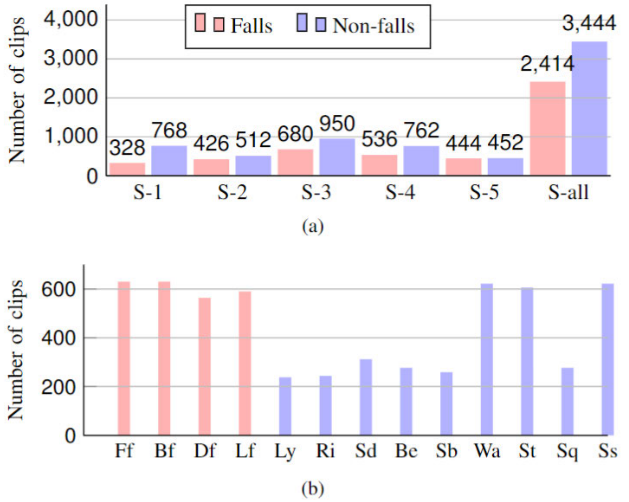


Fig. 2. Distributions of the number of clips (a) with respect to subjects (S-*X* indicates subject *X* and S-all indicates all subjects together), and (b) to the 13 activity types.

B. Subjects, Environment, and Activities

Five young healthy subjects (four males and one female, age: 29.8 ± 4.96, weight: 65.4 kg ± 7.50, height: 174.2 cm ± 7.63) were recruited to collect data for our study. They signed consent forms in which they agreed to use the gathered data in this work and to make it publicly available.

Both indoor and outdoor environments are considered in this study. Seven different indoor locations (two bedrooms, three living rooms, and two kitchens), and eight outdoor scenes (three parks, four streets, one campus) were used to make our data as diverse as possible. All volunteers were required to change their position and direction after the simulation of each activity to increase the diversity in the visual context.

The number of clips collected by the five volunteers is listed in Table I. They were asked to simulate four kinds of falls, namely front_falls (Ff), back_falls (Bf), downside_falls (Df), and lateral_falls (Lf), and nine types of nonfall activities, including lying (Ly), rising (Ri), sitting_down (Sd), bending (Be), stumbling (Sb), walking (Wa), standing (St), squatting (Sq), and sitting_static (Ss). Fig. 2 illustrates the distribution of the collected fall and nonfall activities.

C. Data Collection

Data are collected by switching ON and OFF the two cameras simultaneously for all activities. This is based on the guidelines given in [50], [52], [53], and [54], which indicate the duration (usually 1–3 s), location, and details of falls, how to simulate falls, and how to collect data of daily activities. The work in [54] proposed a common set of trials to test and compare different fall detection systems. They provided instructions on more than 20 types of falls, including different directions of falls (front, back, lateral, and downside). They also proposed a set of fall-like activities of daily living that can lead the system to output false positives. Those fall-related activities include Lying-bed, Rising-bed, Sit-bed, Sit-chair, Sit-sofa, Sit-air, Walking, Jogging, Bending, Bending-pick-up, Stumbling, Limp, Squatting-down, Trip-over, and Coughing-sneezing. Detailed instructions about how to simulate each activity are given in [54].

TABLE I
QUANTITY AND TYPE OF VIDEO CLIPS PER SUBJECT (S)

ID	All	Falls	Non-Falls	Indoor	Outdoor	Neck	Waist
S-1	1096	328	768	554	542	548	548
S-2	938	426	512	562	376	469	469
S-3	1630	680	950	812	818	815	815
S-4	1298	536	762	586	712	649	649
S-5	896	444	452	374	522	448	448
Total	**5858**	**2414**	**3444**	**2888**	**2970**	**2929**	**2929**

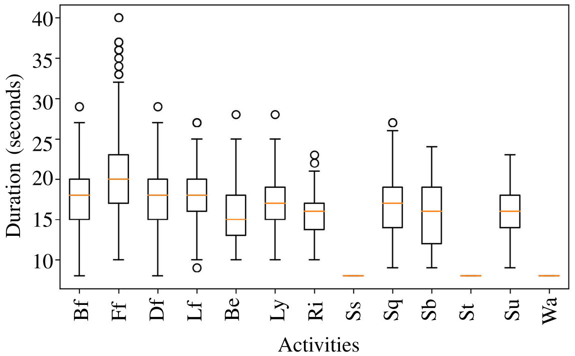


Fig. 3. Distribution of the duration of the collected clips for the 13 types of fall categories.

Although the egocentric camera does not record the subject who wears it, it may capture the people in the environment. That is the reason why the study presented in [46] refused to make their dataset public. We address this privacy-related situation by avoiding capturing the faces of other people and asking all volunteers involved to wear a mask since the cameras may also capture the faces of volunteers during certain activities like downside falls.

We divide the collected data into 13 classes that belong to two main categories: static and dynamic. Static classes comprise Sitting-Static, Standing, and Walking. Dynamic classes are composed of the remaining categories: back-falls (Bf), downside-falls (Df), front-falls (Ff), lateral-falls (Lf), bending (Be), lying (Ly), rising (Ri), sitting-down (Sd), squatting-down (Sq), and stumbling (Sb). These activities are selected based on the previous literature mentioned above. We intentionally omitted more energetic activities as our research specifically targets individuals with mobility problems, particularly the elderly who live alone and face challenges related to their mobility.

For the first group, subjects were instructed to simulate the activity for 5 min, both indoors and outdoors. The collected videos were then divided evenly by a video editor into nonoverlapping 8-s clips, as shown in Fig. 3. As to the events in the other category, the volunteers were asked to wait for 3 s, either standing or walking, before they started simulating the events and finally waited for 4 s before stopping the clip. This ensures that the event of interest is complete. Moreover, all volunteers were asked to change direction and position after

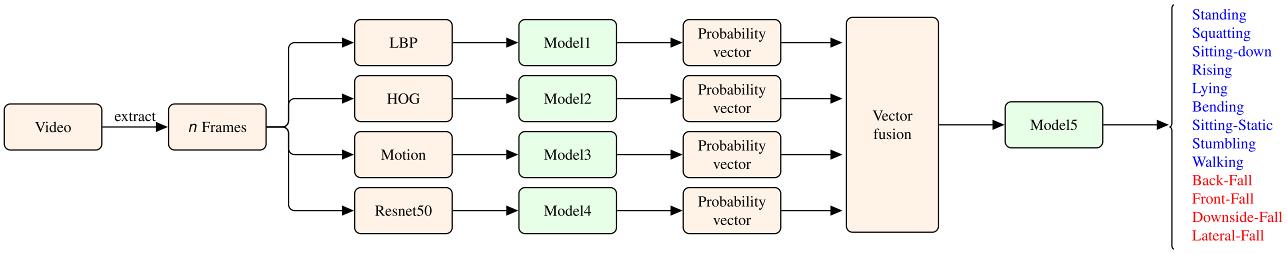


Fig. 4. Flowchart of our proposed pipeline. Given a video, handcrafted and deep features are extracted and fused for the detection and recognition of (red) falls and (blue) nonfalls.

each simulation to increase the diversity of the environment captured by the camera. To more accurately simulate real-world scenarios, each clip is assigned a single label, and the precise beginning and ending times of any fall activity within the clip are not provided to our learning algorithm. Following previous studies [18], [19], [20], [21], [22], [23], [24], the subjects were instructed to use a mat or soft land for both indoors and outdoors, or soft grounds (lawn or sand beach) for outdoors, to avoid any injuries. Some of the nonfall activities were also collected with the same soft ground used in the fall activities. This was done intentionally, to remove any effect the type of ground might have on the activities. Also, different surroundings were used for each subject to prevent the model from learning the scenes instead of the movements.

IV. Proposed Fall Detection Approach

Our proposed method uses a late decision fusion approach that is fed by four different descriptors from the videos: three handcrafted and one set of deep features. The stages of our proposed framework are depicted in Fig. 4.

A. Video Description With Handcrafted and Deep Features

1) Handcrafted Features: We use three types of handcrafted features, two of which, namely HOG and local binary pattern (LBP), concern individual frames, and the other one, optical flow, concerns the movement of features in consecutive frames. All video frames are resized to 128 × 64 pixels and transformed to grayscale before we compute the descriptors described below.

The HOG is a visual descriptor used in object detection, which counts occurrences of different gradient orientations in localized regions of a given image. We apply this approach to extract the HOG features for each video frame. Each frame is divided into smaller cells of size 8 × 8 pixels and a nine-bin orientation histogram is calculated for each cell. A block consisting of four (2 × 2) adjacent cells with a block stride of (8, 8) is used to L2-normalize the concatenated histograms of the four cells within it. Considering the sizes of our video frames, cells, and orientation histograms, our descriptor results in a (8 × 16 × 9 =) 1152-element vector. Finally, we compare two consecutive frames by computing the cosine similarity between the respective 1152-element vectors

$$\text{cosine similarity} = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{k} A_i B_i}{\sqrt{\sum_{i=1}^{k} A_i^2}\sqrt{\sum_{i=1}^{k} B_i^2}} \quad (1)$$

where A and B are two vectors of size k (here $k = 1152$) extracted from two consecutive frames. A given video is, therefore, described with a vector of $n - 1$ elements where n is the number of frames, and each element corresponds to the cosine similarity, which is a number between 0 and 1.

LBP [55] is another visual descriptor used to describe the neighborhood of image elements using binary codes. It describes features, such as edges, lines, spots, and flat areas, by using two complementary measures: local spatial patterns and grayscale contrast. The basic LBP descriptor considers the eight neighbors of each pixel and generates an 8-bit string. The neighbors that are greater than the pixel under consideration result in 1 bit and the others in 0 bits. The 8-bit string is then transformed to a decimal value and finally, a 256-bin histogram is generated for a given image or region. With this approach, the number of bins in the histogram increases exponentially with increasing neighbors P. In this work, we use the uniform LBP approach [56] with rotation and grayscale invariance to have a more concise descriptor. Instead of having a bin for every unique binary pattern, which would result in a 2^P bin, the concept of this LBP variant is to group all nonuniform patterns in one bin and keep a separate bin for each uniform pattern. A uniform pattern is a binary string that contains at most two transitions of the type "01" and "10." In general, there are $P + 1$ uniform binary patterns, two of which are flat (all zeros and all ones). With 56 neighbors per pixel considered along a circle with a radius of 8 pixels, as we use in this work, there are 57 uniform binary patterns, which form the first 57 bins in the histogram. An additional bin is added to represent the remaining or miscellaneous patterns that contain more than two transitions between zeros and ones. The 58-element descriptor is thus constructed by counting the number of instances of each uniform pattern as well as the number of instances of all nonuniform ones in a given video frame. This is followed by L2-normalization. Similar to HOG, we describe a video with a $n - 1$ element vector, where n is the number of frames and each element is the cosine similarity between the respective LBP descriptors of two adjacent frames.

Optical flow is the third handcrafted feature that we use. It calculates the motion between two consecutive video frames at every position. It is based on three assumptions.

1) *Spatial Coherence:* Nearby points in a frame plane move in a similar manner all the time (velocity smoothness constraint).
2) *Intensity Constancy:* The projection of a point is roughly constant over time.

3) *Small Motion:* Points do not move very far between adjacent frames.

Here, we use the Gunnar Farneback method to determine the motion parameters between consecutive frames [57]. It is a dense optical flow technique that considers all pixels instead of corners and edges used by sparse optical flow, and it estimates motion by comparing consecutive frames based on polynomial expansions.

The first step is to approximate some neighborhood of the pixels in the first frame by quadratic polynomials

$$f_1(x) = x^T A_1 x + b_1^T x + c_1 \tag{2}$$

where x refers to the pixels, and A_1, b_1, and c_1 are coefficients whose data structures are a symmetric matrix, a vector, and a scalar, respectively. They can be estimated by a weighted least-squares fit to the signal values in the neighborhood. Considering these quadratic polynomials, a new frame is constructed by a global displacement. The second consecutive frame with a global displacement of d can be estimated as

$$f_2(x) = f_1(x-d) = (x-d)^T A_1 (x-d) + b_1^T (x-d) + c_1. \tag{3}$$

The displacement results will be too noisy if (3) is solved pointwise, so specific neighborhoods of each pixel are combined to determine the displacement d. One could compute the displacement of each pixel for each frame by the steps mentioned above. To speed up the processing, however, we consider one pixel every 8 pixels from the resized frame of 128 × 64 pixels, which results in a feature vector of (16 × 8 =) 128 displacement values. The mean of this vector is used to characterize the motion between two frames. Each video is then represented with a feature vector of $(n-1)$ length, where n is the number of frames.

2) Truncating and Aligning Video Descriptors: In practice, we envisage a system that processes a time window of a user-defined length (depending on energy saving settings, for instance), with a minimum of 8 s, which is sufficient to capture an entire fall. In fact, our dataset consists of video files whose length varies between 8 and 40 s (see Fig. 2). Within that varying time window, we then compute the above three descriptors and truncate them to a fixed length of 238 elements, which represents the smallest time window of 8 s with 30 frames/s (minus the first and the last frames), centered around the respective maximum value.

3) Deep Features: We also consider the extraction of deep features from video frames with the Resnet50 [58] model pretrained on ImageNet [59]. Empirical studies demonstrate that this deep residual network has advantages in terms of feature extraction because of its solid representational capacity for the deep neural network. Resnet is notable for its robustness to network degradation (attributable to the concept of skip connections) and its ability to minimize the risk of vanishing gradients. Due to the requirements of the network, we resize the video frames to 224 × 224 pixels and extract the 2048-element vector from the last fully connected layer as a global descriptor of the image. To avoid high computational intensity, instead of extracting deep features for each frame we only extract such features for ten frames that are equally spaced in time in the given video clip. Each clip is then represented with a 20 480-element vector, which is the concatenation of all feature descriptors of the considered ten frames.

B. Decision Fusion and Classification

We use a decision fusion strategy that stacks the classification output of independent models, which are fed with the above-mentioned descriptors (see Fig. 4). The independent video descriptors based on the handcrafted (HOG, LBP, optical flow) and deep features (Resnet50) are used to train four different classification models: Model1, …, Model4. The output vectors of each of the four models are then fused by the following equation and used to train the fifth classification model Model5:

$$\begin{aligned}\text{DecisionFusion}\Big(&\left[p_1^H,\ldots,p_m^H\right],\left[p_1^L,\ldots,p_m^L\right],\\&\left[p_1^O,\ldots,p_m^O\right],\left[p_1^D,\ldots,p_m^D\right]\Big)\\&=\left[P_1,\ldots,P_m\right]\end{aligned} \tag{4}$$

where p_i^H, p_i^L, P_i^O, and p_i^D are the outputs of each of the HOG, LBP, optical flow, and deep features driven models that the given video belongs to class i, while the resulting P_i represents the final global output that a given video belongs to class i. With this type of decision fusion, we are giving equal importance to all types of features involved. The variable m represents the number of classes. When the proposed approach is used for fall detection, then $m = 2$ (fall or nonfall), and when it is used for the recognition of daily activities then $m = 12$. Finally, a given video is classified with the label of the class index of the maximum value in the vector P.

V. Experiments and Results

A. Design of Experiments

The effectiveness of the proposed approach is evaluated in different scenarios. We start by comparing various types of classifiers in our method and then we evaluate the effect of the *camera location* (neck versus waist) and the *scene environment* (indoor versus outdoor). For each experiment, we use leave-one-subject-out evaluation to test the generalization ability in two applications, namely fall detection (a two-class problem) and activity recognition (a 12-class problem). We merged the classes sitting_static and standing into one for the multiclass activity recognition task since they are both collected from a static viewpoint. We apply the leave-one-subject-out evaluation with a leave-one subject-out cross-validation. For a five-subject problem that we have in this work, this means that the models are trained with the samples of four subjects and tested on the data of the left-out subject. This is repeated five times so that all subjects participate in the testing. This approach evaluates the robustness of the proposed model in generalizing to new subjects. Note that the data collection process in our study was conducted in various environments, where none of the participants collected data in the same location, ensuring that the cross-validation methodology we

TABLE II

Results With Leave-One Subject-Out Evaluation for the Two-Class (Fall Detection) and 12-Class (Activity Recognition) Problems

Handcrafted features	Deep features	Fusion model	Accuracy 2 classes	Accuracy 12 classes
RF	RF	RF	0.901 (± 0.02)	0.657 (± 0.08)
SVM	SVM	SVM	0.875 (± 0.11)	0.534 (± 0.18)
MLP	MLP	MLP	0.898 (± 0.10)	0.602 (± 0.11)
MLP	RF	SVM	0.891 (± 0.03)	0.654 (± 0.04)
MLP	SVM	RF	0.874 (± 0.11)	0.552 (± 0.18)
SVM	MLP	RF	0.875 (± 0.12)	0.561 (± 0.17)
SVM	RF	MLP	0.882 (± 0.06)	0.636 (± 0.04)
RF	MLP	SVM	0.904 (± 0.09)	0.673 (± 0.10)
RF	SVM	MLP	**0.918 (± 0.08)**	**0.709 (± 0.09)**

employed validates the generalization ability across different users and scenarios.

We measure the performance using the accuracy, sensitivity, specificity

$$\begin{aligned}&\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{FP} + \text{FN} + \text{TN})\\&\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})\\&\text{Sensitivity or Recall (Re)} = \text{TP}/(\text{TP} + \text{FN}) \qquad (5)\end{aligned}$$

where TP, TN, FP, and FN refer to the number of true positives, true negatives, false positives, and false negatives, respectively. Sensitivity, also referred to as the true positive rate, measures the system's capability to accurately detect falls. On the other hand, specificity corresponds to the true negative rate, indicating the system's accuracy in correctly identifying nonfall activities. All experiments are conducted on a computing cluster.[2]

B. Model Comparison

The proposed approach illustrated in Fig. 4 includes five classification models, for which we do a comparative analysis. We evaluate the performance of three different supervised learning models, namely random forest (RF), SVM, and multilayer perceptron (MLP), with default hyperparameter values.[3] For the three handcrafted features, we always experiment with the same type of classifier. The decision of investigating the mentioned classifiers is motivated by the diverse capabilities with respect to high dimensions, energy efficiency, and effectiveness. In this analysis, we use all data from our five subjects, including the data coming from both waist and neck cameras as well as indoor and outdoor activities. Table II reports the results of all possible combinations for leave-one subject-out evaluation and for both the two-class and 12-class problems. The results clearly show that the best configuration is to use an RF for each of the handcrafted feature descriptors, an SVM for the deep features, and finally an MLP for fusion in the second layer.

TABLE III

Results of Waist Versus Neck

Camera location	No. of classes	Accuracy
Waist	2	0.915 (± 0.04)
Waist	12	0.704 (± 0.04)
Neck	2	0.940 (± 0.03)
Neck	12	0.650 (± 0.07)

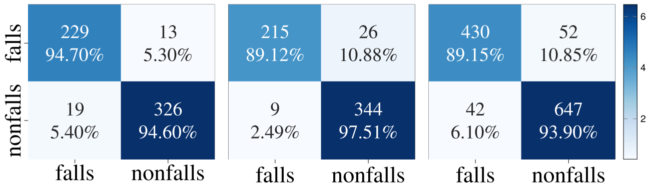


Fig. 5. Average confusion matrices based on external cross-validation for the binary classification problem using (left) the neck camera, (middle) using data in outdoor scenes captured by both neck and waist cameras, and (right) using all data from waist and neck cameras captured in indoor and outdoor scenes. The integers indicate the absolute quantities.

TABLE IV

Results of Indoor Versus Outdoor

Environment location	No. of classes	Accuracy
Indoor	2	0.901 (± 0.09)
Indoor	12	0.650 (± 0.07)
Outdoor	2	0.926 (± 0.08)
Outdoor	12	0.703 (± 0.07)

C. Waist Versus Neck

Next, we investigate whether there is a difference in performance concerning the location (waist or neck) of the attached camera. For this experiment, we evaluate the fusion method with the classifiers established in Section IV-B. Table III reports the results for the two-class and 12-class problems with leave-one-subject-out evaluation. Figs. 5 (left) and 6 illustrate the confusion matrices of the binary and multiclass problems for this experiment. The results demonstrate that for fall detection, and hence the most important problem, a camera attached to the neck yields better performance, while for the 12-class problem (activity recognition) the opposite is true. This conclusion is substantiated by a thorough manual inspection of the video clips obtained from the neck and waist locations.

D. Indoor Versus Outdoor

Since the proposed approach is aimed to be used in both indoor and outdoor environments, we evaluate the effect of the environment on performance. The data are split into indoor and outdoor subsets covering data from both neck and waist locations, the accuracy of which is calculated separately. We list the experimental results of the indoor and outdoor environments in Table IV, and the corresponding confusion matrices in Figs. 5 (middle) and 7. Note that we achieved slightly better performance for the outdoor environment using leave-one-subject-out cross-validation.

[2] The implementation is in Python, and the models are trained using NVIDIA Tesla K40 GPUs.

[3] RF: criterion = "entropy," class_weight = "balanced," n_estimators = 40; SVM: kernel = "linear"; and MLP: hidden_layer_sizes = 100, learning_rate_init = 0.001, alpha = 1, max_iter = 1000.

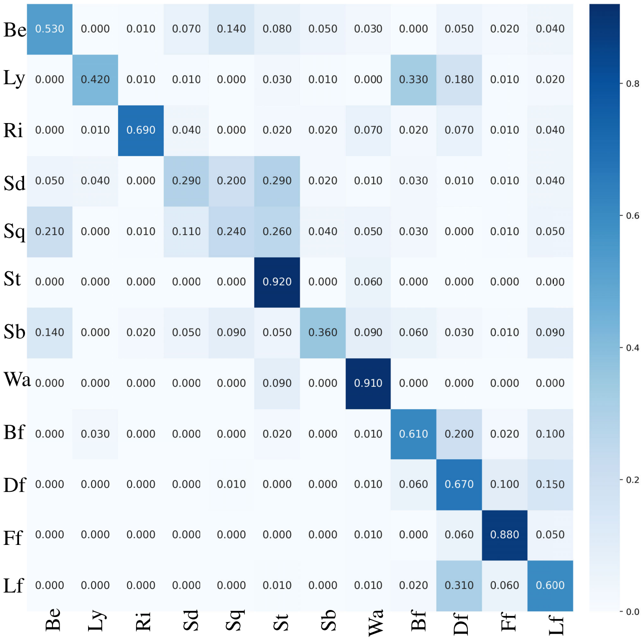


Fig. 6. Average confusion matrix achieved by leave-one-subject-out evaluation for the 12-class problem with the neck camera.

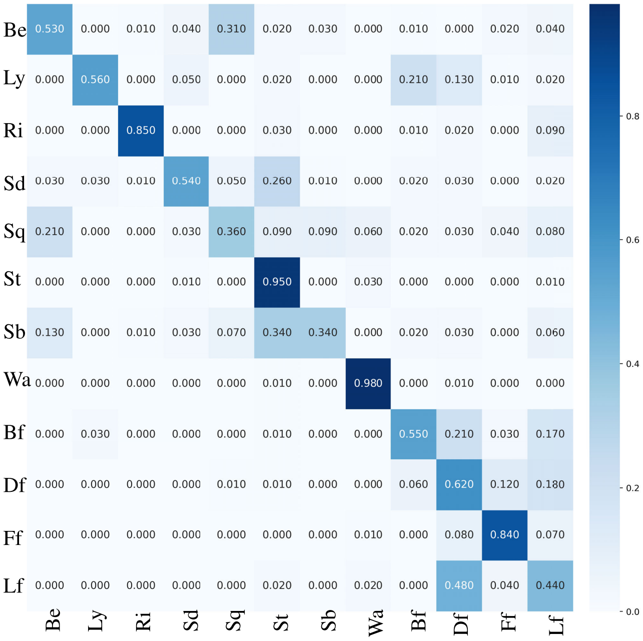


Fig. 7. Average confusion matrix achieved by leave-one-subject-out evaluation for the 12-class problem and outdoor scenes.

E. Our Approach Versus Baseline Frameworks

We evaluate the robustness of our framework against three baselines derived from it.

1) *B1:* We assess the simplest scenario, individual feature descriptor, where independent classifiers are learned for each of the three types of traditional feature descriptors (HOG, LBP, and optical flow) by RF and one with deep features (Resnet50) by SVM.
2) *B2:* The different descriptors are combined by concatenation and fed into a single SVM classifier.
3) *B3:* The three handcrafted descriptors are concatenated as one feature vector and used to train an RF classifier. The deep features are used to train an SVM classifier. Subsequently, the output vectors are concatenated and an MLP classifier is finally trained with these fused descriptors.

TABLE V
RESULTS FOR THREE BASELINE MODELS USING LEAVE-ONE-SUBJECT-OUT EVALUATION FOR BOTH THE TWO-CLASS (FALL DETECTION) AND 12-CLASS (DAILY ACTIVITY RECOGNITION) PROBLEMS. "ALL" REFERS TO THE THREE KINDS OF HANDCRAFTED FEATURES (HOG, LBP, AND OPTICAL FLOW) FOR B2, B3, AND THE PROPOSED FUSION APPROACH "OURS"

Fusion features	Handcrafted features	Deep model	Accuracy 2 classes	Accuracy 12 classes
B1	HOG	×	0.831 (± 0.06)	0.556 (± 0.09)
B1	LBP	×	0.840 (± 0.07)	0.547 (± 0.09)
B1	Optical flow	×	0.817 (± 0.10)	0.537 (± 0.08)
B1	×	✓	0.875 (± 0.12)	0.540 (± 0.17)
B2	All	✓	0.869 (± 0.05)	0.589 (± 0.06)
B3	All	✓	0.896 (± 0.05)	0.659 (± 0.11)
Ours	All	✓	**0.918 (± 0.08)**	**0.709 (± 0.09)**

Table V reports the comparative results, which indicate that our fusion approach outperforms these baselines for both the fall detection and activity recognition problems. Figs. 5 (right) and 8 show the confusion matrices of the binary and multiclass problems based on the results of our fusion approach obtained by leave-one-subject-out cross-validation. In Fig. 9, we illustrate an example set of temporally equidistant ten frames (out 240) for each of the 12 activities acquired by a neck camera, together with the ground truth and classified labels. We also compute the receiver-operating characteristic (ROC) curves for the five individual subjects along with the average ROC and show them in Fig. 10. The ROC curve is generated through the calculation of the true and false-positive rates for certain thresholds applied to the output of the underlying classification model. In our case, we applied an MLP in the last step of the decision fusion methods, so the various thresholds were applied to the output of its activation layer to generate the ROC in this study.

While the focus of fall detection systems is on binary decision-making (fall or no fall), we believe that studying activity recognition serves two important purposes. First, the inclusion of diverse activities allows us to gain a better understanding of the system's performance and its potential limitations. By evaluating the system's ability to recognize different activities, including those that may bear similarities to falls, we can identify any potential areas of confusion or misclassification. This analysis provides valuable insights into the system's robustness and helps to refine fall detection algorithms. Second, the recognition of specific types of falls holds significance in preparing primary caregivers for the expected health injuries. Different types of falls can result in varying degrees of injury and may require different response protocols. By accurately identifying and categorizing specific types of falls, the system can provide crucial information

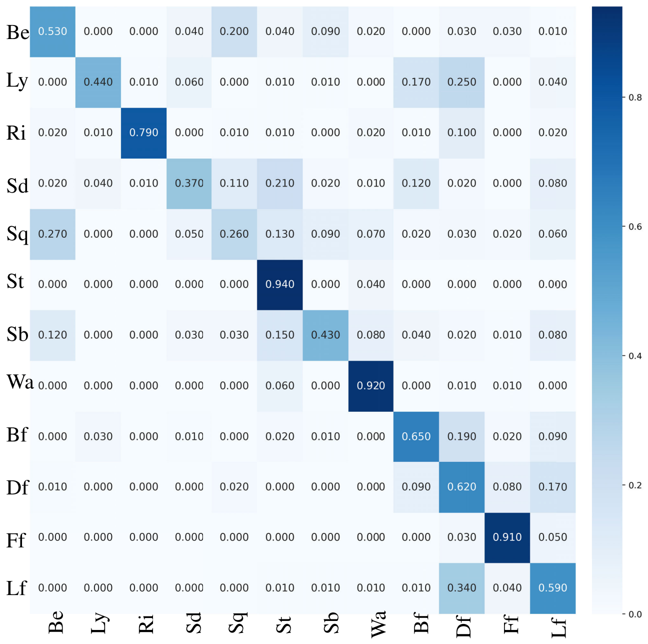


Fig. 8. Leave-one-subject-out cross-validation for 12 classes classification by all data from the waist and the neck, indoor and outdoor.

to caregivers, enabling them to respond appropriately and promptly to the individual's needs. This can lead to improved healthcare outcomes and potentially reduce the risk of further injury or complications.

The confusion matrices in Figs. 6–8 indicate that the majority of the errors are within the same two major classes. For instance, the last four rows reveal that most of the errors in certain fall activities are misclassified to other types of falls, such as Lateral falls (Lf) that are sometimes erroneously classified as Downside falls (Df). We do not consider this to be a major problem as it does not affect the primary goal of this work. The most notable errors are obtained with the Lying (Ly) activities (i.e., nonfall) where a bit less than half of them are misclassified as falls, mainly Backside falls (Bf) or Downside falls (Df). These activities do indeed share common movements, and they are the main cause for not achieving even higher specificity. In Fig. 9, we illustrate examples of three misclassifications including a Lying activity misclassified as a Downside fall. While we aim to investigate this further in our future work, we do note that this anomaly does not affect the sensitivity of our approach, which is more important in this application. We achieve a sensitivity of 0.936 (±0.09) and a specificity of 0.892 (±0.10). One may, however, move the operating point along the ROC curve to tune the approach according to the desired tradeoff between sensitivity and specificity.

While this study primarily focused on evaluating the potential of visual information for fall detection, we acknowledge the importance of addressing false alarms, including those caused by accidental drops of the camera itself. In the fully developed system that we envision, we propose a mechanism where, upon detecting a fall, the user is prompted on their smartphone to confirm whether it was indeed a real fall or not. This interactive step allows for user input and verification, reducing the likelihood of false alarms caused by accidental movements or drops. In cases where the user does not provide a response within a certain time window, the system can assume it is a real fall as a precautionary measure. Additionally, through the utilization of reinforcement learning techniques, the system could adapt and learn from the patterns and behaviors of each individual user, further refining the fall detection capabilities over time.

VI. Discussion

The first-person and vision-based approach with egocentric RGB cameras that we propose has several benefits. Besides its portability character and the effective results it achieves in fall detection, which is our primary concern here, it has the potential to be used in other applications. For instance, by logging the activities over time, one can monitor the social interaction of the individual, which can then be used for the early detection of psychological problems or cognitive declines, such as anxiety, loneliness, relapse of depression, and the onset of dementia. Unlike static surveillance systems, such an egocentric approach protects the identity of the individual wearing the camera, and techniques for blurring faces and vehicle number plates can easily be integrated to protect the identity of the people in the scenes.

Given that no similar dataset is publicly available in the literature, the comparison against other works was not possible. Hence, another contribution of this work is the dataset that we collected, which we made publicly available with this publication. Our dataset, collected by five volunteers who wore waist and neck cameras, describes a wide range of situations. The camera users recorded a total of 5858 video clips in indoor and outdoor environments, describing 13 different daily activities (four falls and nine nonfalls). Notable is the fact that besides the new dataset is the only publicly available one, it is also the largest compared to other unpublished datasets from previous works [16], [17], [46] of fall detection by wearable cameras. In this regard, and for comparison purposes, we implemented a set of different baseline methods against which we evaluated our fusion approach. The results demonstrate that the proposed fusion method achieves the best results.

In our experiments, we emphasize the importance of applying a method with a high generalization capacity. In particular, we used leave-one-subject-out evaluation to quantify the performance of our method. This means that the evaluation was carried out on data from subjects that never participated in the model configuration. The generalization capabilities of the model are very important in our project because we cannot expect to collect fall-related data from new subjects *before* they can use the system. In practice, once the system detects a fall, which can then be verified by the primary caregivers, that fall event will be used to improve the model. Moreover, automatically adjusting to personal styles of movements from different subjects is expected.

As previously mentioned, the dataset was collected from the point of view of the waist and neck. The results from our external evaluation indicate that the neck is the recommended location where to wear the camera for the detection of falls.

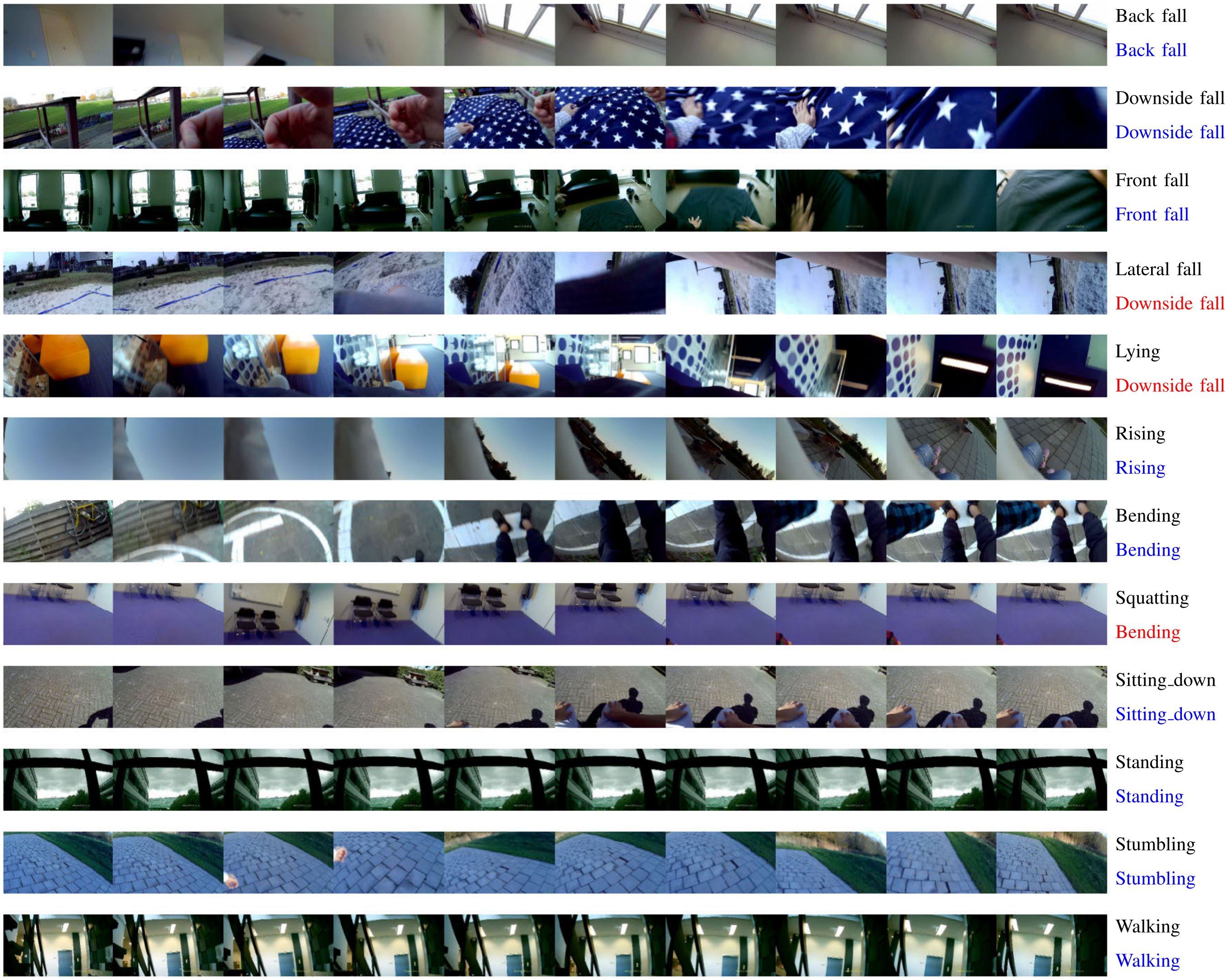


Fig. 9. We present examples of frame sequences of all 12 kinds of activities. The black labels indicate the ground truth, and the blue and red labels indicate the correct and incorrect classifications given by our proposed approach. Example activities from both indoor (back falls, front falls, lying, bending, standing, and walking) and outdoor (downside falls, lateral falls, rising, squatting, sitting, and stumbling) are presented here.

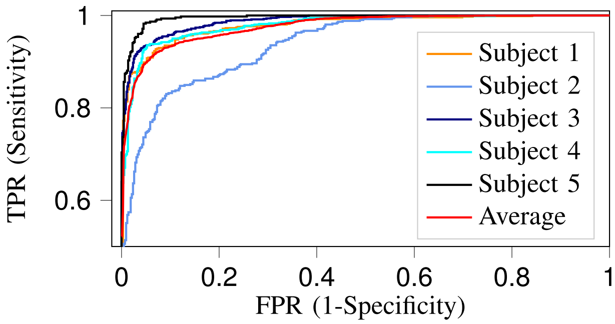


Fig. 10. ROC curves for the five individual subjects when their data are only used for testing, along with the average ROC.

This is, in fact, also more practical, as it turns out that a camera attached to the waist is more likely to have its view occluded than a neck camera, which might have played a role in some of the misclassifications. As to the environment is concerned, our approach achieves very comparable results between indoor and outdoor activities. A portable smart system that can be used indoors and outdoors would be key to improving the well-being of a large part of the population.

A. Practical Implications

We also explore the response time of the proposed approach. Given the hardware that we used to run our experiments (NVIDIA Tesla K40), our approach takes 0.069, 0.190, and 1.07 s to extract HOG, LBP, and Resnet50 features per frame, respectively. In our experiments, given a time window, we chose to extract Resnet50 features for ten frames and hand-crafted features for 240 frames to build the video descriptor. In total, an 8-s time window takes 1.53 s for HOG, 4.56 s for LBP, 3.59 s for optical flow, and 4.90 s for Resnet50 to process. These features can be extracted in parallel. In practice, we do not foresee a system to continuously extract features with every frame. A two-layer system similar to the one in [17] and [43] would be more practical, where a low-energy sensor (e.g., accelerometer or IMU) can be used to detect the onset of an event, and consequently trigger the methodology that we propose to take a decision of a time window surrounding the time point concerned.

We want to note that having a physical GPU processing data in a real scenario is complicated. This would affect the extraction of global descriptors using deep networks such as Resnet. For this reason, handcrafted features, such as the ones evaluated in this work, seem the most suitable option. The HOG is the fastest feature to compute and achieves reasonable results in comparison to the best performance achieved in our experiments, especially in the fall detection task. Relying on only HOG features would allow a more lightweight model that can be implemented in a portable device. Another consideration is the camera battery life. The wearable cameras have a battery life of 3 h when used with a resolution of 1080 pixels and 30 frames/s. While we acknowledge that the developed prototype currently operates for a duration of 3 h before requiring recharging, it is worth mentioning that this is just the current state and does not represent the limitations of future camera technologies. As battery technologies evolve and camera manufacturers prioritize power efficiency, we anticipate that wearable cameras will offer longer and more practical operation times in the near future.

For the time being, we focused our current efforts on investigating the potential of egocentric (RGB) visual information alone for fall detection. We aimed to assess whether the video frames captured by wearable cameras contain sufficient information to detect falls accurately. By focusing on this aspect, we were able to establish a foundation and understand the potential capabilities of egocentric visual information in fall detection scenarios. In future research, we will investigate the integration of low-cost wearable inertial sensors and other modalities to further enhance fall detection systems. The combination of image analysis and wearable inertial sensors can provide a comprehensive and robust solution, taking advantage of the strengths of each modality. For instance, such wearable inertial sensors may be crucial to compensate for certain limitations of wearable cameras, such as when the camera's view may be accidentally obstructed or covered.

There are several promising avenues for future research, encompassing both the sensors used for data capture and the underlying algorithms. In this work, we primarily focused on RGB wearable cameras, which have certain limitations in low-illumination environments. To address this issue, we propose two potential directions: incorporating infrared cameras or exploring the fusion of audio and vision modalities. In our future investigations, we intend to extend our research by incorporating infrared and event-based cameras [60]. Infrared cameras are unaffected by scene illumination, providing improved performance in low-light conditions. On the other hand, event-based cameras, also known as neuromorphic cameras, offer a highly energy-efficient solution, resulting in extended battery life. These cameras acquire data asynchronously at the pixel level, offering distinct advantages. Moreover, event-based cameras demonstrate enhanced privacy features, as their asynchronous and binary data acquisition makes it significantly more challenging to discern privacy-sensitive visual information compared to RGB cameras. In [61], we demonstrate preliminary results on the application of event-based cameras for the fall detection problem. Furthermore, we plan to explore the development of a multimodal system that integrates both visual and audio data. Incorporating audio information can complement visual cues and potentially enhance the overall performance of the fall detection system. By considering both modalities, we aim to leverage the complementary strengths of visual and audio data for more robust and accurate fall detection. These proposed research directions, encompassing the integration of infrared cameras, event-based cameras, and multimodal systems, hold promise for advancing the field of fall detection and addressing the limitations associated with RGB wearable cameras.

As to algorithms, one may investigate a tradeoff between the effectiveness of end-to-end heavyweight deep-learning approaches, such as LSTM and Transformers, concerning the power consumption of a wearable and portable solution. This would require first extending the current dataset substantially from more young healthy adults but also with safe (i.e., nonfall) activities by elderly people.

Future work will also include addressing the hardware constraints in a full system and exploring options such as implementing the detection algorithm in a transportable node or utilizing smartphones as external processing units. We will carefully evaluate the memory, battery, and computation requirements of the algorithm to develop a practical and efficient fall detection system. Additionally, we will consider the transmission of signals from the wearable camera to external devices, taking into account wireless communication protocols and optimizing energy efficiency.

VII. Conclusion

The approach that we propose is effective for the detection of falls and the classification of other daily activities. It is based on the use of an RGB wearable camera and an algorithmic approach that is conceptually simple to implement. In particular, the leave-one-subject-out cross-validation indicates that our fusion approach generalizes to new subjects and performs equally well in indoor and outdoor environments. It turned out that a neck-worn camera yields better performance and is also more practical than a waist-worn camera. The proposed system demonstrated its potential for fall detection and recognition and will be further explored using data collected by different cameras and relying on the multimodal data provided by a video.

Acknowledgment

The authors thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high-performance computing cluster. There is nothing declared about the Conflict of Interest Statement.

References

[1] S. Elliott, J. Painter, and S. Hudson, "Living alone and fall risk factors in community-dwelling middle age and older adults," *J. Community Health*, vol. 34, no. 4, pp. 301–310, Aug. 2009.

[2] G. Yavuz et al., "A smartphone based fall detector with online location support," in *Proc. Int. Workshop Sens. App Phones*, Zurich, Switzerland, 2010, pp. 31–35.

[3] J. A. Haagsma et al., "Falls in older aged adults in 22 European countries: Incidence, mortality and burden of disease from 1990 to 2017," *Injury Prevention*, vol. 26, no. 2, pp. i67–i74, 2020.

[4] S. Sadigh, A. Reimers, R. Andersson, and L. Laflamme, "Falls and fall-related injuries among the elderly: A survey of residential-care facilities in a Swedish municipality," *J. Community Health*, vol. 29, no. 2, pp. 129–140, Apr. 2004.

[5] X. Wang, J. Ellul, and G. Azzopardi, "Elderly fall detection systems: A literature survey," *Frontiers Robot. AI*, vol. 7, p. 71, Jun. 2020.

[6] K. Ozcan and S. Velipasalar, "Wearable camera- and accelerometer-based fall detection on portable devices," *IEEE Embedded Syst. Lett.*, vol. 8, no. 1, pp. 6–9, Mar. 2016.

[7] A. Fathi, A. Farhadi, and J. M. Rehg, "Understanding egocentric activities," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 407–414.

[8] E. H. Spriggs, F. De La Torre, and M. Hebert, "Temporal segmentation and activity classification from first-person sensing," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2009, pp. 17–24.

[9] D. Tome, P. Peluse, L. Agapito, and H. Badino, "xR-EgoPose: Egocentric 3D human pose from an HMD camera," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7728–7738.

[10] E. Talavera, N. Strisciuglio, N. Petkov, and P. Radeva, "Sentiment recognition in egocentric photostreams," in *Proc. Iberian Conf. Pattern Recognit. Image Anal.* Cham, Switzerland: Springer, 2017, pp. 471–479.

[11] A. Furnari and G. Farinella, "What would you expect? Anticipating egocentric actions with rolling-unrolling LSTMs and modality attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6251–6260.

[12] E. Talavera, A. Cola, N. Petkov, and P. Radeva, "Towards egocentric person re-identification and social pattern analysis," *Appl. Intell. Syst.*, vol. 310, no. 5, pp. 203–211, 2019.

[13] F. M. Li, D. L. Chen, M. Fan, and K. N. Truong, "FMT: A wearable camera-based object tracking memory aid for older adults," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 3, pp. 1–25, Sep. 2019.

[14] T. G. Stavropoulos, A. Papastergiou, L. Mpaltadoros, S. Nikolopoulos, and I. Kompatsiaris, "IoT wearable sensors and devices in elderly care: A literature review," *Sensors*, vol. 20, no. 10, p. 2826, May 2020.

[15] K. Md. Shahiduzzaman, X. Hei, C. Guo, and W. Cheng, "Enhancing fall detection for elderly with smart helmet in a cloud-network-edge architecture," in *Proc. IEEE Int. Conf. Consum. Electron.*, May 2019, pp. 1–2.

[16] M. Casares, K. Ozcan, A. Almagambetov, and S. Velipasalar, "Automatic fall detection by a wearable embedded smart camera," in *Proc. 6th Int. Conf. Distrib. Smart Cameras (ICDSC)*, Oct. 2012, pp. 1–6.

[17] K. Ozcan, A. K. Mahabalagiri, M. Casares, and S. Velipasalar, "Automatic fall detection and activity classification by a wearable embedded smart camera," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 3, no. 2, pp. 125–136, Jun. 2013.

[18] Y. Li, T. Banerjee, M. Popescu, and M. Skubic, "Improvement of acoustic fall detection using Kinect depth sensing," in *Proc. 35th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2013, pp. 6736–6739.

[19] M. Kangas, A. Konttila, P. Lindgren, I. Winblad, and T. Jämsä, "Comparison of low-complexity fall detection algorithms for body attached accelerometers," *Gait Posture*, vol. 28, no. 2, pp. 285–291, Aug. 2008.

[20] A. K. Bourke, J. V. O'Brien, and G. M. Lyons, "Evaluation of a threshold-based tri-axial accelerometer fall detection algorithm," *Gait Posture*, vol. 26, no. 2, pp. 194–199, Jul. 2007.

[21] X. Ma, H. Wang, B. Xue, M. Zhou, B. Ji, and Y. Li, "Depth-based human fall detection via shape features and improved extreme learning machine," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 6, pp. 1915–1922, Nov. 2014.

[22] M. Saleh and R. L. B. Jeannès, "Elderly fall detection using wearable sensors: A low cost highly accurate algorithm," *IEEE Sensors J.*, vol. 19, no. 8, pp. 3156–3164, Apr. 2019.

[23] M. Zitouni, Q. Pan, D. Brulin, and E. Campo, "Design of a smart sole with advanced fall detection algorithm," *J. Sensor Technol.*, vol. 9, no. 4, pp. 71–90, 2019.

[24] M. Cheffena, "Fall detection using smartphone audio features," *IEEE J. Biomed. Health Informat.*, vol. 20, no. 4, pp. 1073–1080, Jul. 2016.

[25] X. Wang, E. Talavera, D. Karastoyanova, and G. Azzopardi, "Fall detection and recognition from egocentric visual data: A case study," in *Proc. Int. Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2021, pp. 431–443.

[26] E. Cippitelli, F. Fioranelli, E. Gambi, and S. Spinsante, "Radar and RGB-depth sensors for fall detection: A review," *IEEE Sensors J.*, vol. 17, no. 12, pp. 3585–3604, Jun. 2017.

[27] A. Shrestha, J. Le Kernec, F. Fioranelli, E. Cippitelli, E. Gambi, and S. Spinsante, "Feature diversity for fall detection and human indoor activities classification using radar systems," in *Proc. Int. Conf. Radar Syst. (Radar)*, Belfast, Northern Ireland, 2017, pp. 1–6, doi: 10.1049/cp.2017.0381.

[28] G. Mastorakis and D. Makris, "Fall detection system using Kinect's infrared sensor," *J. Real-Time Image Process.*, vol. 9, no. 4, pp. 635–646, 2014.

[29] I. Charfi, J. Miteran, J. Dubois, M. Atri, and R. Tourki, "Definition and performance evaluation of a robust SVM based fall detection solution," in *Proc. SITIS*, vol. 12, 2012, pp. 218–224.

[30] Z. Cai, J. Han, L. Liu, and L. Shao, "RGB-D datasets using Microsoft Kinect or similar sensors: A survey," *Multimedia Tools Appl.*, vol. 76, no. 3, pp. 4313–4355, Feb. 2017.

[31] N. Lu, Y. Wu, L. Feng, and J. Song, "Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 1, pp. 314–323, Jan. 2019.

[32] E. E. Stone and M. Skubic, "Fall detection in homes of older adults using the Microsoft Kinect," *IEEE J. Biomed. Health Informat.*, vol. 19, no. 1, pp. 290–301, Jan. 2015.

[33] T. Banerjee, J. M. Keller, and M. Skubic, "Resident identification using Kinect depth image data and fuzzy clustering techniques," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2012, pp. 5102–5105.

[34] C. Mosquera-Lopez et al., "Automated detection of real-world falls: Modeled from people with multiple sclerosis," *IEEE J. Biomed. Health Informat.*, vol. 25, no. 6, pp. 1975–1984, Jun. 2021.

[35] X. Xi, M. Tang, S. M. Miran, and Z. Luo, "Evaluation of feature extraction and recognition for activity monitoring and fall detection based on wearable sEMG sensors," *Sensors*, vol. 17, no. 6, p. 1229, May 2017.

[36] O. Kerdjidj, N. Ramzan, K. Ghanem, A. Amira, and F. Chouireb, "Fall detection and human activity classification using wearable sensors and compressed sensing," *J. Ambient Intell. Humanized Comput.*, vol. 11, no. 1, pp. 349–361, Jan. 2020.

[37] X. Xi, W. Jiang, Z. Lü, S. M. Miran, and Z.-Z. Luo, "Daily activity monitoring and fall detection based on surface electromyography and plantar pressure," *Complexity*, vol. 2020, pp. 1–12, Jan. 2020.

[38] T. de Quadros, A. E. Lazzaretti, and F. K. Schneider, "A movement decomposition and machine learning-based fall detection system using wrist wearable device," *IEEE Sensors J.*, vol. 18, no. 12, pp. 5082–5089, Jun. 2018.

[39] K. Ozcan, S. Velipasalar, and P. K. Varshney, "Autonomous fall detection with wearable cameras by using relative entropy distance measure," *IEEE Trans. Hum.-Mach. Syst.*, vol. 47, no. 1, pp. 31–39, Feb. 2017.

[40] I. Boudouane, A. Makhlouf, N. Saadia, and A. Ramdane-Cherif, "Wearable camera for fall detection embedded system," in *Proc. 4th Int. Conf. Smart City Appl.*, Oct. 2019, pp. 1–6.

[41] I. Boudouane, A. Makhlouf, M. A. Harkat, M. Z. Hammouche, N. Saadia, and A. R. Cherif, "Fall detection system with portable camera," *J. Ambient Intell. Humanized Comput.*, vol. 11, pp. 1–13, Jul. 2019.

[42] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 886–893.

[43] K. Ozcan, A. K. Mahabalagiri, and S. Velipasalar, "Fall detection and activity classification using a wearable smart camera," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2013, pp. 1–6.

[44] Y. Li, Z. Ye, and J. M. Rehg, "Delving into egocentric actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 287–295.

[45] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vis.*, vol. 103, no. 1, pp. 60–79, May 2013.

[46] M. Masuda, R. Hachiuma, R. Fujii, and H. Saito, "Unsupervised anomaly detection of the first person in gait from an egocentric camera," in *Proc. Int. Symp. Vis. Comput.*, 2020, pp. 1–4.

[47] E. Stack, "Falls are unintentional: Studying simulations is a waste of faking time," *J. Rehabil. Assistive Technol. Eng.*, vol. 4, Jan. 2017, Art. no. 2055668317732945.

[48] O. Aziz et al., “Validation of accuracy of SVM-based fall detection system using real-world fall and non-fall datasets,” *PLoS ONE*, vol. 12, no. 7, Jul. 2017, Art. no. e0180318.

[49] G. Demiris, B. K. Hensel, M. Skubic, and M. Rantz, “Senior residents’ perceived need of and preferences for ‘smart home’ sensor technologies,” *Int. J. Technol. Assessment Health Care*, vol. 24, no. 1, pp. 120–124, 2008.

[50] S. N. Robinovitch et al., “Video capture of the circumstances of falls in elderly people residing in long-term care: An observational study,” *Lancet*, vol. 381, no. 9860, pp. 47–54, Jan. 2013.

[51] G. Debard et al., “Camera-based fall detection using real-world versus simulated data: How far are we from the solution?” *J. Ambient Intell. Smart Environ.*, vol. 8, no. 2, pp. 149–168, Mar. 2016.

[52] V. Komisar et al., “Injuries from falls by older adults in long-term care captured on video: Prevalence of impacts and injuries to body parts,” *BMC Geriatrics*, vol. 22, no. 1, pp. 1–11, Apr. 2022.

[53] S. Abbate, M. Avvenuti, P. Corsini, J. Light, and A. Vecchio, “Monitoring of human movements for fall detection and activities recognition in elderly care using wireless sensor network: A survey,” in *Wireless Sensor Networks: Application-Centric Design*. Rijeka, Croatia: InTech, 2010, pp. 147–166.

[54] X. Yu, “Approaches and principles of fall detection for elderly and patient,” in *Proc. 10th Int. Conf. E-Health Netw., Appl. Services*, Jul. 2008, pp. 42–47.

[55] T. Ojala, M. Pietikainen, and D. Harwood, “Performance evaluation of texture measures with classification based on Kullback discrimination of distributions,” in *Proc. 12th Int. Conf. Pattern Recognit.*, 1994, pp. 582–585.

[56] T. Ojala, M. Pietikainen, and T. Maenpaa, “Multiresolution gray-scale and rotation invariant texture classification with local binary patterns,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, Jul. 2002.

[57] G. Farnebäck, “Two-frame motion estimation based on polynomial expansion,” in *Proc. Scandin. Conf. Image Anal.* Cham, Switzerland: Springer, 2003, pp. 363–370.

[58] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[59] O. Russakovsky et al., “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[60] S. U. Innocenti, F. Becattini, F. Pernici, and A. D. Bimbo, “Temporal binary representation for event-based action recognition,” in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 10426–10432.

[61] X. Wang, N. Risi, E. T. Martínez, E. Chicca, and G. Azzopardi, “Fall detection with event-based data: A case study,” in *Proc. 20th Int. Conf. Comput. Anal. Images Patterns*. Cham, Switzerland: Springer, Sep. 2003, pp. 1–12.