TESIS DOCTORAL

2023

FALL DETECTION SYSTEM BASED ON INFRARED IMAGES

SISTEMA DE DETECCIÓN DE CAIDAS BASADO EN IMÁGENES INFRARROJAS

JESÚS GUTIÉRREZ GALLEGO

PROGRAMA DE DOCTORADO EN TECNOLOGÍAS INDUSTRIALES

Director **DR. SERGIO MARTÍN GUTIÉRREZ**

Codirector **DR. VICTOR RODRÍGUEZ ONTIVEROS**

UNED

EIDUNED
Escuela
Internacional
de Doctorado

TESIS DOCTORAL

2023

FALL DETECTION SYSTEM BASED ON INFRARED IMAGES

SISTEMA DE DETECCIÓN DE CAIDAS BASADO EN IMÁGENES INFRARROJAS

JESÚS GUTIÉRREZ GALLEGO

PROGRAMA DE DOCTORADO EN TECNOLOGÍAS INDUSTRIALES

Director **DR. SERGIO MARTÍN GUTIÉRREZ**

Codirector **DR. VICTOR RODRÍGUEZ ONTIVEROS**

UNED

EIDUNED
Escuela
Internacional
de Doctorado

Table of content

UNED

EIDUNED
Escuela
Internacional
de Doctorado

Abstract

In a world whose population is becoming older and older the field of elderly care, already very relevant, is expected to become extraordinary important. The resources and personnel devoted to tasks in this field will exponentially grow during the next three decades, as the number of people aged 65 and above will double during this period.

The sector has, so far, automated a very low number of tasks. However, a higher number of them will need to be automated if universal elderly care at reasonable costs is desired. One of the tasks candidate to this automation will be elderly surveillance and, within that field, fall detection. This area has attracted a considerable amount of research interest over the last few years. However, the disconnection between the communities of system researchers and users has hampered the widespread use of this kind of systems.

This thesis encompasses the identification of real user's needs and the design and development of an innovative system adapted to them and their real needs. To do it a thorough revision of all the research work in the field published from 2015 is carried out. This revision identifies two major shortfalls; a deep disconnection between researchers and users and an almost complete absence of real data to develop systems.

To address the first problem, a major study among users is carried out. This work, the largest of its kind, identifies what are the real needs and perceptions of the different communities integrated in the elderly care field. It sheds some light on this area and although clearly suggests that human supervision is always preferred due to the added value it provides, it also identifies the circumtances under which the users would accept the use of fall detection systems.

Some of these situations are not covered by any of the already developed systems, as they imply surveillance under no light conditions and the use of sensors carried by the monitored person would not be a reasonable option. However, vision-based technology using far infrared imagery (FIR) is ideal to address these particular circumstances.

This way, and in order to evaluate the most significant human pose estimation models developed to process visual spectrum (RGB) imagery, a major dataset composed by far infrared video clips of a number of volunteers executing different activities is compiled. This dataset, called FIR-Human, also contains the annotations of joints positions, so model training an evaluation becomes possible.

To address the second problem, the absence of real data and the generalization problem associated to it, an alternative approach to automatic fall detection is proposed in this work. Present vision-based fall detection systems are developed using datasets recorded by young actors. Given the differences in the way young and old people move, the kinematic descriptors used by these systems in order to assess fall, which are a generalization from descritpors used for young people to old patients, could be inappropriate to determine whether a real fall has taken place. Our system, which uses dynamic descriptors instead of kinematic ones, approaches the human body in terms of balance and stability, thus, differences between real and simulated falls become irrelevant, as all falls are a direct result of a failure in the continuous effort of the body to keep balance, regardless of other considerations.

Then, the performances of a system, which integrates human pose estimation, based on far infrared imagery and dynamic descriptors are evaluated using the FIR-Human dataset fall section.

Finally, a number of general conclusions are reached and some suggestions for further research are suggested.

Resumen

En un mundo cuya población está envejeciendo cada vez más, el campo del cuidado de las personas mayores, que ya es muy relevante, se espera que se vuelva extraordinariamente importante. Los recursos y el personal dedicado a tareas en este campo crecerán de forma exponencial durante las próximas tres décadas, ya que el número de personas mayores de 65 años o más se duplicará durante este período.

Hasta ahora, el sector ha automatizado muy pocas tareas. Sin embargo, un mayor número de ellas necesitarán ser automatizadas si se desea un cuidado universal de personas mayores a costos razonables. Una de las tareas candidatas para esta automatización será la vigilancia de personas mayores y, dentro de ese campo, la detección de caídas. Esta área ha atraído un considerable interés de investigación en los últimos años. Sin embargo, la desconexión entre las comunidades de investigadores y usuarios ha dificultado el uso generalizado de este tipo de sistemas.

Esta tesis abarca la identificación de las necesidades reales de los usuarios y el diseño y desarrollo de un sistema innovador adaptado a ellos y sus necesidades reales. Para hacerlo, se lleva a cabo una revisión exhaustiva de todos los trabajos de investigación en el campo publicados desde 2015. Esta revisión identifica dos deficiencias principales: una profunda desconexión entre los investigadores y los usuarios, y una casi completa ausencia de datos reales para desarrollar sistemas.

Para abordar el primer problema, se lleva a cabo un importante estudio entre los usuarios. Este trabajo, el más grande de su tipo, identifica cuáles son las necesidades y percepciones reales de las diferentes comunidades integradas en el campo del cuidado de personas mayores. El estudio sugiere claramente que, aunque la supervisión humana siempre es preferida debido al valor añadido que proporciona el contacto humano, existen circunstancias en las que los usuarios aceptarían el uso de sistemas de detección de caídas.

Algunas de estas situaciones no están cubiertas por ninguno de los sistemas ya desarrollados, ya que están asociados a entornos nocturnos no iluminados en los que el uso de sensores adheridos al cuerpo no sería una opción razonable. Sin embargo, la tecnología basada en imágenes de infrarrojo lejano (FIR) es ideal para abordar estos escenarios.

De esta manera, y con el fin de evaluar las prestaciones de los modelos más significativos de estimación de la postura humana desarrollados para procesar imágenes del espectro visible (RGB) cuando trabajan con imágenes FIR, se compila una base de datos compuesta por clips de video de infrarrojo lejano de varios voluntarios realizando diferentes actividades. Esta base de datos, llamada FIR-Human, también contiene las anotaciones de las posiciones de las articulaciones, lo que permite el entrenamiento y evaluación de los modelos.

Para abordar la segunda cuestión, la falta de datos reales, y el problema de la generalización asociado a ella, se propone en este trabajo un enfoque alternativo para la detección automática de caídas. Los sistemas actuales de detección de caídas basados en visión se desarrollan utilizando bases de datos grabados por actores o voluntarios jóvenes. Dadas las diferencias en la forma en que se mueven las personas jóvenes y mayores, los descriptores cinemáticos utilizados por estos sistemas para evaluar las caídas, que son una generalización de los descriptores determinados para personas jóvenes, podrían ser inapropiados para establecer si ha ocurrido una caída real. Nuestro sistema, que utiliza descriptores dinámicos en lugar de

cinemáticos, aborda el cuerpo humano en términos de equilibrio y estabilidad, por lo que las diferencias entre caídas reales y simuladas se vuelven irrelevantes, ya que todas las caídas son el resultado directo de un fallo en el esfuerzo continuo del cuerpo por mantener el equilibrio, independientemente de otras consideraciones.

A continuación, se evalúan las prestaciones de un sistema que integra la estimación de la postura humana sobre imagen ifrarroja lejana y los descriptores dinámicos, utilizando para ello el bloque de caídas de la base de datos FIR-Human.

Por último, se llega a una serie de conclusiones generales y se hacen algunas sugerencias para futuras investigaciones.

List of tables

List of figures

List of acronyms

1D - One-dimensional.

2D - Two-dimensional.

3D - Three-dimensional.

AC - Accuracy.

ADL - Activities of daily life.

ANN - Artificial neural networks.

ARIMA - Auto-regression and moving average model.

ARMA - Autoregressive-moving-average.

ASM - Active shape models.

AUC - Area under the curve.

BB - Bounding box.

BDM - Bayesian decision making.

Bi-LSTM - Bi-directional long-short term memory.

BoS - Base of support.

CN-ANN - Capsule network – artificial neural network.

CNN – ANN - Convolutional neural network - artificial neural network.

CNN - Convolutional neural network.

COG - center of gravity.

COM - Center of mass.

COM - Center of mass.

COP - Center of pressure.

CPM - Convolutional pose machines.

CS - Compressed sensing.

CSS - Curvature scale space.

DA - Discriminant analysis.

DCNN-ANN - Deep convolution neural networks –artificial neural network.

DN-AAN - Deepnet – artificial neural network.

DRN-ANN - Deep residual network.

DT - Decision tree.

DTM - Dynamic template matching.

DTW - Dynamic time warping.

EBT - Ensemble bagged tree.

EEG – Electroencephalography.

EKF - End key frame.

ELM-ANN - Extreme learning machine – artificial neural network.

EMG – Electrocardiography.

ESAEs-OCCCH-ANN - Ensemble stacked auto encoders - one-class classification based on the convex hull – artificial neural network.

F-ANN - Feed-forward neural network.

FC – Forecasted.

FCM - Fuzzy C-means algorithms.

FCNN - Fully connected neural networks.

FDA - Fisher linear discriminant analysis.

FFT - Fast Fourier transformer.

FIR - Far infrared.

FLOP - Floating-point operations.

FMCW - Frequency modulated continuous wave.

FMM-ANN - Fuzzy min-max neural network.

FN - False negative.

FNR - False negative rate.

FP - False positive.

FPE - Foot placement estimator.

FPR - False positive ratio.

GBDT - Gradient boosting decision trees.

GK-FDA - Gaussian kernel fisher linear discriminant analysis.

GK-SVM - Gaussian kernel support vector machine.

GM-HMM - Gaussian mixture- hidden Markov model.

GMM - Gaussian mixed models.

GMS - Gaussian mean super vectors.

GNB - Gaussian naive Bayes.

GRU-ANN - Gated recurrent unit artificial neural network.

GSVM - Gaussian support vector machine.

GT - Ground truth.

HMM - Hidden Markov model.

HOG - Histogram of oriented gradients.

ICA - Illumination change-resistant algorithm.

INN - Increased nearest neighbor.

IP - Internet Protocol.

IR – Infrared.

ISVM - Incremental support vector machine.

Iv3 – ANN - Inception-v3 artificial neural network.

KNN - K-Nearest Neighbor.

LBP - Local binary pattern.

LBP-TOP - Local binary pattern histograms from three orthogonal planes.

LCD - Liquid Cristal Display.

LDA - Linear discriminant analysis.

LED - Light-emitting diode.

LFF - Left forefoot.

L-GEM - Localized-generalization error model.

LH - Left heel.

LKT - Lucas–Kanade–Tomasi.

LMB-ANN - Levenberg–Marquardt back-propagation - artificial neural network.

LOF - Local outlier factor.

LR - Logistic regression.

LSM - Least squares method.

LS-SVM - Least square- SVM.

LSTM – Long-short term memory.

LSTM-ANN – Long-short term memory artificial neural network.

M/L – Mediolateral.

MAD - Mean absolute deviation.

MCF - Motion co-occurrence feature.

MEWMA - Multivariate exponentially weighted moving average.

MFCC - Mel frequencies cepstral coefficients.

MHI - Motion history image.

MLNFNN-ANN - Multi-layer neuro-fuzzy neural network - artificial neural network.

MLP - Multilayer perceptron.

MLP-ANN - Multilayer perceptron – artificial neural network.

MPJPE - Mean per joint position error.

MSE - Mean square error.

NB - Naïve Bayes.

NDT - Normal distributions transformer.

NFNN-ANN - Neuro-fuzzy neural network - artificial neural network.

OCNN - One-class nearest neighbor.

OCSVM - One-class SVM.

OF - Optical flow.

O-SVM - One-class support vector machine.

PCK - Percentage of correct key-points.

PCP - Percentage of correct parts.

PDJ - Percentage of detected joints.

PID - Proportional-integral-differential.

PIR - Pyroelectric IR.

PPC - Probabilistic principal component.

PPG – Photoplethismography.

PRF - Pulse repetition frequency.

QSVM - Quadratic support vector machine.

RB - Rule based.

RBF-ANN Radial basis function - artificial neural network.

RBFNN - Radial basis function neural network.

RBPF - Rao–Blackwellized particle filter.

RBS - Rule-based systems.

ReNN-ANN - Recurrent neural network – artificial neural network.

RF - Random forest.

RFF - Right forefoot.

RGB - Red green blue.

RH - Right heel.

RNN - Recurrent neural networks.

RPCA - Robust principal component analysis.

SDK - Software development kit.

SE - Sensitivity.

SIFT - Scale invariant feature transform.

SOM-ANN – Self-organizing maps - artificial neural network.

SP - Specificity.

SRC - Sparse representations classification.

SSD - Single-shot detector.

SVDD - Support vector data description.

SVM - Support vector machine.

TN - True negative.

TP - True positive.

TPR - True positive ratio.

XCoM - Extrapolated center of mass over X axis.

YCoM - Extrapolated center of mass over Y axis.

1 Introduction

The UN report on world population [1] estimates that the number of people over 60 doubled its number in 2017 compared to 1980 and that number is expected to double again by 2050, when it will hit the 2 billion mark. By that time, it will exceed the combined number of teenagers and young adults with ages between 10 and 24.

Although the aging phenomenon is more intense in the developed world, it is actually a global one, affecting the entire humankind. Under this perspective, the amount of resources devoted to elderly care is expected to rise significantly in the years to come and, in a near future, it will likely become a very relevant sector. For that reason, all the areas related to elderly care have attracted a growing amount of interest over the last few years.

The sector of elderly care includes a number of areas that could accept automation, contributing, this way, to cost reductions and better service provision. Two areas able to be automated are stability assessment and fall detection. Those two areas are extraordinary relevant for the aforementioned community because they contribute to fall prevention and immediate reaction after fall, key areas for survival, as for this group, 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].

Both fall detection and fall prevention systems use an array of different technologies and, in broad terms, they can be classified in four categories; inertial, radio-frequency, fusion and context-based [3].

Inertial based systems include those ones whose sensors are carried by the monitored person. They assess fall probability based on the information provided by accelerometer and gyroscopic sensors. Leiros-Rodriguez et al. [4] thoroughly review the use of accelerometers as a method of early diagnoses of balance alterations and, therefore, of fall prevention, concluding that methods exploiting accelerometer signal analysis positively influence interventions based on physical exercise, improving balance and preventing falls. On the other hand, Ramachandran et al. [5] review recent advances in the use of accelerometer and gyroscopic sensors and its applications in the field of fall detection.

Radio frequency-based systems include those ones using WiFi or radar signal analysis and most of them are used for fall detection. Different authors [6]-[8], present systems able to process radar signals, both continuous and pulsed, in order to detect human movement and human fall. Other systems [9], [10], use WiFi frequency displacement as a result of human movement to assess fall detection probability.

The fusion block groups all systems that use signals coming from different types of sensors in order to improve performances. These systems are reviewed by Wang et al. [11], concluding that sensor fusion is key to reach optimal performances, as alternative technologies can cover the weak points of a specific one, getting, this way, more robust systems.

Finally, context-based systems include those ones whose sensors use pressure, acoustic, infrared and vision information coming from sensors placed around the monitored person. All systems included in this group have attracted research attention over the course of the last few years as a consequence of the development of artificial neural networks (ANN), as they have been used for signal processing in very different ways. Among all context-based systems, the vision-based ones have a special relevance because of the introduction of artificial vision techniques, which have experienced a major boost during the last decade. In [12], vison-basedfall detection systems are extensively reviewed, concluding that, although their

performances are highly satisfactory, the significant differences between simulated and real falls, and between falls of elderly and young people, documented in [13], [14], as well as the difficulty to access real-world data as a consequence of privacy protection, yield reasonable doubts on the performance of these systems operating in the wild.

Inertial systems, together with context-based ones based on artificial vision techniques are the two blocks of systems that have attracted higher research interest over the last few years. Most of recent research papers of this area belong to one of those two sets of systems as the state-of-the-art chapter will clearly show. In the same chapter, an extensive study regarding all the other technologies is conducted, concluding that the low amount of research effort devoted to their development place them in a position of lower maturity than inertial or artificial vision-based systems. This reality, together with the fact that all commercial fall detection systems are based either on inertial or vision-based technologies, seem to recommend that, until further research proves that those alternative technologies are valid for fall detection in the real world, new system developments should be based on inertial or artificial vision techniques.

The state-of-the-artchapter also shows that, although inertial-based systems present a number of advantages over vison-basedones, they require constantly carrying a sensor that needs to have a functional battery at all times. Under certain circumstances, as the ones identified in the user's needs chapter, these requirements cannot be met and, in those cases, vision-based systems should be considered the best option.

However, all reviewed vison-basedsystems in the state-of-the-artchapter show suboptimal performances under poor illumination conditions as the ones associated to the user's needs presented in the systems requirements chapter. Additionally, today's vison-basedfall detection systems present two important problems, insufficient amount of human falls real world data and privacy protection [12].

The hypothesis of this doctoral thesis is that an artificial vison-basedfall detection system able to overcome these three problems can be implemented. This way, the user's requirements presented in the user's needs chapter will be met.

The main contributions made to the state-of-the-artof vison-basedfall detection systems during the development of the system are the following ones:

- An evaluation of performances of the most significant state-of-the-art two-dimensional (2D) pose estimation architectures able to regress human pose from imagery when the input images are not conventional but Far infrared (FIR). This evaluation is presented in chapter 6.
- A neural network able to determine the projection on the horizontal plane of heels, forefeet and body center of mass (COM) allowing a quick determination of stability indexes and base of support. The theory behind the stability indexes and the design of the network itself is fully described in chapter 7.
- A FIR video dataset made to train the two-dimensional neural network. This data set, the only existing one labeling all human joints both in the 2D and three-dimensional (3D) spaces, includes FIR videoclips and can be used to train human 2D and 3D pose estimation neural networks. It will be released to the research community at the end of the project under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 license. The potential audiences interested in this product include not only the community of researchers of the field of human fall detection, but also other

communities like the human activity recognition, the security and surveillance and the autonomous driving ones.

- A FIR video fall dataset designed to validate the fall detection system. As in the previous case, it is labeled in the 2D and 3D spaces. There is no other fall detection dataset of its kind and will be released to the research community at the end of the project under the same conditions. The potential audiences interested in this product could include not only the community of researchers of the field of human fall detection but also the human activity recognition one.

In addition, an extensive research to determine real user's needs has been carried out. With the exception of the articles by Thilo et al. [15] and Demiris et al. [16], where recommendations to system developers are given, this is the first time such a research is made.

1.1 Motivation

The growing economic significance of the elderly care field invites us to think that certain tasks could be automated reducing this way the burden on caregivers and therefore the cost-of-service provision. Fall detection is one of the tasks suitable for this automation, as proper technology to do it already exists.

Part of this thesis focuses on identifying what is the perception of the different communities integrated in the elderly care field about automatic fall detection, identifying under what circumstances the automation of this task is acceptable and what should be the requirements a system devoted to this purpose must meet.

The study clearly determines when the use of automatic fall detection systems could be accepted and under what circumstances they would be used. After a thorough revision of the state-of-the-artof the field, the main shortfalls of the current systems are identified and it is concluded that no already developed system is able to properly work in the low light enviroments users point out as the most likely ones associated to situations when the use of these systems becomes acceptable.

Therefore, in this thesis, we develop a system fully adapted to the needs specified by the communities of the elderly care field and solve some of the identified shortfalls of the field. In doing so, we intend to contribute to improving the quality of life for both the elderly community and their caregivers by providing a tool well adapted to the necessities they express.

1.2 Structure of this document

This thesis is organized as follows:

Chapter 3 reviews the state-of-the-artof the field of automatic fall detection systems, considering wearable, ambient, and vision-based ones. The chapter concludes with some observations, pointing out advantages, disadvantages, and limitations.

Chapter 4 is devoted to presenting the most extensive research made in this area, which main goal is to provide developers with awareness of the perception that the different communities of the elderly care field have regarding automatic fall detection systems and the requirements that they should meet.

Chapter 5 presents the FIR-Human dataset, the only one of its kind. It includes far infrared video clips recorded by volunteers in different activities, along with the 2D and 3D annotations associated with their joint positions. This chapter also includes a comparative performance

evaluation of the most significant 2D human pose estimation architectures after they have been trained using the FIR-Human dataset.

Chapter 6 presents an alternative approach to the classical use of kinematic descriptors in automatic fall detection systems. For the first time, to the best of our knowledge, we propose the introduction of human dynamic stability descriptors used in other fields to determine whether a fall has occurred. These descriptors approach the human body in terms of balance and stability. This way, differences between real and simulated falls become irrelevant, as all falls are a direct result of a failure in the continuous effort of the body to maintain balance, regardless of other considerations. The descriptors are determined using the information provided by a neural network capable of estimating the body's center of mass and the feet projections onto the ground plane, as well as the feet contact status.

Chapter 7 is centered on the validation of a fall detection system that, using far infrared images, is able to assess whether a fall has occurred based on the use of dynamic descriptors. This system responds to the user's needs identified in Chapter 4 and addresses some of the limitations and shortcomings of present systems.

Chapter 8 is devoted to exposing the main conclusions reached during the development of the system and potential future lines of research.

2 State-of-the-art

2.1 Methods

In this chapter, we focus on establishing the state-of-the-art in the field of automatic human fall detection systems. To fulfill this final goal, a deep review of all published papers present in public databases of research documentation (ScienceDirect, IEEE Explorer, and Sensors databases) has been made. This documentary search was based on different text string searches and was executed from February to December 2020. The period of publication was established between 2015 and 2020, so the latest developments in the field can be identified, and the study serves its purpose of being a guideline for the identification of shortfalls and deficiencies in present systems that should be addressed in new ones.

The terms used in the bibliographical Boolean exploration were "fall detection," and depending on the specific technological block, an additional term was added. This way, for the wearable systems, "wearable" was considered, "ambient" for the case of the ambient block, and "vision" in the last case, so artificial vision technologies were considered. A secondary search was carried out to complete the first one by using other search engines of scholarly literature focused on health (PubMed, MedLine). All searches have been limited to articles and publications in English, the language used by most researchers in this area.

After an initial analysis of papers fulfilling these searching criteria, 77 articles were considered in the area of wearable systems, 40 for the case of ambient technologies, and 81 papers were selected to identify the fall detection state-of-the-art based on artificial vision. This way, 198 articles were considered in the attempt to illustrate how fall detection systems have evolved over the course of the studied period.

The entire process is summarized in the flow diagram shown in figure 1.

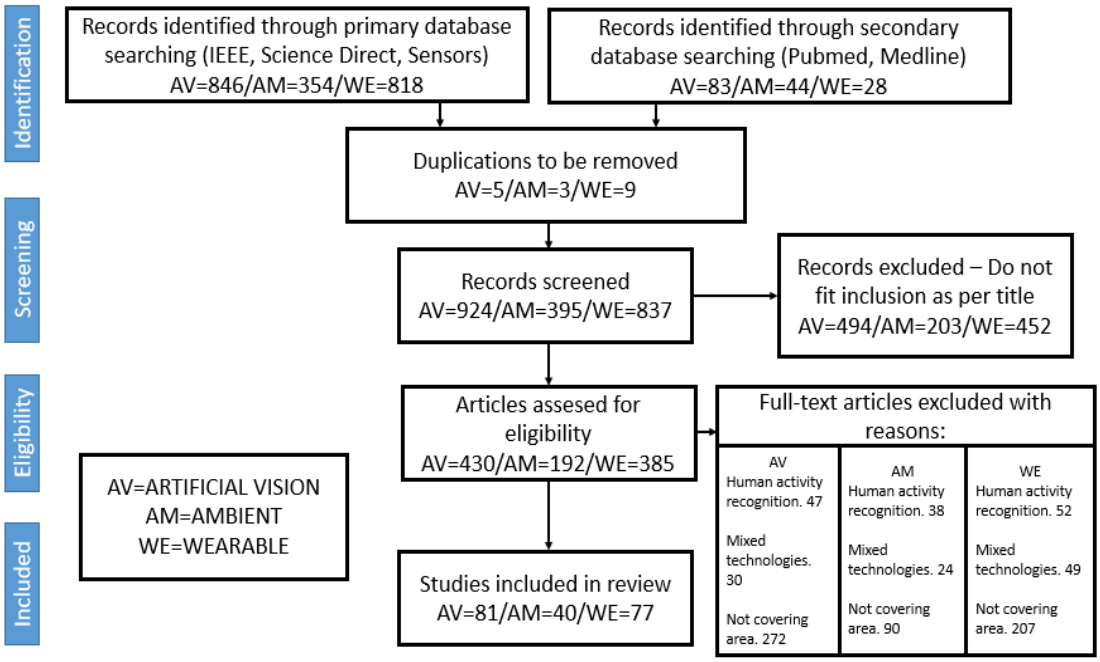


Figure 1. Flow diagram of adopted search and selection strategy for paper selection.

The selection process included an initial screening made through reference management software to guarantee no duplication, and a manual screening, whose objective was making sure the article covered the field, did not fall within the fields of fall prevention or human activity recognition (HAR), and did not mix vision technologies with others.

All selected systems were studied one-by-one to determine their characterization and classification techniques, describing them in-depth in the Discussion Section, so a full taxonomy can be made based on their characteristics. In addition, performance comparisons are also included, so conclusions on which ones are the most suitable systems can be reached.

Additionally, given the crucial importance of power consumption for the specific case of wearable systems, this point is covered in the study of this specific technology. For the case of vision-based technologies, the object tracking algorithms, essential to deliver proper system performances, are also studied.

2.2 Results

The article search and selection process started with an initial identification of 2173 potential articles. Duplicated ones and those whose title clearly did not match the required content were discarded, leaving 852 articles that were assessed for eligibility. These articles were then reviewed and those related to HAR, mixed technologies, and the ones that did not cover the area of automatic human fall detection or gait analysis were discarded. This way, 198 articles are considered in the review: 77 covering wearable systems, 40 covering ambient ones, and 81 covering vision-based technologies.

Tables 1 and 2 group both wearable fall detection and gait analysis systems classified by the type of signal used, with an indication of their performances, while table 3 includes a comparative study of their classification algorithm performances.

Table 1. Fall detection systems classified by type of signal.

SIGNAL TYPE	AUTHOR AND YEAR	SENSOR	PERFORMANCE
ACCELEROMETER	L. Kau et al. (2015) [17]	phone	Specificity - 99.75% Sensibility - 92%
	P. Pierleoni et al. (2015) [18]	accelerometer	Specificity - 100% Sensibility - 80.74%
	P. Kostopoulos et al. (2015) [19]	smartwatch	Specificity - 87.29% Sensibility - 92.18%
	Luca Palmerini et al. (2015) [20]	accelerometer	Specificity - 89.7% Sensibility - 90%
	A. Kurniawan et al. (2016) [21]	accelerometer	Sensibility: - Frontal fall 95% - Backwards fall 75%
	Changhong Wang et al. (2016) [22]	accelerometer	Specificity - 93.2% Sensibility - 97.5%

	T. N. Gia et al. (2016) [23]	accelerometer	Not published
	A. K. Bourke et al. (2016) [24]	accelerometer	Specificity - 87% Sensibility - 88%
	S. Abdelhedi et al. (2016) [25]	accelerometer	Specificity - 99.6% Sensibility - 98.33%
	N. Otanasap et al. (2016) [26]	accelerometer	Specificity - 95.31% Sensibility - 99.48%
	Jian He et al. (2016) [27]	accelerometer	Specificity - 99.1% Sensibility - 93.8%
	A. Sucerquia et al. (2016) [28]	accelerometer	Specificity - 98.75% Sensibility - 95.52%
	M. A. Alvarez de la Concepción et al. (2017) [29]	phone	Specificity – 95%
	P. Jatesiktat et al. (2017) [30]	accelerometer	Specificity - 98.2%
	N. Pannurat et al. (2017) [31]	accelerometer	Specificity - 86.54%
	Putra IPES et al. (2017) [32]	accelerometer	F-score - 98% y 92%
	C. Medrano et al. (2017) [33]	accelerometer	Not published
	R. Shen et al. (2017) [34]	phone	Precision - 79.69%
	D. Yacchirema et al. (2018) [35]	accelerometer	Precision - 93.75% Accuracy - 91.67%
	B. Kaudki et al. (2018) [36]	accelerometer	Specificity - 65% Sensibility - 40%
	A. Sucerquia et al. (2018) [37]	accelerometer	Accuracy -99.4%
	W. Saadeh et al. (2019) [38]	accelerometer	Specificity - 99.1% Sensibility - 97.8%
	Lin Chen et al. (2019) [39]	smartwatch	Specificity –96.36% Sensibility – 99.3%
	A. Shahzad et al. (2019) [40]	phone	Specifity - 88.01% Sensibility - 95.83%
ACCELEROMETER AND GYROSCOPE	H. W. Guo et al. (2015) [41]	Accelerometer & gyroscope	Not published
	H. Jian et al. (2015) [42]	Accelerometer & gyroscope	Specificity - 96.67% Sensibility - 95%

	M. I. Nari et al. (2016) [43]	Accelerometer & gyroscope	Specificity - 86.7% Sensibility - 90%
	T. Sivaranjani et al. (2017) [44]	Accelerometer & gyroscope	Not published
	A. Jefiza et al. (2017) [45]	Accelerometer & gyroscope	Specificity - 99.367% Sensibility - 95.161%
GYROSCOPE	Y. Su et al. (2016) [46]	Gyroscope	Specificity - 98.8% Sensibility - 98.1%
ACCELEROMETER AND PRESSURE SENSOR	A. M. Sabatini et al. (2016) [47]	Accelerometer & pressure sensor	Specificity - 100% Sensibility - 80%
	W. Lu et al. (2020) [48]	Accelerometer & pressure sensor	Specificity - 96.5% Sensibility - 97.5%
PRESSURE SENSOR	J. Light et al. (2015) [49]	Pressure sensor	Precision - 0.889
	W. Lu et al. (2016) [50]	Pressure sensor	Specificity - 90% Sensibility - 94%
INCLINATION SENSOR	J. Sun et al. (2015) [51]	Inclination sensor	Detection rate - 92%
	J. Sun et al. (2016) [52]	Inclination sensor	Detection rate - 85.4%
SOUND SENSOR	M. Cheffena et al. (2016) [53]	Phone	Specificity - 98.46 % Sensibility - 98.97%
EMG SENSOR	A. Leone et al. (2015) [54]	EMG sensor	Specificity - 77.9 % Sensibility - 82.8%
	A. Leone et al. (2015) [55]	EMG sensor	Specificity - 81.5 % Sensibility - 87.3 %
	G. Rescio et al. (2015) [56]	EMG sensor	Specificity - 70.3 % Sensibility - 73.4 %
	V. F. Annese et al. (2016) [57]	EMG and EEG sensors	Not published
	Xugang xi et al. (2017) [58]	EMG sensor	Specificity - 98.59 % Sensibility - 98.7 %

	A. Leone et al. (2017) [59]	EMG sensor	Specificity - 87.1 % Sensibility - 89.1 %
	J. Xiao et al. (2018) [60]	EMG sensor	Specificity - 92.67 % Sensibility - 93.71 %
	Gabriele Rescio et al. (2018) [61]	EMG sensor	Specificity - 89.5% Sensibility - 91.3 %
	G. Mezzina et al. (2019) [62]	EMG and EEG sensors	Specificity - 99.82 % Sensibility - 93.33 %
ECG	Adnan Nadeem et al. (2019) [63]	ECG sensor and accelerometer	Data base generation

Table 2. Gait analysis systems classified by type of signal.

WORKING SIGNAL	**AUTHOR AND YEAR**	**SENSOR**	**PERFORMANCE**
ACCELEROMETER	A. Shahzad et al. (2017) [64]	Accelerometer	Not published
ACCELEROMETER & GYROSCOPE	M. Hemmatpour et al. (2017) [65]	Accelerometer and gyroscope	Accuracy 99.2%
	Mirko Di Rosa et al. (2017) [66]	Accelerometer and gyroscope	Gait analysis
PRESSURE SENSOR	K. Chaccour et al. (2016) [67]	Pressure sensor	Gait analysis
ACCELEROMETER & PRESSURE SENSOR	M. A. Brodie et al. (2015) [68]	Accelerometer and pressure sensor	Gait analysis

Table 3. Machine learning classification algorithms comparative studies.

AUTHOR AND YEAR	**ALGORITHM**	**OPTIMAL ALGORITHM**
R. Igual et al. (2015) [69]	Support Vector Machine (SVM) and K-Nearest Neighbor (K-NN)	SVM
M. Manikandan et al. (2015) [70]	Naïve-Bayes (NB), SVM, Multilayer Perceptron – Artificial Neural Network (MLP - ANN) and Decision Tree (DT)	SVM
A. Lisowska et al. (2015) [71]	SVM, K-NN, Random Forest (RF) and Convolutional Neural Network - Artificial Neural Network (CNN – ANN)	CNN - ANN

M. Vidigal et al. (2015) [72]	MLP - ANN, Radial Basis Function - Artificial Neural Network (RBF - ANN) and Self Organizing Maps - Artificial Neural Network (SOM- ANN)	MLP - ANN
Ryan M. Gibson et al. (2016) [73]	Levenberg–Marquardt back-propagation - Artificial Neural Network (LMB – ANN), K-NN, RBF - ANN, Probabilistic Principal Component (PPC) and Linear Discriminant Analysis (LDA)	K-NN
C. Medrano et al. (2016) [74]	K-NN, Local Outlier Factor (LOF), One-Class SupportVector Machine (O-SVM) and SVM	K-NN
A. T Özdemir (2016) [75]	K-NN, Bayesian decision making (BDM), SVM, least squares method (LSM), dynamic time warping (DTW) and artificial neural networks (ANNs)	K-NN
P. Vallabh et al. (2016) [76]	NB, K-NN, LSM, SVM and ANN	K-NN
M. Ahmed et al. (2017) [77]	SVM, K-NN and ANN	K-NN
B. Ando et al. (2017) [78]	Neuro-Fuzzy Neural Network - Artificial Neural Network (NFNN – ANN) and threshold algorithms	NFNN-ANN
O Aziz et al. (2017) [79]	Logistic Regression (LR), NB, DT, K-NN and SVM	SVM
A. Hakim et al. (2017) [80]	SVM, DT, K-NN and Discriminant Analysis (DA)	SVM
A. Jahanjoo et al. (2017) [81]	Multi-layer Neuro-Fuzzy Neural Network - Artificial Neural Network (MLNFNN – ANN), MLP, K-NN and SVM	MLNFNN – ANN
V. Carletti et al. (2017) [82]	SOM- ANN, O-SVM and One-Class Nearest Neighbor (OCNN)	SOM - ANN
Xugang et al. (2017) [58]	Fisher Linear Discriminant Analysis (FDA), Fuzzy Min-Max Neural Network (FMM-ANN), Gaussian Kernel Fisher Linear Discriminant Analysis (GK-FDA), Gaussian Kernel Support Vector Machine (GK-SVM) and Fuzzy C-means algorithms (FCM)	FMM-ANN
T. Xie et al. (2017) [83]	Extreme Learning Machine – Artificial Neural Network (ELM – ANN), SVM and NB	ELM - ANN
S. B. Khojasteh et al. (2018) [84]	Feed-forward Neural Network (F-ANN) , SVM, DT, Rule-Based Systems (RBS)	SVM

A. Lisowska et al. (2018) [85]	Incremental Support Vector Machine (ISVM), SVM, Increased Nearest Neighbour (INN), K-NN, RF, Recurrent Neural Network – Artificial Neural Network (ReNN-ANN) and CNN – ANN	INN and ReNN-ANN
T. R. Mauldin et al. (2018) [86]	SVM, NB and ReNN -ANN	ReNN - ANN
M. Musci et al. (2018) [87]	ReNN - ANN and threshold algorithms	ReNN - ANN
L. Nguyen et al. (2018) [88]	MLP – ANN, SVM, K-NN and threshold algorithms	MLP-ANN and SVM
J. Ramón et al. (2018) [89]	SVM, K-NN, NB and DT	SVM
A. Chelli et al. (2019) [90]	MLP-ANN, K-NN, quadratic support vector machine (QSVM) y ensemble bagged tree (EBT)	EBT y QSVM
L. Chen et al. (2019) [39]	One-class SVM (OCSVM), SVM, K-NN and ensemble stacked autoencoders - one-class classification based on the convex hull – Artificial Neural Network (ESAEs-OCCCH-ANN)	ESAE-OCCCH-ANN
D. Yacchirema et al. (2019) [91]	LR, Deepnet – Artificial Neural Network (DN-AAN), DT and RF	RF

Table 4 includes all reviewed ambient fall detection systems, as no gait analysis system of this kind has been identified. Table 5 is a compared study of their classification algorithms performances.

Table 4. Ambient fall detection systems by type of signal.

WORKING SIGNAL	AUTHOR AND YEAR	SENSOR	PERFORMANCE
ACOUSTIC	M. Salman et al. (2015) [92]	Group of microphones	AUC – 0.9928 (No interferences)
	A. Díaz-Ramírez et al. (2015) [93]	Microphone	Sensibility=90% (No interferences)
	E. Principi et al. (2016) [94]	Ground acoustic sensor	F_1=98.06%
	D. Droghini et al. (2017) [95]	Ground acoustic sensor	F_1=94.03%
	A. Irtaza et al. (2017) [96]	Microphone	F_1=97.41%

	Syed M. Adnan et al. (2018) [97]	Microphone	F_1=97.41%
	D. Droghini et al. (2019) [98]	Ground acoustic sensor	F_1=93.58%
CONTACT (PRESSURE, DEFORMATION OR CAPACITANCE)	K. Chaccour et al. (2015) [99]	Intelligent carpet	Specificity: 94.9% Sensibility: 88.8%
	Seung-Bae Jeon et al. (2017) [100]	Smart ground	Specificity: 96% Sensibility: 95.5%
	M. Daher et al. (2017) [101]	Smart ground	Sensibility: 94.1%
	Julien Haffner et al. (2018) [102]	Intelligent capacitative ground	Sensibility: 89%
PASSIVE IR (PIR)	W. Chen et al. (2015) [103]	IR sensor	Specificity: 90.75% Sensibility: 95.25%
	Q. Guan et al. (2017) [104]	Pyroelectric IR (PIR) sensor	Specificity: 93% Sensibility: 98%
	X. Fan et al. (2017) [105]	IR sensor	F_1= 99%
	A. Hayashida et al. (2017) [106]	Group of IR sensors	Precision = 94%
	J. Adolf et al. (2018) [107]	IR sensor	Specificity: 93% Sensibility: 85%
	Y. Ogawa et al. (2020) [108]	Group of IR sensors	Precision: 95.75%
	Z. Liu et al. (2020) [109]	Group of IR sensors	F_1= 96%
RADAR	B. Y. Su et al. (2015) [110]	Doppler radar	Specificity: 92.2% Sensibility: 97.1%
	C. Garripoli et al. (2015) [111]	Doppler radar	Sensibility = 100%
	B. Erol et al. (2017) [112]	Range-doppler radar	Precision = 95.95%

	H. Sadreazami et al. (2018) [113]	Ultra wide band radar	Precision: 95.04% Sensibility: 88.46%
	S. Chen et al. (2019) [114]	Low frequency and PRF pulsed radar	Precision = 99.346%
	Y. Sun et al. (2019) [115]	FMCW radar	F_1= 98.9%
	H. Sadreazami et al. (2019) [116]	Ultra wide band radar	Precision = 96.12%
	Y. Shankar et al. (2019) [117]	FMCW radar	Precision = 99.5%
	A. Chelli et al. (2019) [118]	RF simulator	Specificity: 100% Sensibility: 100%
	H. Sadreazami et al. (2019) [119]	Ultra wide band radar	Precision = 96.15%
	C. Ding et al. (2019) [120]	FMCW radar	Precision = 95.5%
	H. Sadreazami et al. (2019) [121]	Ultra wide band radar	Specificity: 91.78% Sensibility: 90.38%
	H. Sadreazami et al. (2020) [122]	Ultra wide band radar	Precision: 95.28% Specificity: 92.91%
	A. Bhattacharya et al. (2020) [123]	FMCW radar	Precision: 95%
	H. Sadreazami et al. (2020) [124]	Doppler radar	Specificity: 91.67% Sensibility: 93.44%
WIFI	H. Wang et al. (2017) [125]	Wifi system	Specificity: 92% Sensibility: 89%
	H. Cheng et al. (2019) [126]	Wifi system	F_1=96%
	M. Huang et al. (2019) [127]	Wifi system	Precision: 95% FNR: 2.44%
	Y. Hu et al. (2020) [128]	Wifi system	Sensibility: 96%

	M. Keaton et al. (2020) [129]	Wifi system	Precision: 81.8%
	T. H. Nguyen et al. (2020) [130]	Wifi system	Precision: 66.03%
	J. Ding et al. (2020) [131]	Wifi system	Precision: 93%

Table 5. Classification algorithms of ambient fall detection systems.

WORKING SIGNAL	AUTHOR AND YEAR	ALGORITHMS	OPTIMAL ALGORITHM
ACOUSTIC	M. Salman et al. (2015)) [92]	OCSVM	
	A. Díaz-Ramírez et al. (2015) [93]	Dynamic Time Warping (DTW)	
	E. Principi et al. (2016) [94]	Gaussian Mean Supervectors (GMS) + SVM	
	D. Droghini et al. (2017) [95]	Gaussian Mean Supervectors (GMSs) + OCSVM	
	A. Irtaza et al. (2017) [96]	SVM	
	Syed M. Adnan et al. (2018) [97]	SVM	
	D. Droghini et al. (2019) [98]	Siamese Autoencoder Neural Network + K-NN, SVM, OCSVM y Siamese Neural Network + K-NN	Siamese Autoencoder Neural Network + K-NN
	A. Collado et al. (2017) [132]	DT, K-NN, LG, NB, rule based classifier PART, RF and SVM	RF
CONTACT (PRESSURE,	K. Chaccour et al. (2015) [99]	Rule based (RB)	

DEFORMATION OR CAPACITANCE)	Seung-Bae Jeon et al. (2017) [100]	Bayesian decision making (BDM)	
	M. Daher et al. (2017) [101]	For tiles : RB	
	Julien Haffner et al. (2018) [102]	K-NN, SVM and LS	SVM
PASSIVE IR	W. Chen et al. (2015) [103]	K-NN	
	Q. Guan et al. (2017) [104]	K-NN, Gaussian Mixture- Hidden Markov Model (GM-HMM), NB, SVM	SVM
	X. Fan et al. (2017) [105]	long short term memory artificial neural network (LSTM-ANN), gated recurrent unit artificial neural network (GRU-ANN) and Multilayer Perceptron – Artificial Neural Network (MLP - ANN)	LST-ANN
	A. Hayashida et al. (2017) [106]	DT	
	J. Adolf et al. (2018) [107]	Inception-v3 Artificial Neural Network (Iv3 – ANN)	
	Y. Ogawa et al. (2020) [108]	LDA, K-NN, SVM, NB, AdaBoost, RF, Voting and Bagging	Voting
	Z. Liu et al. (2020) [109]	RF	
RADAR	B. Y. Su et al. (2015) [110]	K-NN and SVM	K-NN

	C. Garripoli et al. (2015) [111]	Least Square- SVM (LS-SVM)	
	B. Erol et al. (2017) [112]	SVM	
	H. Sadreazami et al. (2018) [113]	LSTM – ANN, SVM, K-NN and DTW	LSTM-ANN
	S. Chen et al. (2019) [114]	Convolutional Neural network (CNN)	
	Y. Sun et al. (2019) [115]	LSTM-ANN and 3-D CNN	LSTM-ANN
	H. Sadreazami et al. (2019) [116]	Capsule Nework – Artificial Neural Network (CN-ANN), SVM, decision tree (DT), Gaussian naive Bayes (GNB), multi-layer perceptron (MLP) and CNN-ANN	CN-ANN
	Y. Shankar et al. (2019) [117]	Deep convolution neural networks – Artificial Neural Network (DCNN-ANN)	
	A. Chelli et al. (2019) [118]	K-nearest neighbors (KNN), decision tree (DT), artificial neural network (ANN) and support vector machine (SVM)	SVM
	H. Sadreazami et al. (2019) [119]	Gaussian support vector machine (GSVM), K-nearest neighbors (KNN), multi-layer perception (MLP)	DRN-ANN

		and dynamic time warping (DTW) and Deep Residual Network (DRN-ANN)	
	C. Ding et al. (2019) [120]	K-NN	
	H. Sadreazami et al. (2019) [121]	CNN, SVM, K-NN	CNN
	H. Sadreazami et al. (2020) [122]	SVM, K-NN, DT y Linear Discriminant Analysis (LDA)	K-NN
	A. Bhattacharya et al. (2020) [123]	CNN	
	H. Sadreazami et al. (2020) [124]	multi-layer perceptron (MLP), k-nearest neighbors (KNN), dynamic time warping (DTW), KNN-DTW, long-short-term-memory artificial neural networks (LSTM-ANN) and deep convolutional neural network (DCNN)	DCNN
WIFI	H. Wang et al. (2017) [125]	SVM	
	H. Cheng et al. (2019) [126]	LSTM-ANN, Gated Recurrent Unit artificial neural network (GRU-ANN) and CNN	GRU-ANN
	M. Huang et al. (2019) [127]	Dynamic template matching (DTM)	
	Y. Hu et al. (2020) [128]	DTW	

	M. Keaton et al. (2020) [129]	fully-connected neural networks FCNN and CNN	FCNN
	T. H. Nguyen et al. (2020) [130]	CNN and LSTM-ANN	LSTM-ANN
	J. Ding et al. (2020) [131]	Hidden Markov Model (HMM), Long Short-Term Memory (LSTM) , Random Forest (RF), Support Vector Machine (SVM) and Recurrent Neural Network – Artificial Neural Network (ReNN-ANN)	ReNN-ANN

Finally, all reviewed vision-based fall detection systems are included in table 6, together with their performances. Table 7 compares their performances and table 8 reflects all system performance evaluation datasets.

Table 6. . Reviewed vision-based fall detection systems.

Reference	Year	Characterization (Global/Local/Depth)	Classification	Input Signal	Used Datasets	Performance
A. Yajai et al. [133]	2015	Skeleton joint tracking model provided by MS Kinect® is used to track joints and build a 2D and 3D bounding box around the body/depth characterization	Feature-threshold-based. • Height/width ratio of the bounding box • center of gravity (COG) position in relation to support polygon (defined by ankle joints)	Depth	This system-specific video dataset—no public access at revision time	Accuracy 98.43% Specificity 98.75% Recall 98.12%
C. -J. Chong et al. [134]	2015	Pixel clustering and background (Horprasert)/global characterization	Feature-threshold-based. Method 1: • Bounding box (BB) aspect ratio • CG position Method 2: • Ellipse orientation and aspect ratio • Motion history image (MHI)	Red-green-blue (RGB)	Specific video dataset—no public access at revision time	Method 1 Sensitivity 66.7% Specificity 80% Method 2 Sensitivity 72.2% Specificity 90%
H. Rajabi et al. [135]	2015	Foreground extraction through background subtraction (Gaussian mixed models—GMM) and Sobel filter application/ global characterization	Feature-threshold-based. • BB orientation angle • Change of COG width • Height/width relation of contour • Hu moment invariants	RGB	This system-specific video dataset—no public access at revision time	Fall detection success rate 81%

L. H. Juang et al. [136]	2015	Foreground extraction through background subtraction (optical flow-based) and human joints identified/global characterization	Support vector machine (SVM)	RGB	This system-specific video dataset—no public access at revision time	Accuracy up to 100%
M. A. Mousse al. [137]	2015	Foreground extraction through pixel color and brightness distortion determination and integration of foreground maps through homography/global characterization	Feature-threshold-based. Ratio observed silhouette area/silhouette area projected on the ground plane	RGB—2 ORTHOGONAL VIEWS	Multicam Fall Dataset [138]	Sensitivity 95.8% Specificity 100%
Muzaffer Aslan et al. [139]	2015	Human silhouette is segmented using depth information, and curvature scale space (CSS) is calculated and encoded in a Fisher vector/depth characterization	SVM	Depth	SDUFall [140]	Average accuracy 88.01%
Z. Bian et al. [141]	2015	Silhouette extraction by using depth information. Human body joints identified and tracked with torso rotation/depth characterization	SVM	Depth	This system-specific video dataset—no public access at revision time	Sensitivity 95.8% Specificity 100%
C. Lin et al. [142]	2016	Foreground extraction through background subtraction (GMM)/global characterization	Feature-threshold-based. • Ellipse orientation • Linear and angular acceleration • MHI	RGB	This system-specific video dataset—no public access at revision time	Not published
F. Merrouche et al. [143]	2016	Foreground extraction by using the difference between depth frames and head tracking through particle filter/depth characterization	Feature-threshold-based. • Ratio head vertical position/person height	Depth	SDUFall [140]	Sensitivity 90.76% Specificity 93.52%

			• COG velocity			Accuracy 92.98%
K. G. Gunale et al. [144]	2016	Foreground extraction through background subtraction (direct comparison)/global characterization	K-nearest neighbor (KNN)	RGB	Chute dataset—no public access at revision time	Accuracy Fall 90% No fall 100%
K. R. Bhavya et al. [145]	2016	Foreground extraction through background subtraction (direct comparison)/global characterization + optical flow (OF)/global characterization	KNN on MHI and OF features	RGB	This system-specific video dataset—no public access at revision time	Not published
Kun Wang et al. [146]	2016	Segmentation through vibe [147] and histogram of oriented gradients (HOG) and local binary pattern (LBP)/global characterization + feature maps obtained through convolutional neural network (CNN)/ local characterization	SVM-linear kernel	RGB	Multicam Fall Dataset [138] and SIMPLE Fall Detection Dataset [148] and This system-specific video dataset—no public access at revision time	Sensitivity 93.7% Specificity 92%
U. Pratap et al. [149]	2016	Foreground extraction through background subtraction (GMM)/global characterization	Feature-threshold-based. • Silhouette COG stationary over a threshold time limit	RGB	Specific video datasets—no public access at revision time	Fall detection rate 92% False alarm rate 6.25%
X. Wang et al. [150]	2016	Segmentation through vibe [147] and upper body database populated and sparse OF determined/global characterization	Feature-threshold-based. • Body ratio width/height • Vertical velocity derived from OF • Upper body position history	RGB	LE2I [151]	Average precision 81.55%

A. Y. Alaoui et al. [152]	2017	Foreground extraction through background subtraction (direct comparison)/global characterization + OF/global characterization	No classification algorithm reported	RGB	CHARFI2012 Dataset [153]	Precision 91% Sensitivity 86.66%
Apichet Yajai et al. [154]	2017	Skeleton joint tracking model provided by MS Kinect®/depth characterization	Feature-threshold-based. Aspect ratios: • Bounding box • COG • Bounding box diagonal vs. max. height • Bounding box height vs. max. height	Depth	This system-specific video dataset—no public access at revision time	Accuracy 98.15% Sensitivity 97.75% Specificity 98.25%
B. Lewandowski et al. [155]	2017	Voxels around the point cloud are calculated. The ones classified as human are clustered, and IRON features are calculated/local characterization	Feature-threshold-based. • Mahalanobis distance between cluster IRON features and the distribution of IRON features from fallen bodies	Depth	This system-specific video dataset—no public access at revision time	Sensitivity in operational environments 99%
F. Harrou et al. [156]	2017	Foreground extraction through background subtraction (direct comparison)/depth characterization	Multivariate exponentially weighted moving average (MEWMA)-SVM KNN Artificial neural network (ANN) Naïve Bayes (NB)	RGB	UR Fall Detection [157] & Fall Detection Dataset [158]	Accuracy KNN 91.94% ANN 95.15% NB 93.55% NEWMA-SVM 96.66%

G. M. Basavaraj et al. [159]	2017	Foreground extraction through background subtraction (median)/global characterization	Feature-threshold-based. • Ellipse eccentricity and orientation • MHI	RGB	This system-specific video dataset—no public access at revision time	Accuracy Fall 86.66% Non-fall 90%
K. Adhikari et al. [158]	2017	Foreground extraction through background subtraction (direct comparison) using both RGB techniques and depth ones and Feature maps obtained through CNN/local and depth characterization	Softmax based on features vector from CNN	Depth	This system-specific video dataset—no public access at revision time	Overall, accuracy 74% System sensitivity to lying pose 99%
Koldo De Miguel et al. [160]	2017	Foreground extraction through background subtraction (GMM) + Sparse OF determined/global characterization	KNN on silhouette and OF features	RGB	This system-specific video dataset—no public access at revision time	Accuracy 96.9% Sensitivity 96% Specificity 97.6%
Leiyue Yao et al. [161]	2017	Skeleton joint tracking model provided by MS Kinect®/depth characterization	Feature-threshold-based • Torso angle • Centroid height	Depth	This system-specific video dataset—no public access at revision time	Accuracy 97.5% True positive rate 98% True negative rate 97%
M. Antonello et al. [162]	2017	Voxels around the point cloud are calculated. Then they are segmented in homogeneous patches and the ones classified as human are gathered and classified or not as a human lying body/depth characterization	SVM—radial-based kernel	Depth	IASLAB-RGBD fallen person Dataset [163]	Set A Accuracy: single view (SV) 0.87/SV+map verification (MV) 0.92

						Precision: SV 0.73/SV+MV 0.85 Recall: SV 0.85/SV+MV 0.85 Set B Accuracy: SV 0.88/SV+MV 0.9 Precision: SV 0.8/SV+MV 0.87 Recall: SV 0.86/SV+MV 0.81
M. N. H. Mohd et al. [164]	2017	Skeleton joint tracking model provided by MS Kinect® is used to determine joint positions and speeds/depth characterization	SVM based on joints speeds and rule-based decision-based on joints position in relation to knees	Depth	TST Fall Detection [165] , UR Fall Detection [157] and Falling Detection [166]	Accuracy 97.39% Specificity 96.61% Sensitivity 100%
N. B. Joshi et al. [167]	2017	Foreground extraction through background subtraction (GMM)/global characterization	Feature-threshold-based. • BB width/height ratio • COG position	RGB	LE2I [151]	Specificity 92.98% Accuracy 91.89%

			• Orientation • Hu moments			
N. Otanasap et al. [168]	2017	Skeleton joint tracking model provided by MS Kinect®/depth characterization	Feature-threshold-based. • Head velocity • CG position in relation to ankle joints	Depth	This system-specific video dataset—no public access at revision time	Sensitivity 97% Accuracy 100%
Q. Feng et al. [169]	2017	CNN is used to detect and track people, and Sub-MHI are correlated to each person BB/local characterization	SVM	RGB	UR Fall Detection [157]	Precision 96.8% Recall 98.1% F_1 97.4%
S. Hernandez-Mendez et al. [170]	2017	Foreground extraction through background subtraction (direct comparison) and silhouette tracking. Then centroid and features are determined/depth characterization	Feature-threshold-based. • Angles and ratio height/width of the BB	Depth	Depth And Accelerometric Dataset [171] and this system-specific video dataset—no public access at revision time	The fallen pose is detected correctly on 100% of occasions.
S. Kasturi et al. [172]	2017	Foreground extraction through background subtraction (direct comparison)/depth characterization	SVM	Depth	UR Fall Detection [157]	Sensitivity 100% Specificity 88.33%
S. Kasturi et al. [173]	2017	Foreground extraction through background subtraction (direct comparison)/depth characterization	SVM	Depth	UR Fall Detection [157]	Accuracy Total testing accuracy 96.34%

S. Pattamaset et al. [174]	2017	Body vector construction and CG identification taking as starting point 16 parts of the human body/depth characterization	Feature-threshold-based. • COG acceleration • Body vector/vertical angle	Depth	This system-specific video dataset—no public access at revision time	Accuracy 100%
Sajjad Taghvaei et al. [175]	2017	Foreground extraction through background subtraction/depth characterization	Hidden Markov model (HMM)	Depth	This system-specific video dataset—no public access at revision time	Accuracy 84.72%
Y. M. Galvão et al. [176]	2017	Median square error (MSE) every 3 frames/global characterization	Multilayer perceptron (MLP) KNN SVM—polynomial kernel	RGB	UR Fall Detection [157]	F1 score: MLP 0.991 KNN 0.988 SVM—polynomial kernel 0.988
Thanh-Hai Tran et al. [177]	2017	Skeleton joint tracking model provided by MS Kinect®/depth characterization or Motion map extraction from RGB images and gradient kernel descriptor calculated/global characterization	Feature-threshold-based. • Height of hip joint • Vertical body velocity Or • SVM classification	Depth or RGB	UR Fall Detection [157] and LE2I [151] and Multimodal Multiview Dataset of Human Activities [178]	UR Dataset Sensitivity 100% Specificity 99.23% LE2I Dataset Sensitivity 97.95% Specificity 97.87%

						MULTIMODAL Dataset (Average) Sensitivity 92.62% Specificity 100%
X. Li et al. [179]	2017	Foreground extraction through background subtraction (direct comparison) and feature maps obtained through CNN/ local characterization	Softmax based on features vector from CNN	RGB	UR Fall Detection [157]	Sensitivity 100% Specificity 99.98% Accuracy 99.98%
Yaxiang Fan et al. [180]	2017	Feature maps obtained through CNN from dynamic images/local characterization	Classification made by fully connected last layers of CNNs	RGB	Multicam Fall Dataset [138] & LE2I [151], High-Quality Dataset [181] and This system-specific video dataset—no public access at revision time	Sensitivity LE2I 98.43% Multicam 97.1% HIGH-QUALITY FALL SIM 74.2% SYSTEM Dataset 63.7%
A. Abobakr et al. [182]	2018	Silhouette extraction by using depth information. A feature vector of different body pixels based on depth difference between pairs of	Random decision forest for pose recognition and SVM for movement identification	Depth	UR Fall Detection [157] and CMU Graphics Lab—motion capture library [183]	Accuracy 96% Precision 91%

		points is created/depth characterization				Sensitivity 100% Specificity 93%
B. Dai et al. [184]	2018	Foreground extraction through background subtraction (direct comparison)/global characterization	Feature-threshold-based. • BB segmented areas occupancy. • COG/height ratio • COG vertical speed	RGB	UR Fall Detection [157] and This system-specific video dataset—no public access at revision time	Sensitivity 95% Specificity 96.7%
Georgios Mastorakis et al. [185]	2018	Depth images are used to determine head velocity profile/depth characterization	Feature-threshold-based. • Hausdorff distance between real head velocity profile and database ones	Depth	Specific video dataset developed for [171] (A) and [140] (B)– no public access at revision time	A Dataset Sensitivity 100% Specificity 100% B Dataset Sensitivity 90.88% Specificity 98.48%
K. Sehairi et al. [186]	2018	Foreground extraction through background subtraction (self-organizing maps) and feature extraction associated with each silhouette/global characterization	• SVM-radial basis function (SVM-RBF) • KNN • Fully connected ANN trained through background propagation ANN	RGB	LE2I [151]	Accuracy SVM-RBF 99.27% KNN 98.91% ANN 99.61%

Kun-Lin Lu et al. [187]	2018	Person detection through CNN YoLOv3 and feature extraction of the generated bounding box/local characterization	Feature-threshold-based • Bounding box height evolution in 1.5 s periods	RGB	This system-specific video dataset—no public access at revision time	Recall 100% Precision 93.94% Accuracy 95.96%
Leila Panahi et al. [188]	2018	Foreground extraction through background subtraction (depth information) and silhouette tracking. Then ellipse is established around the silhouette, and features are determined/depth characterization	SVM & Threshold-based decision • Centroid elevation • Centroid speed • Ellipse aspect ratio	Depth	Depth and Accelerometric Dataset [171]	Average results SVM Sensitivity 98.52% Specificity 97.35% Threshold-based decision Sensitivity 98.52% Specificity 97.35%
M. Rahnemoonfar et al. [189]	2018	Feature maps obtained through CNN/depth characterization	Softmax based on features vector from CNN	Depth	SDUFall [140]	Accuracy 97.58%
Manola Ricciuti et al. [190]	2018	Foreground extraction through background subtraction (direct comparison)/depth characterization	SVM	Depth	This system-specific video dataset—no public access at revision time	Accuracy 98.6%

Myeongseob Ko et al. [191]	2018	Depth map from monocular images and silhouette detection through particle swarm optimization/global characterization	Feature-threshold-based • Vertical velocity • BB aspect ratio • BB height • Top depth/bottom depth ratio	RGB	This system-specific video dataset—no public access at revision time	Accuracy 97.7% Sensitivity 95.7% Specificity 98.7%
Syed F. Ali et al. [192]	2018	Foreground extraction through background subtraction (GMM)/global characterization	Boosted J48	RGB	UR Fall Detection [157] Mand Multicam Fall Dataset [138]	Accuracies Multicam (2 classes) 99.2% Multicam (2 classes) 99.25% UR FALL 99%
W. Min et al. [193]	2018	Skeleton joint tracking model provided by MS Kinect® is used to estimate vertical/torso angle/depth characterization	SVM	Depth	TST Fall Detection [165]	Accuracy 92.05%
W. Min et al. [194]	2018	Object recognition through CNN and features of human shape sorted out as well as their spatial relations with furniture in the image/local characterization	Automatic engine classifier based on similarities (minimum quadratic error) between real-time actions and activity class features	RGB	This system-specific video dataset—no public access at revision time and UR Fall Detection [157]	Precision 94.44% Recall 94.95% Accuracy 95.5%
X. ShanShan et al. [195]	2018	Foreground extraction through background subtraction (GMM)/global characterization	SVM-radial kernel	RGB	Center For Digital Home Dataset– MMU [196]	Sensitivity 96.87% Accuracy 86.79%

Amal El Kaid et al. [197]	2019	Feature maps obtained through convolutional layers of a CNN/local characterization	Softmax based on features vector from CNN	RGB	This system-specific video dataset—no public access at revision time	Reduces false positives of angel assistance system by 17% by discarding positives assigned to people in a wheelchair
Chao Ma et al. [198]	2019	Face masking to preserve privacy and feature maps obtained through CNN/local characterization	Autoencoder SVM	RGB + IR	UR Fall Detection [157] and Multicam Fall Dataset [138] and Fall Detection Dataset [158] and This system-specific video Dataset—no public access at revision time	Autoencoder Sensitivity 93.3% Specificity 92.8% SVM Sensitivity 90.8% Specificity 89.6%
D. Kumar et al. [199]	2019	Silhouette segmentation by edge detection through HOG/global characterization + silhouette center angular velocity determined by long short-term memory (LSTM) model/local characterization	feature-threshold-based. • Silhouette center point angular velocity	RGB	MOT Dataset [200] and UR Fall Detection [157] and COCO Dataset [201]	Accuracy 98.1%

F. Harrou et al. [202]	2019	Foreground extraction through background subtraction (direct comparison)/global characterization	SVM • Linear kernel • Polynomial kernel • Radial kernel	RGB	UR Fall Detection [157] & Fall Detection Dataset [158]	Accuracy: Linear kernel 93.93% Polynomial kernel 94.35% Radial kernel 96.66%
J. Brieva et al. [203]	2019	Feature maps obtained through CNN from OF/ local characterization	Softmax based on features vector from CNN	RGB	This system-specific video dataset—no public access at revision time	Precision 95.27% Recall 95.42% F_1 95.34%
M. Hua et al. [204]	2019	Human keypoints identified by OpenPose (convolutional pose machines and human body vector construction) and recurrent neural network (RNN)-LSTM ANN used for pose prediction/local characterization	Fully connected layer	RGB	LE2I [151]	Precision 90.8% Recall 98.3% F_1 0.944
M. M. Hasan et al. [205]	2019	Human keypoints identified by OpenPose (convolutional pose machines and human body vector construction) and RNN-LSTM ANN/local characterization	Softmax based on features vector from RNN-LSTM	RGB	UR Fall Detection [157] & Fall Detection Dataset [158] & Multicam Fall Dataset [138]	URFD Sensitivity 99% Specificity 96% FDD Sensitivity 99% Specificity 97%

						Multicam Sensitivity 98% Specificity 96%
P. K. Soni et al. [206]	2019	Foreground extraction through background subtraction (GMM)/global characterization	SVM	RGB	UR Fall Detection [157]	Specificity 97.1% Sensitivity 98.15%
Ricardo Espinosa et al. [207]	2019	OF extracted from 1-s windows/global characterization + Feature maps obtained through CNN/local characterization	• Softmax based on features vector from CNN • SVM • Random forest (RF) • MLP • KNN	RGB	UPFALL [208]	Sensitivity Softmax 97.95% SVM 14.1% RF 14.3% MLP 11.03% KNN 14.35%
S. Kalita et al. [209]	2019	BBs established in hands, head and legs through extended core9 framework/local characterization	SVM	RGB	UR Fall Detection [157]	Sensitivity 93.33% Specificity 95% Accuracy 94.28%
Saturnino Maldonado -Bascón et al. [210]	2019	Person detection through CNN YoLOv3 and feature extraction of the generated BB /local characterization	SVM	RGB	IASLAB-RGBD fallen person dataset [163] and This system-specific video dataset—no public access at revision time	Average results Precision 88.75%

						Recall 77.7%
X. Cai et al. [211]	2019	OF/global characterization + Wide residual network/local characterization	Softmax classifier implemented in the last layer of the ANN	RGB	UR Fall Detection [157]	accuracy 92.6%
Xiangbo Kong et al. [212]	2019	Segmentation by model provided by MS Kinect® + depth map and CNN used for feature maps creation/depth characterization	Softmax based on features vector from CNN implemented in its last layer	Depth	This system-specific video dataset—no public access at revision time	Depending on the camera height accuracy, results between 80.1% and 100% are obtained
Xiangbo Kong et al. [213]	2019	Foreground extraction through background subtraction (Depth information) and HOG is calculated as a classifying feature	SVM-linear kernel	Depth	This system-specific video Dataset—no public access at revision time	Sensitivity 97.6% Specificity 100%
A. CARLIER et al. [214]	2020	Dense OF/global characterization + feature maps obtained through CNN/ local characterization	Fully connected layer	RGB	UR Fall Detection [157] and Multicam Fall Dataset [138] and LE2I [151]	Sensitivity 86.2% False discovery rate 11.6%
B. Wang et al. [215]	2020	Human keypoints identified by OpenPose (convolutional pose machines and human body vector construction) and followed by DeepSORT (CNN able to track numerous objects simultaneously)/local characterization	Classifiers are used to sort out falling state and fallen state • Gradient boosted tree (GDBT) • Decision tree (DT) • RF	RGB	UR Fall Detection [157] & Fall Detection Dataset [158] & LE2I [151]	F1-score Falling state GDBT 95.69% DT 84.85% RF 95.92%

			• SVM • KNN • MLP			SVM 96.1% KNN 93.78% MLP 97.41% Fallen state GDBT 95.27% DT 95.45% RF 96.8% SVM 95.22% KNN 94.22% MLP 94.46%
C. Menacho et al. [216]	2020	Dense OF/global characterization and feature maps obtained through CNN/ local characterization	Fully connected layer	RGB	UR Fall Detection [157]	Accuracy 88.55%
C. Zhong et al. [217]	2020	Binarization based on IR threshold + edge identification/global characterization + feature maps obtained through convolutional layers of an ANN/local characterization	Based on features maps from CNN: • Radial basis function neural network (RBFNN) • SVM • Softmax • DT	IR	This system-specific video dataset—no public access at revision time	Multi-occupancy scenarios F1 score: RBFNN 89.57 (+/−0.62) SVM 88.74% (+/−1.75) Softmax 87.37% (+/−1.4)

						DT 88.9% (+/−0.68)
G. Sun et al. [218]	2020	pose estimation through OpenPose (convolutional pose machines and human body vector construction) and single-shot multibox detector-MobileNet (SSD-MobileNet)/local characterization	• Support vector data description (SVDD) • SVM • KNN	RGB	COCO Dataset [201] and a specific video dataset—no public access at revision time	Sensitivity SVM 92.5% KNN 93.8% SVDD 94.6%
J. Liu et al. [219]	2020	Local binary pattern histograms from three orthogonal planes (LBP-TOP) applied over optical Flow after robust principal component analysis (RPCA) techniques have been applied over incoming video signals.	Sparse representations classification (SRC)	RGB	UR Fall Detection [157] & Fall Detection Dataset [158]	Accuracy: FDD dataset 98% URF dataset 99.2%
J. Thummala et al. [220]	2020	Foreground extraction through background subtraction (GMM)/global characterization	Feature-threshold-based. Object height/width ratio, ratio change speed and MHI.	RGB	LE2I [151]	Accuracy 95.16%
Jin Zhang et al. [221]	2020	Human keypoints identified by CNN (convolutional pose machines and human body vector construction)/local characterization	Logistic regression classifier based on : • Rotation energy sequence • Generalized force sequence	RGB	This system-specific video dataset—no public access at revision time	Fall detection rate 98.7% False alarm rate 1.05%
K. N. Kottari et al. [222]	2020	Segmentation through vibe [147] and illumination change-resistant algorithm (ICA) [223] then main silhouette axis determination	Feature-threshold-based. • Silhouette main axis angle with vertical axis	RGB	This system-specific video dataset—no public access at revision time and PIROPO [224]	Specific database accuracy ICA—87%–96.34%

						VIBE—78.05%–86.5% PIROPO—ICA Walk accuracy 95% Seat accuracy 98.65%
Qi Feng et al. [225]	2020	Feature maps obtained through convolutional layers of a CNN and LSTM/local characterization	Softmax based on features vector from ANN implemented in its last layer	RGB	Multicam Fall Dataset [138], UR Fall Detection [157] and this system-specific video dataset—no public access at revision time	Multicam Dataset Sensitivity 91.6% Specificity 93.5% UR Dataset Precision 94.8% Recall 91.4% THIS SYSTEM Dataset Precision 89.8% Recall 83.5%

Qingzhen Xu et al. [226]	2020	Human keypoints identified by OpenPose (convolutional pose machines and human body vector construction) and CNN used for feature maps creation/local characterization	Softmax based on features vector from CNN implemented in its last layer	RGB	UR Fall Detection [157] and Multicam Fall Dataset [138] and NTU RGB+D Dataset [227]	Accuracy rate 91.7%
Swe N. Htun et al. [228]	2020	Foreground extraction through background subtraction (GMM)/global characterization	Hidden Markov model (HMM) based on Observable data : • Silhouette surface • Centroid height • Bounding box aspect ratio	RGB	LE2I [151]	Precision 99.05% Recall 98.37% Accuracy 99.8%
T. Kalinga et al. [229]	2020	Skeleton joint tracking model provided by MS Kinect® is used to determine joint speeds and angles of different body parts/depth characterization	Feature-threshold-based. • Joint speeds and angles of body parts	Depth	This system-specific video dataset—no public access at revision time	Accuracy 92.5% Sensitivity 95.45% Specificity 88%
Weiming Chen et al. [230]	2020	Human keypoints identified by OpenPose (convolutional pose machines and human body vector construction)/local characterization	Feature-threshold-based • Hip vertical velocity • Spine/ground plane angle • BB aspect ratio	RGB	This system-specific video dataset—no public access at revision time	Accuracy 97% Sensitivity 98.3% Specificity 95%
X. Cai et al. [231]	2020	Feature maps obtained through hourglass convolutional auto-encoder (HCAE) ANN/local characterization	Softmax based on features vector from HCAE	RGB	UR Fall Detection [157]	Sensitivity 100% Specificity 93% Accuracy 96.2%

						URFD Precision 0.897 Recall 0.813 F_1 0.852 Specific dataset Precision 0.981 Recall 0.923 F_1 0.948
Yuxi Chen et al. [233]	2020	Feature maps obtained through 3 different CNNs (LeNet, AlexNet y GoogLeNet)/depth characterization	Classification made by fully connected last layers of CNNs	Depth	Video dataset developed for the system in [212]	Average values Lenet Sensitivity 82.78% Specificity 98.07% AlexNet Sensitivity 86.84% Specificity 98.41% GoogLeNet Sensitivity 92.87%

						Specificity 99%
X. Wang et al. [234]	2020	Feature maps obtained through convolutional layers of an ANN/local characterization	Logistic function to identify frame-by-frame two classes in the prediction layer (person and fallen)	RGB	UR Fall Detection [157] & Fall Detection Dataset [158]	Average precision (AP) for fallen 0.97 mean average precision (mAP) for both classes 0.83

Table 7. Vision-based system performance comparison.

Reference	Year	Input Signal	ANN/Classifiers and Performance			
C. -J. Chong et al. [134]	2015	RGB	Method 1 BB aspect ratio and CG position			
			Sensitivity 66.7%			
			Specificity 80%			
			Method 2 Ellipse orientation and aspect ratio + MHI			
			Sensitivity 72.2%			
			Specificity 90%			
F. Harrou et al. [156]	2017	RGB		Accuracy	Sensitivity	Specificity
			KNN	91.94%	100%	86.00%
			ANN	95.15%	100%	91.00%
			NB	93.55%	100%	88.60%
			MEWMA-SVM	**96.66%**	**100%**	**94.93%**

Y. M. Galvão et al. [176]	2017	RGB	F1 score **Multilayer perceptron (MLP) 0.991** K-nearest neighbors (KNN) 0.988 SVM—polynomial kernel 0.988
Leila Panahi et al. [188]	2018	Depth	Average results SVM Sensitivity 98.52% Specificity 97.35% Threshold-based decision Sensitivity 98.52% Specificity 97.35%
K. Sehairi et al. [186]	2018	RGB	Accuracy SVM-RBF 99.27% KNN 98.91% **ANN 99.61%**
Chao Ma et al. [198]	2019	RGB+IR	**Autoencoder** **Sensitivity 93.3%** **Specificity 92.8%** SVM

			Sensitivity 90.8%				
			Specificity 89.6%				
F. Harrou et al. [202]	2019	RGB	Accuracy :				
			K-NN 91.94%				
			ANN 95.16%				
			Naïve Bayes 93.55%				
			Decision tree 90.48%				
			SVM 96.66%				
Ricardo Espinosa et al. [207]	2019	RGB			Sensitivity		Specificity
			Softmax		**97.95%**		**83.08%**
			SVM		14.10%		90.03%
			RF		14.30%		91.26%
			MLP		11.03%		93.65%
			KNN		14.35%		90.96%
Xiangbo Kong et al. [212]	2019	Depth		HOG+SVM	LeNet	AlexNet	GoogLeNet	ETDA-Net
			Average accuracy	89.48%	88.28%	93.53%	**96.59%**	95.66%
			Average specificity	95.43%	97.18%	97.56%	98.76%	**99.35%**
			Average sensitivity	83.75%	74.54%	87.10%	88.74%	**91.87%**
B. Wang et al. [215]	2020	RGB	F1 score					

		Falling state
		GDBT 95.69%
		DT 84.85%
		RF 95.92%
		SVM 96.1%
		KNN 93.78%
		MLP 97.41%
		Fallen state
		GDBT 95.27%
		DT 95.45%
		RF 96.8%
		SVM 95.22%
		KNN 94.22%
		MLP 94.46%
C. Zhong et al. [217] 2020	IR	F1 score
		RBFNN 89.57 (+/−0.62)
		SVM 88.74% (+/−1.75)
		Softmax 87.37% (+/−1.4)
		DT 88.9% (+/−0.68)

C. Menacho et al. [216]	2020	RGB	Accuracy		
			VGG-16 87.81%		
			VGG-19 88.66%		
			Inception V3 92.57%		
			ResNet50 92.57%		
			Xception 92.57%		
			ANN proposed in this system 88.55%		
G. Sun et al. [218]	2020	RGB		Sensitivity	Specificity
			SVM	92.50%	93.70%
			KNN	93.80%	92.30%
			SVDD	**94.60%**	**93.80%**
Yuxi Chen et al. [233]	2020	Depth	Average values		
			Lenet		
			Sensitivity 82.78%		
			Specificity 98.07%		
			AlexNet		
			Sensitivity 86.84%		
			Specificity 98.41%		
			GoogLeNet		

	Sensitivity 92.87%
	Specificity 99%

Table 8. Vision-based system performance evaluation datasets.

Signal Type	**Dataset Name**	**Characteristics**
Accelerometric and electroencephalogram (EEG) and RGB and passive infrared (IR)	Upfall [208]	17 volunteers execute falls and activities of daily life (ADL) of different types recorded by an accelerometer, EEG, RGB and passive IR systems
Depth and Accelerometric	Depth and accelerometric dataset [171]	Volunteers execute several activities, and falls are recorded by a depth system and accelerometers.
	TST fall detection [165]	11 volunteers execute 4 fall types and 4 ADLs recorded by RGB-depth (RGB-D) and accelerometer systems
	UR fall detection [157]	30 falls and 40 ADLs recorded by RGB-D and accelerometer systems
RGB	Center for digital home data set—MMU [196]	20 videos, including 31 falls and several ADLs
	LE2I [151]	191 different activities, including ADLs and 143 falls
	Charfi2012 dataset [153]	250 video sequences in four different locations, 192 containing falls, and 57 containing ADLs. Actors, under different light conditions, move in environments where occlusion exits and cluttered and textured background is common
	High-quality dataset [181]	It is a fall detection dataset that attempts to approach the quality of a real-life fall dataset. It has realistic settings and fall scenarios. In detail, 55

		fall scenarios and 17 normal activity scenarios were filmed by five web-cameras in a room similar to one in a nursing home
	Multicam fall dataset [138]	The video data set is composed of several simulated normal daily activities and falls viewed from 8 different cameras and performed by one subject in 24 scenarios
	Simple fall detection dataset [148]	The dataset contains 30 daily activities such as walking, sitting down, squatting down, and 21 fall activities such as forward falls, backward falls and sideway falls
	MOT dataset [200]	MOT dataset intends to be a framework for the fair evaluation of multiple people tracking algorithms. In this framework, the designers provide : • Detections for all the sequences; • A common evaluation tool providing several measures, from recall to precision to running time; • An easy way to compare the performance of state-of-the-art tracking methods; • Several challenges with subsets of data for specific tasks such as 3D tracking and surveillance.
	COCO dataset [201]	COCO is a large-scale object detection, segmentation, and captioning dataset designed to show common objects in context
	Piropo [224]	Multiple activities recorded in two different scenarios with both conventional and fish eye cameras
Depth	IASLAB-RGB fallen person dataset [163]	It consists of several static and dynamic sequences with 15 different people and 2 different environments

	Multimodal multiview dataset of human activities [178]	It consists of 2 datasets recorded simultaneously by 2 Kinect systems including ADLs and falls in a living room equipped with a bed, a cupboard, a chair and surrounding office objects illuminated by neon lamps on the ceiling or by sunlight
	Sdufall [140]	10 volunteers develop 6 activities recorded by RGB-D systems
	Falling detection [166]	6 volunteers perform 26 falls and similar activities recorded by RGB-D systems.
	Fall detection dataset [158]	5 volunteers execute 5 different types of fall
	NTU RGB+ dataset [227]	It is a large-scale dataset for human action recognition. It contains 56,880 action samples and includes 4 different modalities of data for each sample: RGB videos, depth map sequences, 3D skeletal data and IR videos
Synthetic Movement Databases	CMU Graphics Lab—motion capture library [183]	Library that captures synthetic movements through movement capture (MoCap) technology

2.3 Discussion

2.3.1 Wearable systems

The monitored person, either directly attached to the body itself or integrated into clothes or accessories, carries wearable systems used for human fall detection. These systems monitor acceleration, linear or angular speed, angles, orientation, pressure, or distances to determine what events of the person's daily life could be assessed as a fall. Once an event has been identified as a potential fall, it can be used in the contexts of fall detection or fall prevention.

Wearable technologies present, in comparison with the rest of the technologies used in the field of automatic fall detection, an array of specific characteristics that will be presented in depth and that are mainly related to the effects of the system's power supply through batteries. This reality implies the need for reducing power consumption as much as practicable, which, in turn, requires algorithms optimized to switch from idle monitoring in low fall probability states to intense monitoring in those situations when a fall is likely.

This way, wearable fall detection systems could be grouped into two essential categories, the ones whose main objective is fall detection and the ones whose main goal is fall prevention. This differentiation can be observed in different articles of the state-of-the-art of this technology such as [235] and [236]. In accordance with this classification all systems able to detect a fall once it has taken place can be included in the group of fall detection systems while all other systems able to identify a fall in its very early states or able to identify gaits with high fall probabilities should be part of the second group.

An alternative taxonomy to the one described in the previous paragraph would be the one grouping systems as a function of its capabilities to detect a fall in any of its states or its ability to discriminate human gaits with higher fall probabilities.

According to this alternative classification, the first group of systems aims to detect a fall at any given state, prior either to impact or after it. The vast majority of collected papers describing systems belonging to this group, although start an intense monitoring process prior to ground impact, only declare the fall after this event has taken place, as the objective of these systems is requesting emergency help. However, a few of the considered systems declare the fall in previous stages, during the initial phases of the fall, so the fall can be prevented, as in [62], or its effects mitigated, as in [44].

The second block of systems includes all systems that are able to classify gait styles and determine whether the fall probability for a particular way of walking is high. If that is the case, the system informs the user or their care providers, so measures can be taken to avoid a potential future fall.

2.3.1.1 Technologies

The technologies used to prevent and detect falls in wearable devices are based on the analysis of signals provided by sensors carried by the monitored person, either directly attached to the body itself or integrated in clothes or accessories.

These sensors can provide a wide range of different kinds of signals including:

- Acceleration
- Pressure
- Inclination
- Sound

- Electromyography
- Electrocardiography

Additionally, a number of systems use a combination of signals as a method to improve their performances compared to the ones obtained when a single signal type is used.

Tables 2 and 3 include the systems considered in this study grouped by technology clusters with indication of their measured performance.

2.3.1.2 Position and number of sensors

Sensor position highly determines system's performances.

For the case of systems using accelerometer or gyroscope signals this impact is documented. This way, Nor Surayahani [237] places inertial sensors on three different positions; hip, thigh and foot, concluding that the optimal sensor placement is the hip. In addition, sensors placed on feet offer the lowest system's performances while sensors on thigh deliver performances in between both.

J. Jacob [238] studies system's performances placing inertial sensors at different points on the backbone. This study concludes that, in overall terms, higher positions offer better performances, with optimal results when sensors are placed on T-4 and diminishing ones as their position lowers.

N. Pannurat [31] proposes a specific algorithm for signal analysis and evaluates its results when sensors are placed at different body positions. This way, optimal sensor placement is, according to this study and in a diminishing order of performance, hip, head, wrist, thigh, chest, ankle and arm.

Finally, C. Krupitzer [239] develops a system able to optimize signal analysis as a function of sensor position (thigh, hip or chest). System's test results are optimal when the sensor is placed on chest while hip and thigh are, in this order, less desirable sensor placements.

In general terms it is accepted that the higher the number of sensors is, the better the performances of the system are. However, this concept has its limits, as an excessive number of them leads to diminishing performances because of overfitting phenomena. This effect is documented by E. Casirali [240] in a work where, in addition, the importance of placing at least one sensor on the wrist is demonstrated in order to optimize system performances.

2.3.1.3 Signal analysis and classification algorithms

Once sensors have generated a proper signal, its analysis will determine whether a fall has taken place or whether a specific gait meets the criteria to assign it a high fall probability. The steps associated to this analysis are the following ones:

- Signal pre-processing. The objective of this phase is to prepare the signal for the analysis itself. This way, noise is diminished by using Kalman or passband filtering, outliers are disregarded by applying statistical analysis, and the signal domain is switched from time to frequency by employing Fast Fourier Transform techniques.
- Signal characterization. Speed and acceleration, both linear and angular, will be inferred from sensor signal after pre-processing and these elements will be used to feed the classification algorithm which will determine whether a fall has taken place or whether a gait has a high fall probability.

- Classification. All classification algorithms identified in this work are part of one of the following blocks:
  - Threshold based:
    - Fixed threshold.
    - Adaptative threshold. The threshold is adapted to a specific environment, situation or person by using statistical or machine-learning techniques.
  - Non-threshold based:
    - Based on statistical methods.
    - Based on machine-learning methods.
    - Mixed algorithms:
      - Homogeneus. Algorithms belonging to a single block are fused.
      - Heterogeneus. Algorithms belonging to different blocks are fused.

2.3.1.3.1 Signal pre-processing

During this phase, a number of processes aiming to improve signal characteristics take place before the analysis itself is executed.

The main objective of this phase is often noise reduction or elimination, especially when accelerometers, gyroscopes or inclination sensors generate the signal.

Mechanisms used for this purpose are diverse and include Kalman filtering, as proposed by J. He [241], [242], where signals coming from accelerometers and gyroscopes are filtered using Kalman techniques.

Pass band filtering is often used to eliminate, at the same time, the gravity vector, placed in the lower band area, and noise, placed in the higher band sector. This technique is used by A. Sucerquia [37], [243].

Some other procedures are also used to reduce noise. Among them, the use of some frequency domain techniques, such as the Fast Fourier Transformer (FFT), is very relevant. This way, D. Bersch uses FFT's in [244] to filter out high-frequency harmonics, mitigating this way the amount of noise present in the signal.

The validity of these techniques is a function of the specific characteristics of every signal and is subject to different processing costs. Pass Band filters are simple and require reduced amounts of power, while more sophisticated techniques, such as FFT's, are much more powerful and resource consuming. Costs and performances must be carefully evaluated at design time to reach a satisfactory trade-off point. Although high computing costs could be bypassed by using online processing, constant network linking has power costs not in line with the low power consumption philosophy guiding the design of wearable systems.

2.3.1.3.2 Signal analysis and characterization

Once signal has been pre-processed, it needs to be characterized.

The following variables can be inferred from sensor signal:

- Accelerometer: Linear acceleration and velocity [17] - [40].

- Gyroscope: Angular acceleration and velocity [41] - [46].
- Pressure sensor: Pressure difference registered at inertial event times [49] - [50].
- Inclination sensor: Velocity variation [51] - [52].
- Sound sensor: Frequencies obtained from signal samples through the use of FFT techniques [53].
- Electromiography (EMG): Nervous signal intensity commanding muscles [54] - [61]
- Electrocardiography (ECG): Nervous signal intensity commanding heart [63].

Some classification algorithms, based on Artificial Neural Networks (ANN), as the one used by E. Casilari-Pérez [245] , do not require signal characterization because the pre-processed signal can directly feed the network.

2.3.1.3.3 Classification algorithms

A. Threshold based

Threshold based algorithms try to determine whether a fall has taken place or a specific gait meets the criteria to assign it a high fall probability by determining whether signal values over-exceed certain thresholds. This way, if that is the case, it is assumed that the event has taken place.

Threshold value selection is critical, as too low values increase sensibility, raising false positives at the same time and too high values increment specificity while simultaneously reduce sensibility. Threshold determination has therefore a critical impact on system performance.

In overall terms, classification algorithms based on thresholds have low processing requirements and, consequently, hardware and power consumption requirements are low, vital characteristics for wearable systems.

Thresholds can be established for all system users, and in this case, they are called fixed thresholds, or they can be modified in order to adapt them to specific users, being called in this case adaptive thresholds. Systems based on fixed thresholds have lower power consumption and processing requirements than adaptive ones, as the former ones are much simpler.

a. Fixed threshold

These algorithms are the simplest and, therefore, they are the most suitable ones to be used by devices with important processing or power limitations. However, they are not user adaptive, which implies lower performances than more capable algorithms.

Fudickar [246] identifies several fall phases establishing an initial free fall followed by ground impact. He also describes an ulterior stabilization phase associated to fall shock during which the person lies still on the ground and a final phase, called critical, which starts when the fallen person starts moving to stand up again.

Fudickar [247] establishes the duration of each window, as well as, for inertial devices, the thresholds of the impact and critical phases.

Using the fixed threshold philosophy Abdelhedi [25] develops a system able to evaluate axis acceleration and, based on this information, it determines angular variation.

Razum [248] proposes a methodology to determine optimal fixed thresholds, establishing all the system's thresholds at the same time instead of doing it sequentally, as traditionally had been the case. This way, both sensibility and specificity are improved.

Pham [249] determines an optimal procedure to establish fixed thresholds, so false positives are diminished as much as possible.

Thella [250] studies how difficult it is to determine thresholds for the free fall phase, moment when any system aiming to mitigate fall damages must trigger actions, which, in this system specific case, would be airbag initiation. Sivaranjani [44] also proposes an airbag damage mitigation system triggered by a combination of a substantial diminish of the vertical acceleration and a significant increase of the angular velocity measured at the ankle.

In the area of gait analysis Hemmatpour [251] presents a gait non-linear model able to establish detection thresholds of abnormal gaits with high fall likelihood.

b. Adaptive threshold

These algorithms, more complex than the previous ones, require higher processing power and are more power consuming. However, they are more capable and are able to fit individual's peculiarities improving, this way, system's performances.

Yinfeng [252] proposes an algorithm based on multivariable statistical analysis to adjust standard threshold values to specific individuals.

Lingmei [253] classifies system's users as a function of age, sex, height and weight. This way, thresholds can be modified and performances are improved, both in the area of fall detection and fall prevention.

Otanasap [168] presents an algorithm, which, starting from standard thresholds is able to adjust them to individuals using user's history.

As expected, adaptive threshold performances over-exceed fixed ones in all the studied systems and, therefore, thresholds should be adapted to users whenever that is possible.

B. Non-threshold based

This kind of methods use more complex algorithms than the previous ones to determine whether a fall event has taken place.

As Aziz [254] documents, these algorithms offer better performances but require higher processing power, and therefore, more powerful hardware is required if services are provided as edge applications, or constant network connection is required. In both cases, power consumption is higher than in the case of threshold-based algorithm.

Thus, a number of studies have been carried out to evaluate whether algorithms based on statistical models and machine-learning techniques have practical applications in the fields of fall detection and gait analysis. These algorithms, due to their improved performances, have gained momentum in the last five years, as evidenced by Xu [255]. He evaluates systems designed over this period and finds out that non-threshold algorithms have been used 30% more frequently than threshold-based ones.

a. Statistical models

A number of statistical models have been used to recognize falls using signals provided by wearable sensors. Although this kind of technique has not been extensively used in the studied systems, it is a quick method to classify activities and determine whether an event meets the criteria to consider it a fall or whether a certain gait has the characteristics to assign it a high fall probability.

This way, the system proposed by Xinyao [256] uses an auto-regression and moving average model (ARIMA) to evaluate if the combined signals provided by sensors placed on the monitored person's head, trunk and arms are within the limits established for non-falls events or if they are exceeded, assuming in the latter case that a fall has taken place.

Finally, Su [46] presents a system, which uses a linear Fisher discriminant to determine, in a time as reduced as possible, whether the signals produced by inertial sensors placed on the trunk and the leg are within parameters of normality.

b. Machine learning

Machine learning techniques have been extensively used in this area and a number of comparative studies try to determine which ones offer better performances.

Table 3 offers an insightful view of the performances of different machine learning classification algorithms used by fall detection or fall prevention wearable systems.

S. Ray [257] classifies machine-learning algoritms as follows:

- Descend gradient algorithms.
- Linear regression algorithms.
- Multivariable regression algorithms.
- Logistic regression algorithms.
- Decision tree algorithms.
- SVM algorithms.
- Naïve-Bayes algorithms.
- K-NN algorithms.
- K means clustering algorithms.
- Neural network training algorithms.

The first block of algorithms tries to minimize a defined cost function. Their coefficients are updated at every iteration until convergence of the cost function is obtained, and further iterations cannot reduce its value any more. This type of algorithms could be directly used for classification purposes, but in the reviewed systems, they are used to train neural networks. For that reason, this first block could be considered as part of the last one.

Due to its simplicity, linear regression is not useful in this area, but multivariable regression has been used in the form of both linear discriminant and Fisher algorithms. In multivariable regression analysis, relations are established between dependent variables (system outputs) and independent variables (system inputs). Linear discriminant and Fisher algorithms are part of the same block, as both classify objects by establishing linear combinations of features that

characterize the class they belong to. Although algorithms of this block are used in some of the reviewed papers, the systems using them have not demonstrated good performance.

Logistic regression algorithms are also used to determine fall probability. To do this, a logistic function is used to model the fall probability variable. Its parameters are estimated using fall data, and once they have been established, the function is used to determine fall probability. As in the previous case, logistic regression algorithms have proven to deliver poor performance.

Bayesian algorithms are also used in the reviewed systems. This technique uses Bayes' theorem to determine event probability using previous information (a priori). For the specific case of the Naïve-Bayes algorithm, a number of simplifications are made in order to establish a fall probability, as D. Berrar details [258]. As in previous cases, the performances of this technique are quite low.

Decision tree algorithms and aggregations of this method, known as random forests, can be found in the reviewed papers to determine whether a fall event has taken place. Because of the simultaneous consideration of different elements in a single tree, they tend to over fit. To avoid this problem, multiple trees can be used, implementing, this way, a concept known as random forest. Decision trees performances, although better than the ones obtained by previous algorithms, continue being low compared to the ones of other reviewed algorithms.

SVM algorithms are extensively used in the reviewed papers and their performances are excellent. They establish a separation surface limiting classes in order to classify objects. When this separation cannot be linear, surfaces are created by using complex functions.

The block of KNN algorithms is often used by the reviewed systems with good results. During their training phase, these algorithms cluster object classes in order to classify events in their operational phase. This way, in this second phase items are assigned to the closer class after mean distance to all of them is calculated.

Neural network training algorithms are used to obtain network optimal coefficients. Neural networks are an aggregation of items, called neurons, connected to each other through links. Each neuron outputs values that are multiplied by a factor called weight. The result is fed into the connected neuron and, in turn, it outputs values, which, after being multiplied by new weights, are pumped into ulterior neurons. Additionally, value or intensity conditions are set to activate inter-neural links. These conditions are established through functions called activation functions. This way, proper weights are key to obtain good network performances and, to determine them, training algorithms are used during network training.

The following table summarizes the results of the three best performing algorithms.

Table 9. Machine learning classification algorithms performance comparison.

	STUDIES IT APPEARS IN	BETTER PERFORMANCES
SVM	20	8 (40%)
KNN	15	5 (33%)
ANN	20	12 (60%)

However, these results should be cautiously analyzed, as no common reference benchmark was used for comparison. Additionally, real world fall data is scarce and the vast majority of

it is, in reality, data associated to simulated falls performed by young volunteers. Both aspects are well documented by L. Ren [259] in his article.

2.3.1.4 System performance indicators

The two essential variables determining fall detection system performances are system sensibility and system specificity. They are defined as follows:

$$Sensibility = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives}$$

Some studies reviewed reference system performance to its accuracy, defining it as:

$$Accuracy = \frac{True\ positives + True\ negatives}{True\ positives + True\ negatives + False\ positives + False\ negatives}$$

The set of system's performance variables is completed by defining Prevalence as:

$$Prevalence = \frac{True\ positives + False\ negatives}{True\ positives + True\ negatives + False\ positives + False\ negatives}$$

In other words, Prevalence is the relation between the total number of falls and the total number of events.

This way an alternative way to define Accuracy is:

Accuracy=Sensibility * Prevalence * Specificity * (1 – Prevalence)

In simple terms, sensibility is a measure of the system capability to determine what events are a fall while Accuracy determines the system capability to establish what events are not a fall.

Accuracy reduces granularity, as it just measures how accurate the system is at classification tasks. This way, as the number of falls is low compared to the rest of daily events, the most relevant parameter for Accuracy is Specificity. This way, a fall detection system could have high Accuracies as long as its Specificity is high, even if its Sensibility is low.

Additionally, a last parameter, F1, is defined as:

$$F1 = \frac{2\ x\ True\ positives}{2\ x\ True\ positives + False\ positives + False\ negatives}$$

In any case, independently of the used parameter to measure system performance, it depends on the type of signal, the number of sensors and the signal processing (pre-processing, signal characterization and classification algorithm)

However, system comparison is difficult, as no common benchmark has ever been established and the used fall data has been extracted from simulated falls whose characteristics are not necessarily identical to elderly's ones.

The differences between young volunteers falls, who usually simulate the falls, and elderly's ones is studied by M. Kangas [260], concluding that there are significant differences between both. Additionally, Klenk [261] finds substantial differences between simulated and real falls contradicting the results of the studies made by Timo Jämsä [262] and E Casilari [263], which report statistical similarities. Anyhow, all these studies conclusions are extracted from a very reduced number of experiments and, therefore, they cannot be considered decisive.

Finally, the absence of common test benchmarks is highlighted in a number of papers [259], [245].

2.3.1.5 Low power consumption

Low power consumption is key for wearable fall detection systems as, for the time being, they operate on batteries.

Power consumption depends, essentially, on system architecture and on fall detection algorithm implementation.

Distributed system architectures, which process signal nearby sensors, are optimal when sampling rates are reduced and detection algorithms are simple as, this way, power consumption, processing capabilities and hardware requirements will be low. Power consumption will be even lower if the used detection algorithm implements power saving modes (sleep mode) whenever that is possible. These systems, as expected, have lower performances than other ones whose requirements are higher.

Systems implementing algorithms that are more complex have improved performances, requiring higher processing power and hardware requirements, which, in turn, means that power consumption will be higher. Additionally, cloud computing, which could diminish processing needs, requires constant network linking that implies higher power consumption. This way, systems like the one presented by E Casilari [263], which requires constant Bluetooth link between a smarwatch and a telephone, reduces watch battery time between charges from a week down to around 10 hours.

Important efforts are being made to recharge batteries or feed systems through alternative methods. This way, energy collecting systems based on the use of the piezoelectric effect have been proposed, as Q. Cheng did in [264], where he uses this effect to feed a LCD screen and a LED.

T. Wu [265] presents a system able to energize inertial and photoplethismographic (PPG) sensors for 12 hours using the solar energy collected during a period of 3 hours and K. Li [266] introduces one able to harvest energy from limbs movements.

Y. Cai [267] designs a power device which collects solar, thermal and mechanic energy to feed wearable systems while S. M. Noghabaei's one [268] harvests electromagnetic energy

and S. Roundy [269] studies mechanic energy harvesting comparing outputs of piezoelectric, electromagnetic and electrostatic systems.

Finally, M. Mohammadifar [270] proposes a power device based on microbial power cells which explodes bacteria's ability living on the human skin to transform the chemical energy of the human sweat into electrical one in order to power wearable systems.

All these power generation alternative technologies are still in a very immature stage and, for the time being, they are not an option to power any real-world system. This way, for the time being, the only credible option to extend system operational period between charges is diminishing power consumption as much as possible.

2.3.1.6 Fall detection wearable systems acceptance

In overall terms, wearable fall detection systems have not been well accepted by the potential main beneficiary community, the elderly.

Thilo [15] studies this phenomenon and interviews a number of nursing home residents concluding that introducing this kind of systems in that community can only be done if the final user is included in the system design and development phases, making them feel part of the team. It also requires commitment of the caregivers community, as they should be the bridge connecting the system development group and the final user. This way, care givers, who are the final user's most trusted group, could transmit elderly's needs and system requirements to the development teams while, at the same time, help the elderly to understand the advantages of this kind of systems.

Surprisingly, in the reviewed papers there are only two references to actions of this nature taken by the system development community in order to favor system introduction. These references are the above-mentioned paper by Thilo and a study by Demiris [16], where a number of recommendations made by nursing home residents to the system development community are gathered.

2.3.1.7 Conclusions

Wearable fall detection systems' maturity has experienced significant advances over the last few years. The research community's effort to achieve a level of maturity high enough to make them commercially viable is evident from the number of papers published on wearable systems, which is the second highest after the one associated with vision-based systems.

Wearable systems use sensors carried by the monitored person to evaluate their movements. The vast majority of system sensors are either accelerometers or gyroscopes, although some other kinds, such as microphones, pressure sensors, ECG, and EMG, are also used.

After an initial pre-processing phase, whose main goal is noise reduction, the systems analyze the signal to determine its characteristics. These characteristics are then used for classification purposes to evaluate whether a fall event has taken place.

There are several classification algorithms, with threshold-based ones being commonly used due to their simplicity and low processing power requirements. However, their performances are limited when compared to other more capable algorithms. Classification based on machine learning has demonstrated optimal outcomes, closely followed by SVM algorithms. However, unlike the threshold-based algorithms, their power consumption and processing requirements are high.

Some of the reviewed systems use multiple sensors, and, in general, the higher the number of sensors, the better the system's performance. However, this general rule has limitations, and an excessive number of sensors could reduce performances due to overfitting phenomena. Additionally, sensor position is relevant, especially in the case of inertial ones. They offer better results when placed in positions where movement during a fall is maximum.

Despite efforts to power wearable devices through alternative means, such as mechanical, biochemist, or sunlight energy harvesting systems, batteries continue to be the only credible devices to energize this kind of systems. This fact favors edge-computing techniques and simple processing algorithms, as these characteristics reduce power consumption. Additionally, further consumption reductions imply diminished sampling rates and keeping systems in dormant states for periods as long as practical. While these features are very positive in terms of power consumption reduction, they may diminish system performances. Therefore, a reasonable design trade-off point must be reached.

The vast majority of reviewed systems use validation databases to evaluate performances. Some of those databases have been developed to check specific systems and have never been released to the research community, while some others are public ones. The only one among them containing real fall data is FARSEEING [271]. The rest has been made by volunteers and actors a lot younger than the elderly community, theoretical main beneficiary of these systems. This fact, together with the conclusions of the studies made by Kangas [260] and Klenk [261], which document significant differences between real and simulated falls, and between the ones of young and elderly people, highlight how reduced is the amount of valid data that can be used to design and validate fall detection systems and rises doubts about the reliability of these systems in the real world.

In addition, there are no standard evaluation benchmarks globally accepted by the community of developers. This fact makes system comparison very difficult, which, in turn, substantially complicates reaching conclusions about system performances.

Finally, and although most commercial fall detection systems are based on wearable technologies, mainly inertial ones, its acceptance rate among the elderly continues low, as shown in the system requirement chapter. Surprisingly, there is an almost absolute lack of studies aiming to assist system development, as only the paper by Thilo et al. [15] and the one by Demiris et al. [16] describe the needs of the elderly community and make recommendations to guide developers. Given that none of the reviewed papers devotes any effort to collect user's or system requirements, it becomes reasonable to assume that low acceptance rates are, at least partially, a consequence of the disregard for the real user's needs.

In any case, and in spite of the problems associated to the disadvantages of carrying a sensor at all times and paying attention to keeping its battery operational, the maturity of wearable technologies is responsible for being, together with the artificial vision one, the technology used by commercial systems.

2.3.2 Ambient systems

Ambient fall detection systems, unlike wearable ones, place sensors around the monitored person. Although in strict terms, artificial vision techniques should be part of this group, most taxonomies establish a specific block for artificial vision systems, which includes both the ones able to process visual signals and the infrared-based ones. In this work, for practical reasons, the formal division between ambient and artificial systems will be maintained.

Ambient systems present advantages compared to wearable ones, as they lack the power consumption, processing power, and network linking limitations associated with the latter

ones. However, they also have substantial limitations, as these systems are only present in a very limited number of environments, usually the ones where the life of the monitored people takes place most of the time.

None of the reviewed ambient systems can conduct gait analysis, so all of them should be classified as fall detection systems.

2.3.2.1 Technologies

Different types of sensors could provide the signal to be fed into the system depending on the used technology. This way, signal could come from:

- Acoustic sensors
- Contact sensors:
  - Pressure
  - Deformation
  - Capacitance
- Passive infrared sensors
- Radiofrequency sensors:
  - Radar
  - WiFi

Table 5 includes all reviewed systems classified in accordance with the previous ambient fall detection system taxonomy.

2.3.2.2 Signal analysis and classification

After sensors generate a signal, like in the case of the wearable systems, it must be processed, so whether a fall has taken place can be concluded. This analysis is divided into the following blocks:

- Signal pre-processing. In overall terms, this phase goal is decluttering and, when the signal is an aggregation of components, the separation is done during this phase.
- Signal characterization. This process phase aims to infer the main signal characteristics so the event can be classified. A number of systems using deep neural networks inject the pre-processed signal straight into them, so signal characterization and classification phases are fused in a single one.
- Classification. Its objective is determining whether a fall has taken place. To do it the relevant characteristics of the signal, extracted in the previous phase, are used.

2.3.2.2.1 Signal pre-processing

Signal pre-processing varies depending on the type of signal.

For the case of sound signals, sound direction can be determined if an array of microphones is used. This is the case of the system presented by Mungamuru [272], which uses signal phase displacement to determine sound direction. Additionally, a number of the reviewed systems use low-pass filters to reduce noise. In the case of [96] and [97] a Hidden Markov Model to suppress periods of silence and separate superposed sounds is employed.

The reviewed systems using contact signals do not pre-process it.

PIR systems pre-process signal in order to reduce noise, usually by using moving average techniques or other statistical methods that try to extract human silhouette out of clutter by using the temperature difference between background and human body. These processes are detailed in [103]. Gaussian filters and high frequency component suppression after wavelet transform application over signal [105] are also useful methods for noise reduction.

Radar based systems need signal comparison between broadcasted and received signals, so Doppler shifting can be established. This way, signal is decluttered before sampling it. Additionally, Kalman filter is used in some systems like [117] to track moving objects.

Finally, WiFi systems evaluate wave propagation modifications because of human body interposition between emitting and receiving antennas. In [125] amplitude and phase displacement of the carrier wave of a domestic WiFi is evaluated, as well as its Doppler shifting as a consequence of hunan activity. To do it, as different devices request router services and, therefore, broadcasted signal is not linearly distributed over time, an initial interpolation is executed to apply then a pass-band filter, so all non-relevant frequencies are discarded. This way, in [127] a Butterworth passband filter is used with this purpose and in [131] a wavelet transform is applied to filter out noise before signal is passed to the following block.

2.3.2.2.2 Signal analysis

As in the previous case, used analysis techniques are signal dependant.

For the case of sound signals, Mel Frequencies Cepstral Coefficients (MFCCs) will be determined, following an analogous process to the human hearing one, which allows us to recognize sounds and voices. To do this, the following steps are followed:

Signal is divided into time steps.

- A FFT is applied and spectral power is determined in every time step. Then the spectral power distribution is calculated using frequency steps, which are referred to a Mel scale.
- Power logarithms are taken in each frequency step.
- A discrete cosine transform is applied to the logarithms obtained in the previous step.
- MFCC's will be the obtained power spectrum amplitudes.

This technique, used in [92]- [95] and in [98] ,replies the way human hearing works, as it groups sound signals by frequency steps in order to recognize them and its sensibility is not linear but logarithmic.

An alternative characterization process is proposed in [96] and [97], where acoustic local ternary patterns are proposed for sound recognition, a technique imported from the artificial vision field and used by these systems with good results.

Finally, in [98] a Siamese neural network, able to infer fall sound characteristics from a limited number of real fall data, is used with fall detection purposes. This network, as in the previous case, is imported from the world of artificial vision and most classification algorithms can easily handle its output. Furthermore, this is the only system among all the reviewed ones, which uses a neural network with fall sound recognition purposes.

For the case of contact signals, all reviewed papers use voltage for signal characterization.

PIR systems establish human body detection angle. Once it has been determined both vertical and horizontal velocities, as well as trajectory deviations and their variances are calculated, as it happens in [103]. Alternative approaches, as the one in [106], use other data such as sensor distance, activity duration or trajectory. Some other approaches, such as [107], use deep neural networks to visually recognize signal characteristics.

In radar systems, sampled Doppler shifting is treated to obtain signal characteristics. In [118] mean, maxima, increasing and decreasing rates and variance are used with classificatory objectives. In [110] a wavelet transform is applied to the pre-processed signal to calculate spectral power distribution over time, which is used with classification purposes. In [111] power spectrum is also determined after sequencing signal in two-second blocks and applying an FFT to each one of them. A similar process is used in [116], where a Short-time Fourier Transform (STFT) is used. In [112] the integrated map of Doppler shift vs distance is the parameter used to classify events. In [115] a similar map is used but, in this case, both vertical and horizontal axis are considered. A similar process is followed in [120], where those maps are used to evaluate energy at the points where a fall event could have taken place in order to corroborate it. In [113] and [116] time domain Doppler Shift is used with classification purposes.

WiFi systems use a number of techniques to evaluate fall probability. This way, in [125] start and end points of potential fall events are determined using phase displacement. Once the time slot assigned to the fall has been destermined, phase displacement variation, mean, total deviation, maximum value, event duration, phase changing rate and signal energy diminution during the event are used to determine whether it has been a fall. In some other systems like [126] signal is directly fed into a neural network.

2.3.2.2.3 Classification algorithms

Table 5 offers a good perspective of the classification methods used in this field, as it contains all the classification algorithms used by the reviewed ambient fall detection systems. It must be noted that, although in general terms, they are similar to the ones used in wearable systems, some of them slightly differ.

HMM is part of the semi-supervised algorithms block, Adaboost is part of the numeric supervised algorithms one, and K-NN and Voting and bagging belong to the algorithm regression group. DTW is a temporal series algorithm, which identifies similarities between series.

In table 10, which has been elaborated using data from table 5, the best-performing algorithms can be identified.

Table 10. Best classification machine learning techniques for ambient fall detection systems.

CLASSIFICATOR	BEST OPTION
ANN	66,7%
RF	5,6%
SVM	16,7%
KNN	11,1%

This way, although this data must be cautiously approached because there is not a widely accepted comparison benchmark, the results could suggest that neural networks, followed by SVM and KNN classifiers, are the best performing options.

2.3.2.3 System performance indicators

In addition to the metrics used to evaluate wearable system performances, two extra ones, FNR and AUC, are utilized to assess ambient ones.

FNR (False Negative Rate) measures the ratio of non-detected real falls or false negatives.

Additionally, and prior to explain what AUC is, ROC needs to be defined. ROC (Receiver Operating Characteristic Curve) is a curve that reflects performances of a binary classification system for its different settings. The False Positive Ratio (FPR) and the True Positive Ratio (TPR) are the used axis to represent the curve.

$$TPR = \frac{True\ positives}{True\ positives + False\ negatives}$$

$$FPR = \frac{False\ positives}{True\ negatives + False\ positives}$$

AUC (Area Under the Curve) is defined as the integral of the curve along its entire domain. This way, AUC, which will be a number between 0 and 1, will indicate improved system performances as it approaches 1.

2.3.2.4 Acceptance

The number of published papers devoted to ambient fall detection systems is lower than the ones describing wearable ones. However, they present a number of advantages that have allowed them to be better accepted by the elderly community. The main one is that the monitored person does not need to pay constant attention to wearable systems, which require ensuring at all times that their batteries are able to power them. However, the deployment of ambient systems around the monitored person is not instantaneous, as it is often limited to the facilities where the elderly live.

As in the previous case, the lack of interest from the developing community to understand the real problems of potential users of this kind of systems is clearly reflected by the number of papers alluding to this area. In the particular case of ambient fall detection systems, no paper has been found during the review process trying to assess these problems or having the intention to provide any feedback or guidance from the community of users and their caregivers.

Improvements in acceptance rates, as in the case of wearable systems, require better connections between the communities of users and developers.

2.3.2.5 Conclusions

Ambient fall detection systems' main difference, compared to wearable ones, is sensor position. While the monitored person carries wearable system sensors, ambient ones are placed around them. However, the development of ambient systems over the last few years has not been as intense as that experienced by wearable technologies. A good indicator of this weaker research interest is the number of published papers associated with these technologies, which is around half the number of articles associated with wearable or vision-based devices.

Ambient systems are based on contact, passive infrared, acoustic, radar, or Wi-Fi technologies. Although artificial vision technologies could be included in this block, as their sensors are deployed around the monitored person, most taxonomies put them in a different group.

In these systems, the signal passes through a first block where noise is reduced. Additionally, in acoustic systems, sound direction is determined, and often, silent periods are eliminated. For Wi-Fi systems, all irrelevant frequencies are blanked.

Signal characterization is different depending on the type of technology employed. For example, acoustic signal characterization is based on Cepstral Coefficients determination, a technique that, based on spectral distribution power comparison, tries to determine whether a fall has taken place. Passive infrared techniques extract human silhouette from the background using temperature differences between the human body and its surrounding environment. Contact systems evaluate whether a specific pressure distribution could be associated with a fallen body, while devices based on radar or Wi-Fi use Doppler shifting to evaluate fall probability.

Classification techniques are diverse. The most performant ones include K-NN algorithms, which use distance from the element to be classified to known samples, and SVM algorithms, whose performances are slightly better than K-NN ones. These algorithms are able to establish limits separating classes in multidimensional spaces. Finally, systems using neural networks have proved to be the most performant ones at successfully evaluating a potential fall event.

Ambient systems do not present problems associated with power consumption. This way, the limitations associated with processing power limits or continuous data transmission to cloud systems are suppressed. This clear advantage over wearable systems is countered by the limitations associated with sensor deployment, which, very often, is limited to facilities used by the elderly community.

The number of public databases associated with ambient technologies is very limited, and most reviewed systems have been validated using databases specifically developed for them. This fact, together with the lack of accepted common benchmarks, makes system comparison extremely complex.

Ambient systems' acceptance rate is even lower than that of wearable devices. However, they present a clear advantage from the user’s perspective as; unlike in the case of wearable ones, there is no need to pay continuous attention to battery status. In spite of that, ambient systems' acceptance rate remains almost nonexistent with no commercial use, probably because of the immaturity of these technologies.

2.3.3 Vision-based systems

Visual-based fall detection systems have been evolving in a manner similar to other human activity recognition systems that rely on artificial vision. There has been a notable increase in the use of ANN’s in these systems. Additionally, there is a clear trend towards adopting cloud computing systems, except for those integrated into robots.

All the analyzed systems share a common three-step approach to fall detection using artificial vision, with some variations depending on the specific system.

The first step, which is not always necessary, involves preprocessing the video signal to optimize it as much as possible.

The second step is characterization, where image features are abstracted to express what is happening in the images through descriptors. These descriptors will be used in the last step of the process.

The third step is the classification phase, where the observed actions, characterized by abstract descriptors, are labeled as either a fall event or a non-fall event. This allows measures to be taken promptly to assist the fallen person.

Some systems adopt a frame-by-frame approach, where the main goal is to classify the human pose as fallen or not, focusing less on the fall motion itself. For systems, trying to determine if a specific movement may be a fall, silhouette tracking is a fundamental operation performed through various processes. The tracking techniques used by the systems are explained in the following section.

Finally, a comparison of classification algorithm performance and validation datasets is presented in the last two sections.

2.3.3.1 Signal preprocessing

The final objective of this phase is either distortion and noise reduction or format adaptation, so downstream system blocks can extract characteristic features with classification purposes. Image complexity reduction could also be an objective during the preprocessing phase in some systems, so the computational cost can be reduced, or video streaming bandwidth use can be diminished.

The techniques grouped in this Section for decreasing noise are numerous and range from Gaussian smoothing used in [159] to the morphological operations executed in [145], [159], [202] or [152]. They are introduced in subsequent Section as a part of the foreground segmentation process.

Format adaptation processes are present in several of the studied systems, as is the case in [176], where images are converted to grayscale and have their histograms equalized before being transferred to the characterization process.

Image binarization, as in [217], is also introduced as a part of the systematic effort to reduce noise during the segmentation process. While some other systems, like the one presented in [184], pursue image complexity decreasing by transforming video signals from red, green and blue (RGB) to black and white and then applying a median filter, an algorithm which assigns new values to image pixels based on the median of the surrounding ones.

Image complexity reduction is a goal pursued by some systems, as the one proposed in [219], which introduces compressed sensing (CS), an algorithm first proposed by Donoho et al. [273] used in signal processing to acquire and reconstruct a signal. Through this technique, signals, sparse in some domain, are sampled at rates much lower than required by the Nyquist–Shannon sampling theorem. The system uses a three-layered approach to CS by applying it to video signals, which allows privacy preservation and bandwidth use reduction. This technique, however, introduces noise and over-smooths edges, especially those in low contrast regions, leading to information loss and image low-resolution. Therefore, image complexity reduction feature characterization often becomes a challenge.

Signal characterization

The second process step intends to express human pose and/or human motion as abstract features in a qualitative approach, to quantify their intensity in an ulterior quantity approach.

These quantified features are then used with classifying purposes in the last step of the fall detection system.

These abstract pose/action descriptors can globally be classified into three main groups: global, local and depth.

Global descriptors analyze images as a block, segmenting foreground from background, extracting descriptors that define it and encoding them as a whole.

Local descriptors approach the abstraction problem from a different perspective and, instead of segmenting the block of interest, process the images as a collection of local descriptors.

Depth characterization is an alternative way to define descriptors from images containing depth information by either using depth maps or skeleton data extracted from a joint tracking process.

2.3.3.1.1 Global descriptor

Global descriptors try to extract abstract information from the foreground once it has been segmented from the background and encode it as a whole.

This kind of activity descriptors was very commonly used in artificial vision approaches to human activity recognition in general and to fall detection in particular. However, over time, they have been displaced by local descriptors or used in combination with them, as these ones are less sensitive to noise, occlusions and viewpoint changes.

Foreground segmentation is executed in a number of different ways. Some approaches to this concept establish a specific background and subtract it from the original image; some others locate regions of interest by identifying the silhouette edges or use the optical flow, generated because of body movements, as a descriptor. Some global characterization methods segment the human silhouette over time to form a space–time volume, which characterizes the movement. Some other methods extract features from images in a direct way, as in the case of the system described in [176], where every three frames, the mean square error (MSE) is determined and used as an indicator of image similarity.

Silhouette Segmentation

Human shape segmentation can be executed through a number of techniques, but all of them require background identification and subtraction. This process, known as background extraction, is probably the most visually intuitive one, as its product is a human silhouette.

Background estimation is the most important step of the process, and it is addressed in different ways.

In [145], [152], [184], [202], as the background is supposed constant, an image of it is taken during system initialization, and a direct comparison allows segmentation of any new object present in the video. This technique is easy and powerful; however, it is extremely sensitive to light changes. To mitigate this flaw, the system described in [159], where the background is also supposed stable, a median throughout time is calculated for every pixel position in every color channel. Then, it is directly subtracted from the observed image frame-by-frame.

Despite everything, the obtained product still contains a substantial amount of noise associated with shadows and illumination. To reduce it, morphological operators can be used as in [145], [152], [159], [202]. Dilation and/or erosion operations are performed by probing

the image at all possible places with a structuring element. In the dilation operation, this element works as a local maximum filter and, therefore, adds a layer of pixels to both inner and outer boundary areas. In erosion operations, the element works as a local minimum filter and, therefore, strips away a layer of pixels from both regions. Noise reduction after segmentation can also be performed through Kalman filtering, as in [220], where this filtering method is successfully used with this purpose.

An alternative option for background estimation and subtraction is the application of Gaussian mixture models (GMM), a technique used in [135], [139], [142], [206], [220], among others, that models the values associated with specific pixels as a mix of Gaussian distributions.

A different approach is used in [134], where the Horprasert method [274] is applied for background subtraction. It uses a computational color model that separates the brightness from the chromaticity component. By doing it, it is possible to segment the foreground much more efficiently when light disturbances are present than with previous methods, diminishing this way light change sensitiveness. In this particular system, pixels are also clustered by similarity, so computational complexity can be reduced.

Some systems, as the one presented in [135], apply a filter to determine silhouette contours. In this particular case, a Sobel filter is used, which determines a two-dimensional gradient of every image pixel.

Other segmentation methods, like vibe [147], used in [150] and [222], store, associated with specific pixels, previous values of the pixel itself and its vicinity to determine whether its current value should be categorized as foreground or background. Then, the background model is adapted by randomly choosing which values should be substituted and which not, a clearly different perspective from other techniques, which give preference to new values. On top of that, pixel values declared as background are propagated into neighboring pixels part of the background model.

The system in [136] segments the foreground using the technique proposed in [275], where the optical flow (OF), which are presented in later Sections, is calculated to determine what objects are in motion in the image, feature used for foreground segmentation. In a subsequent step, to reduce noise, images are binarized and morphological operators are applied. Finally, the points marking the center of the head and the feet are linked by lines composing a triangle whose area/height ratio will be used as the characteristic classification feature.

Some algorithms, like the illumination change-resistant independent component analysis (ICA), proposed in [223], combine features of different segmentation techniques, like GMM and self-organizing maps, a well-known group of ANN able to classify into low dimensional classes very high dimensional vectors, to overcome the problems of silhouette segmentation associated with illumination phenomena. This algorithm is able to tackle segmentation errors associated with sudden illumination changes due to any kind of light source, both in images taken with omnidirectional dioptric cameras and in plain ones.

ICA and vibe are compared in [222] by using a dataset specifically developed for that system with better results for the ICA algorithm.

In [137], foreground extraction is executed in accordance with the procedure described in [276]. This method integrates the region-based information on color and brightness in a codeword and the collection of all codewords are grouped in an entity called codebook. Pixels are then checked in every single new frame and, when its color or brightness does not match

the region codeword, which encodes area brightness and color bands, it is declared as foreground. Otherwise, the codeword is updated, and the pixel is declared as area background. Once pixels are tagged as foreground, they are clustered together, and codebooks are updated for each one of them. Finally, these regions are approximated by polygons.

Some systems, like the one in [137], use orthogonal cameras and fuse foreground maps by using homography. This way, noise associated with illumination variations and occlusion is greatly reduced. The system also calculates the observed polygon area/ground projected polygon area rate as the main feature to determine whether a fall event has taken place.

Self-organizing maps is a technique, well described in [277], used with segmentation purposes in [186]. When applied, initial background estimation is made based on the first frame at system startup. Every pixel of this initial image is associated with a neuron in an ANN through a weight. Those weights are constantly updated as new frames flow into the system and, therefore, the background model changes. Self-organizing maps have been successfully used to subtract foreground from background, and they have proven to be resilient to light variation noise.

Binarization is a technique used for background subtraction, especially in infrared (IR) systems, as the one presented in [217], where the inputs IR signals pixels are assigned two potential values, zero and one. All pixels above a certain threshold value are assigned a value 1 (human body temperature dependent), and all others are given a value of 0. This way, images are expressed in binary format. However, the resulting image usually has a great amount of noise. To reduce it, the algorithm is able to detect contours through gradient determination. Pixels within closed contours whose dimensions are close to the ones of a person continue being assigned a value 1, while the rest are given a value 0.

Once the foreground has been segmented, it is time to characterize it through abstract descriptors that can be classified at a later step.

This way, after background subtraction, features used for characterization in [159] and [142] are silhouettes eccentricity, orientation and acceleration of the ellipse surrounding the human shape.

Characteristic dimensions of the bounding box surrounding the silhouette are also a common distinctive feature, as is the case in [206]. In [195], a silhouette's horizontal width is estimated at 10 vertically equally spaced points, and, in [202], five regions are defined in the bounding box, being its degree of occupancy by the silhouette is used as the classifying element.

Other features also used for characterization used in [135] and [167] include Hu moments, a group of six image moments in variables to translation, scale, rotation, and reflection, plus a seventh one, which changes sign for image reflection. These moments, assigned to a silhouette, do not change because of the point of view alterations associated with body displacements. However, they dramatically vary as a result of human body pose changes as the ones associated with a fall. This way, a certain resistance to noise due to the point of view change is obtained.

The Feret diameter, the distance from the two most distant points of a closed line when taking a specific reference orientation, is another used distinctive feature. The system described in [186] uses this distance, with a reference orientation of 90°, to characterize the segmented foreground.

Procrustes analysis is a statistical method that uses minimum square methods to determine the needed similarity transformations required to adjust two models. This way, they can be compared, and a Procrustes distance, which quantifies how similar the models are, can be inferred. This distance, employed in some of the studied systems as a characterization feature, is used to determine similarities between silhouettes in consecutive frames and, therefore, as a measure of its deformation because of pose variation.

The system introduced in [150], after identifying in each frame the torso section in the segmented silhouette, stores its position in the last 100 frames in a database and uses this trajectory as a feature for fall recognition.

To decrease sensitiveness to noise because of illumination noise and viewpoint changes, some systems combine RGB global descriptors and depth information.

This is the case of [177], where the system primarily uses depth information, but when it is not available, RGB information is used instead. In that case, images are converted to grayscale and pictures are formed by adding up the difference between consecutive frames. Then, features are extracted at three levels. At the pixel level, where gradients are calculated, at the patch level, where adaptive patches are determined, and at the global level, where a pyramid structure is used to combine patch features from the previous level. The technique is fully described in [278].

A different approach to the same idea is tried in [191], where depth information is derived from monocular images as presented in [140]. This algorithm uses monocular visual cues, such as texture variations, texture gradients, defocus and color/haze. It mixes all these features with range information derived from a laser range finder to generate, through a Markov random field (MRF) model, a depth map. This map is assembled by splitting the image into patches of similar visual cues and assigning them depth information that is related to the one associated with other image patches. Then, and to segment foreground from background, as the human silhouette has an almost constant depth, a particle swarm optimization (PSO) method is used to discover the optical window in which the variance of the image depth is minimum. This way, patches whose depth information is within the band previously defined are segmented as foreground.

This method, first introduced in [279], was designed to simulate collective behaviors like the ones observed in flocks of birds or swarms of insects. It is an iterative method where particles progressively seek optimal values. This way, in every iteration, depth values with the minimum variance associated with connected patches are approximated, increasing until an optimal value is reached.

Space–Time Methods

All previously presented descriptors abstract information linked to specific frames and, therefore, they should be considered as static data, which clustered along time, acquire a dynamic dimension.

Some methods, however, present visual information where the time component is already inserted and, therefore, dynamic descriptors could be inferred from them.

That is the case of the motion history image (MHI) process. Through this method, after silhouette segmentation, a 2-D representation of its movement, which can be used to estimate if the movement has been fast or slow, is built up. It was first introduced by Bobick et al. [280] and reflects motion information as a function of pixel brightness. This way, all pixels represent moving objects bright with an intensity function of how recent movement is. This technique

is used in [144], [145] and [220] to complement other static descriptors and introduce the time component.

Some systems, as the one introduced in [169], split the global MHI feature in sub-MHIs that are linked to the bounding boxes created to track people. This way, a global feature like MHI is actually divided into parts, and the information contained in each one of them is associated with the specific silhouette responsible for the movement. Through this procedure, the system is able to locally capture movement information and, therefore, able to handle several persons at the same time.

Optical Flow

Optical flow can be defined as the perceived motion of elements between two consecutive frames of a video clip resulting from the relative changes in angle and distance between the objects and the recording camera.

OF, as MHI, is a characterization feature that integrates the time dimension in the information abstraction process and, therefore, a dynamic descriptor.

A number of methods to obtain OF have been developed, being the Lucas–Kanade–Tomasi (LKT) feature tracker, presented in [281] and [282], the most used one. This is the OF obtaining procedure used in all the studied systems which use this feature as a dynamic descriptor.

Two main approaches are considered to obtain OF, sparse, where only relevant points are followed, and dense, where all image pixels are taken into consideration to collect their flow vectors.

In [145], [152], [160], [203], [211], [214] and [216], a dense OF is created that will be used as one of the image characteristic features from which descriptors can be extracted.

Some of these systems obtain OF from segmented objects, as is the case in [145], where, after silhouette segmentation, an OF is derived, and its motion co-occurrence feature (MCF), which is the modulus/direction histogram of the OF, is used for classification.

The system in [152] also extracts a dense OF from segmented objects. In this case, after OF determination, it distributes flow vectors on a circle in accordance with their direction. The resulting product is a Von Mises distribution of the OF flow vectors, which is used as the characterization feature for classification.

In some of the studied systems, as the one presented in [211], the dense optical flow is used as the input of a neural network to generate movement descriptors.

In [150], a sparse OF of relevant points on the silhouette edge is derived, and their vertical velocity will be used as a relevant descriptor for fall identification.

OF has proven to be a very robust and effective procedure to segment the foreground, especially in situations where backgrounds are dynamic, as is the case in fall detection systems mounted on robots that patrol an area searching for fallen people.

Feature Descriptors

Local binary patterns (LBP), as used in [146], is an algorithm for feature description. In this technique, an operator iterates over all image pixels and thresholds its neighborhood with the pixel's own value. This way, a binary pattern is composed. Occurrence histograms based on resulted binary patterns of the entire image, or a part of it, are used as feature descriptors.

Local binary pattern histograms from three orthogonal planes (LBP-TOP) are a further development of the LBP concept. They incorporate time and, therefore, movement in the descriptor, transforming it into a dynamic one. This technique computes each pixel LBP over time, building, this way, a three-dimensional characterization of the video signal by integrating space and temporal properties.

The system described in [219] takes, as input for characterization, a video signal which has gone through multilayered compressed sensing (CS) algorithm and that, therefore, has lost information, especially in low contrast areas. To overcome this difficulty, the system obtains the optical flow of the video signal after the CS process has taken place, and the LBP-TOP is applied over that OF, highly improving the characterization this way. As the video quality is so poor, the OF extraction based on pixel motion is ineffective. To obtain it, low rank and sparse decomposition theory, also known as robust principal component analysis (RPCA) [283], is used to reduce noise. This technique is a modification of the statistical method of principal component analysis whose main objective is to separate, in a corrupted signal, a video one, in this case, the real underlying information contained in the original image from the sparse errors introduced by the CS process.

The histogram of oriented gradients (HOG), as used in [146], is another feature descriptor technique introduced by N. Dalal et al. [284] in the field of human detection with success. The algorithm works over grayscale images using edge detection to determine object positions. This approach uses gradient as the main identification feature to establish where body edges are. It takes advantage of the fact that gradients will sharply rise at body edges in comparison with the mean gradient variation of the area they are placed in. To identify those boundaries, a mask is applied on each pixel and gradients are determined through element-wise multiplication. Histograms of gradient orientation are then created for each block, and, in the final stages of the process, they are normalized both locally and globally. These histograms are used as image feature descriptors.

The system proposed in [199] incorporates HOGs as the image descriptor, which, in later stages of the identification algorithm, is used by an ANN to determine whether a fall has occurred.

2.3.3.1.2 Local descriptors

Local descriptors approach the problem of pose and movement abstraction in a different way. Instead of segmenting the foreground and extracting characteristic features from it, encoding them as a block, they focus on area patches from which relevant local features, characteristic of human movement or human pose, can be derived.

Over time, local descriptors have substituted or complemented global ones, as they have proofed to be much more resistant to noise or partial occlusion.

Characterization feature techniques focused on fall detection, pay attention to head motion, body shape changes and absence of motion [285]. The system introduced in [209] uses the two first groups of features. It models body shape changes and head motion by using the extended CORE9 framework [286]. This framework uses minimum bounding rectangles to abstract body movements. The system slaves bounding boxes to legs, hands and head, which is taken as the reference element. Then, directional, topological, and distance relations are established between the reference element and the other ones. All this information is finally used for classification purposes.

The vast majority of studied systems that implement local descriptors do it using ANNs. ANNs are a major research area at the moment, and their application to the artificial vision

and human activity recognition is a hot topic. These networks, which simulate biological neural networks, were first introduced by Rosenblatt [287] through the definition of the perceptron in 1958.

There are two main families of ANNs with application in artificial vision, human pose estimation and human fall detection, which have been identified in this research. These two families are convolutional neural networks (CNN) and recurrent neural networks (RNN).

ANNs are able to extract feature maps out of input images. These maps are local descriptors able to characterize the different local patches that integrate an image.

RNNs are connectionist architectures able to grasp the dynamics of a sequence due to cycles in its structure. Introduced by Hopfield [288], they retain information from previous states and, therefore, they are especially suitable to work with sequential data when its flow is relevant. This effect of information retention through time is obtained by implementing recurrent connections that transfer information from previous time steps either to other nodes or to the originating node itself.

Among RNNs architectures, long short-term memory (LSTM) ones are especially useful in the field of fall detection. Introduced by Hochreiter [289], LSTMs most characteristic feature is the implementation of a hidden layer composed of an aggregation of nodes, called memory cells. These items contain nodes with a self-linked recurrent connection, which guarantees information will be passed along time with no vanishing. Unlike other RNNs, whose long-term memory materializes through weights given to inputs, which change slowly during training, and whose short-term memory is implemented through ephemeral activations, passed from a node to the successive one, LSTMs introduce an intermediate memory step in the memory cells. These elements internally retain information through their self-linked recurrent connections, which include a forget gate. Forget gates allow the ANN to learn how to forget the contents of previous time steps.

LSTM topologies, like the one implemented in [205], allow the system to recall distinctive features from previous frames, incorporating, this way, the time component to the image descriptors. In this particular case, an RNN is built by placing two LSTM layers between batch normalization layers, whose purpose is to make the ANN faster. Finally, a last layer of the network, responsible for classification, implements a Softmax algorithm.

Some LSTMs architectures, as the one described in [199], are used to determine characteristic foreground features. This ANN is able to establish a silhouette center and establish angular speed, which will be used as a reference to determine whether a fall event has taken place.

The system proposed in [204] includes several LSTM layers. This encoding-decoding architecture integrates an encoding block, which encodes the input data, coming from a CNN block used to identify joints and estimate body pose, to a vector of fixed dimensionality, and a decoding block, composed of a layer able to output predictions on future body poses. This architecture is based on the seq2seq model proposed in [290] and has been successfully used in this system with prediction purposes, substantially reducing fall detection time, as it is assessment is made on a prediction, not on observation.

A specific LSTM design is the bidirectional one (Bi-LSTM). This architecture integrates two layers of hidden nodes connected to inputs and outputs. Both layers implement the idea of information retention through time in a different way. While the first layer has recurrent connections, in the second one, connections are flipped and passed backward through the

activation function signal. This topology is incorporated in [232], where Bi-LSTM layers are stacked over CNN layers used to segment incoming images.

CNNs were inspired by the neural structure of the mammal visual system, very especially by the patterns proposed by Hubel et al. [291]. The first neural network model with visual pattern recognition capability was proposed by Fukushima [292], and, based on it, LeCun and some collaborators developed CNNs with excellent results in pattern recognition, as shown in [293] and [294].

This family of ANNs is assembled by integrating three main types of layers; convolutional, pooling and fully connected, each one of them playing a different role. Every layer of the CNN receives an input, transforms it and delivers an output. This way, the initial layers, which are convolutional ones, deliver feature maps out of the input images, whose complexity is reduced by the pooling layers. Eventually, these maps are led to the fully connected layers, where the feature maps are converted into vectors used for classification.

A typical CNN architecture is shown in figure 2.

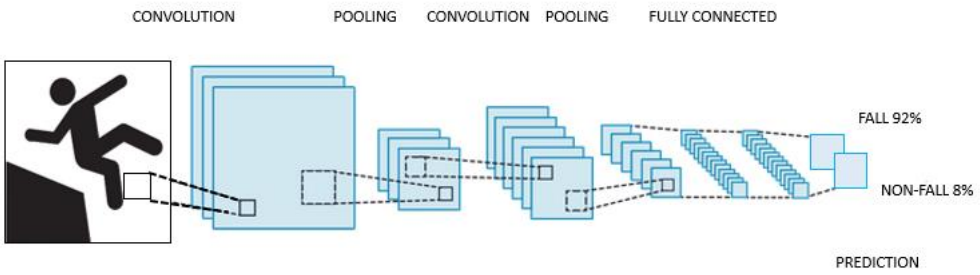


Figure 2. Typical convolutional neural network (CNN) architecture.

Some systems, like the one in [234], where a YoLOv3 CNN is used, take the input image and modify its scale to get several feature maps out of the same image. In this case, the CNN is used to generate three different sets of feature maps, based on three image scales, which eventually, after going through the fully connected layers, will be used for classification.

A similar approach is used in [225], where a YoLOv3 CNN identifies people. Identified people are tracked, and a CNN ANN extracts characteristic features from each person in the image. The feature vectors are passed to an LSTM ANN whose main task is to retain features over time so the temporal dimension can be added to the spatial features obtained by the CNN. The final feature vectors, coming out of the LSTM layers, are sent to a fully connected layer, which implements a Softmax algorithm used for event classification.

In [215], the object detection task, performed by a YoLO CNN, is combined with object tracking, a task developed by DeepSORT [295], a CNN architecture able to track multiple objects after they have been detected.

The approach made in [210] to detect a fallen person uses a YoLOv3 CNN to detect fallen bodies on the ground plane. It maximizes the sensitivity by turning 90 and 270 degrees all images and compare the bounding boxes found in the same image. Then, features are extracted from the bounding box, which will be used as classification features.

In [206] and [214], a wide residual network, which is a type of CNN, takes as input an OF and derives feature maps out of it. These maps are delivered to the fully connected layers, which, in turn, will pass vectors for movement classification to the last layers of the ANN.

A similar procedure is followed by the system in [217], whose ANN mixes layers of CNN, which deliver features maps from the incoming binarized video signal, with layers of radial basis function neural networks (RBFNN), which will be used as a classifier.

Another interesting type of CNN is the hourglass convolutional auto-encoder (HCAE), introduced in [231]. This kind of architecture piles convolutional and pooling layers over fully connected ones to get a feature vector, and then it follows the inverse process to reconstruct the input images. The HCAE compares the error value between the encoded-decoded frames and the original frames, applying back-propagation for self-tuning. Ten consecutive frames are inputted into the system to guarantee it captures both image and action features.

An alternate approach is the one presented in [194], where a CNN identifies objects (including people) and associate vectors to them. These vectors, which measure features, characterize both the human shape itself and its spatial relations with surrounding objects. This way, events are classified not only as a function of geometrical features of the silhouette but also as a function of its spatial relations with other objects present in the image. This approach has proven very useful to detect incomplete falls where pieces of furniture are involved.

A good number of approaches, as in [198], use 3D CNNs to extract spatiotemporal features out of 2D images, like the ones used in this system. This way, ANNs are used not only to extract spatial features associated with pose recognition but also to capture the temporal relation established among successive poses leading to a fall. The system in [180] uses this approach, creating a dynamic image by fusing in a single image all the frames belonging to a time window and passing this image to the ANN as the input from where extracting features.

Certain convolutional architectures, like the ones integrated into OpenPose and used in [215] and [218], can identify human body key points through convolutional pose machines (CPM), a CNN able to identify those features. These key points are used to build a vector model of the human body in a bottom-up approach as shown in figure 3.

Figure 3. Convolutional pose machine presentation.

To correct possible mistakes, this approximation is complemented in [218] by a top-down approach through single shot multibox detector-MobileNet (SSD-MobileNet), another convolutional architecture able to identify multiple objects, human bodies in this case. SSD-MobileNet, lighter and requires less computational power than typical SSDs, is used to remove all key points identified by OpenPose not being part of a human body, correcting this way, inappropriate body vector constructions.

A similar approach is used in [221], where a CNN is used to generate an inverted pendulum based on five human key points, knees, center of the hip line, neck and head. The motion history of these joints is recorded, and a subsequent module calculates the pendulum rotation energy and its generalized force sequences. These features are then codified in a vector and used for classification purposes.

The system in [233] uses several ANNs and selects the most suitable one as a function of the environment and the characteristics of the tracked people. In addition, it uploads wrongly categorized images, which are used to retrain the used models.

2.3.3.1.3 Depth descriptors

Descriptors based on depth information have gained ground thanks to the development of low-cost depth sensors, such as Microsoft Kinect®. This affordable system counts with a software development kit (SDK) and applications able to detect and track joints and construct human body vector models. These elements, together with the depth information from stereoscopic scene observation, have raised great interest among the artificial vision research community in general and the human fall detection system developers in particular.

A good number of the studied systems use depth information, solely or together with RGB one, as the data source in the abstraction process leading to image descriptor construction. These systems have proven to be able to segment foreground, greatly diminishing interference due to illumination interferences up to the distance where stereoscopic vision procedures are able to infer depth data. Fall detection systems use this information either as depth maps or skeleton vector models.

Depth Map Representation

Depth maps, unlike RGB video signals, contain direct three-dimensional information on objects in the image. Therefore, depth map video signals integrate raw 3D information, so three-dimensional characterization features can be directly extracted from them.

This way, the system described in [174] identifies 16 regions of the human body marked with red tape and position them in space through stereoscopic techniques. Taking that information as a base, the system builds the body vector (aligned with spine orientation) and identifies its center of gravity (COG). Acceleration of COG and body vector angle on a vertical axis will be used as features for classification.

Foreground segmentation of human silhouette is made by these systems through depth information, by comparing depth data from images and a reference established at system startup. This way, pixels appearing in an image at a distance different from the one stored for that particular pixel in the reference are declared as foreground. This is the process followed by [172] to segment the human silhouette. In an ulterior step, descriptors based on bounding box, centroid, area and orientation of the silhouette are extracted.

Other systems, like the one in [229], extract background by using the same process and the silhouette is determined as the major connected body in the resulting image. Then, an ellipse

is established around it, and classification will be made as a function of its aspect ratio and centroid position. A similar process is followed in [188], where, after background subtraction, an ellipse is established around the silhouette, and its centroid elevation and velocity, as well as its aspect ratio, are used as classification features.

The system in [185] uses depth maps to segment silhouettes as well and creates a bounding box around them. Box top coordinates are used to determine the head velocity profile during a fall event, and its Hausdorff distance to head trajectories recorded during real fall events is used to determine whether a fall has taken place. The Hausdorff distance quantifies how far two subsets of a metric space are from each other. The novelty of this system, leaving aside the introduction of the Hausdorff distance as described in [296], is the use of a moving capture (MoCap) technique to drive a human model using software to simulate its motion (OpenSim), so profiles of head vertical velocities can be captured in ADLs, and a database can be built. This database is used, by the introduction of the Hausdorff distance, to assess falls.

The system in [213], after foreground extraction by using depth information as in the previous systems, transforms the image to a black and white format and, after de-noising it through filtering, calculates the HOG. To do it, the system determines the gradient vector and its direction for each image pixel. Then, a histogram is constructed, which integrates all pixels' information. This is the feature used for classification purposes.

In [170], silhouettes are tracked by using a proportional-integral-differential (PID) controller. A bounding box is created around the silhouette, and features are extracted in accordance with [297]. A fall will be called if thresholds established for features are exceeded. Faces are searched, and when identified, the tracking will be biased towards them.

Some other systems, like the one in [143], subtracts background by direct use of depth information contained in sequential images, so the difference between consecutive depth frames is used for segmentation. Then, the head is tracked, so the head vertical position/person height ratio can be determined, which, together with COG velocity, is used as a classification feature.

In [182], all background is set to a fixed depth distance. Then, a group of 2000 body pixels is randomly chosen, and for each of them, a vector of 2000 values, calculated as a function of the depth difference between pairs of points, is created. These pairs are determined by establishing 2000 pixel offset sets. The obtained 2000-value vector is used as a characteristic feature for pose classification.

The system introduced in [139], after the human silhouette is segmented by using depth information through a GMM process, calculates its curvature scale space (CSS) features by using the procedures described in [140]. CSS calculation method convolutes a parametric representation of a planar curve, silhouette edge in this case, with a Gaussian function. This way, a representation of the arc length vs. curvature is obtained. Then, silhouettes features are encoded, together with the Gaussian mixture model used in the aforementioned CSS process, in a single Fisher vector, which will be used, after being normalized, for classification purposes.

Finally, a block of systems creates volumes based on normal distributions constructed around point clouds. These distributions, called voxels, are grouped together, and descriptors are extracted out of voxel clusters to determine, first, whether they represent a human body and then to assess if it is in a fallen state.

This way, the system presented in [155] first estimates the ground plane by assuming that most of the pixels belonging to every horizontal line are part of the ground plane. The ground can then be estimated, line per line, attending to the pixel depth values as explained in the procedure described in [298]. To clean up the pictures, all pixels below the ground plane are discarded. Then, normal distributions transform (NDT) maps are created as a cloud of points surrounded by normal distributions with the physical appearance of an ellipsoid. These distributions, created around a minimum number of points, are called voxels and, in this system, are given fixed dimensions. Then, features that describe the local curvature and shape of the local neighborhood are extracted from the distributions. These features, known as IRON [299], allow voxel classification as being part of a human body or not and, this way, voxels tagged as human are clustered together. IRON features are then calculated for the cluster representing a human body, and the Mahalanobis distance between that vector and the distribution associated with fallen bodies is calculated. If the distance is below a threshold, the fall state is declared.

A similar process is used in [162], where, after the point cloud is truncated by removing all points not contained in the area in between the ground plane and a parallel one 0.7 m over it by applying the RANSAC procedure [300], NDTs are created and then segmented in patches of equal dimensions. A support vector machine (SVM) classifier determines which ones of those patches belong to a human body as a function of their geometric characteristics. Close patches tagged as humans are clustered, and a bounding box is created around. A second SVM determines whether clusters should be declared as a fallen person. This classification is refined, taking data from a database of obstacles of the area, so if the cluster is declared as a fallen person, but it is contained in the obstacle database, the declaration is skipped.

Skeleton Representation

Systems implementing this representation are able to detect and track joints and, based on that information, they can build a human body vector model. This block of techniques, as the previous one, strongly diminishes the noise associated with illumination but have problems to build a correct model when occlusion appears, both the one generated by obstacles and the one product of perspective auto-occlusions.

A good number of these systems are built over the Microsoft Kinect® system and take advantage of both de SDK and the applications developed for it. This is the case of the system introduced in [168], where three Kinect® systems cover the same area from different perspectives, and joints are, therefore, followed from different angles, reducing this way the tracking problems associated with occlusion. In this system, human movement is characterized through two main features, head speed and CG situation referenced to ankles position.

The Kinect® system is also used in [193] to follow joints and estimate the vertical distance to the ground plane. Then, the angle between the vertical and the torso vector, which links the neck and spine base, is determined and used to identify a start keyframe (SKF), where a fall starts, and an end keyframe (EKF), where it ends. During this period, vertical distance to the ground plane and vertical velocity of followed upper joints will be the input for classification. A very similar approach is followed in [161], where torso/vertical angle and centroid height are the key features used for classification.

This system is used as well in [133] to build, around identified joints, both 2D and 3D bounding boxes aligned with the spine direction. Then, the ratio width/height is determined, and the relation HCG/PCG, being the former de elevation of the COG over the ground plane

and the latter de distance between the COG projection on the ground and the support polygon defined by ankles position, is calculated. Those features will be the base for event classification.

In [301], human body key points are identified by a CNN whose input is a 2D RGB video signal complemented by depth information. Based on those key points, the system builds a human body vector model. A filter was developed to generate digital terrain models from data captured by airborne systems [302], and the depth data were then used to estimate the ground plane. The system uses all that information to calculate the distance from the body CG and the body region over the shoulders to the ground. These distances will serve to characterize the human pose.

A CNN is also used in [189] to generate feature maps out of the depth images. This network stacks convolution layers to extract features and pooling layers to reduce map complexity, with a philosophy identical to the one used in the RGB local characterization. The output map goes through two layers of fully connected layers to classify the recorded activity, and a Softmax function is implemented in the last layer of the ANN, which determines whether a fall has taken place.

In [212], prior to input images in a CNN to generate feature maps, which will be used for classification, the background is subtracted through an algorithm that combines depth maps and 2D images to enhance segmentation performance. This way, if the pixels of the segmented 2D silhouette experiment sharp changes, but pixels in the depth map do not, pixels subject to those changes are regarded as noise. The system mixes information from both sources, allowing a better track on segmented silhouettes and a quick track regain in case it is lost.

The system in [141]—after identifying human body joints as the key features whose trajectory will be used to determine whether a falling event has taken place—proposes rotating the torso so it is always vertical. This way, joint extraction becomes pose invariant, a technique used in the system with positive results in order to deal with the noise associated with joint identification as a result of rapid movement and occlusion, characteristic of falls.

2.3.3.2 Classification

Once pose/movement abstract descriptors have been extracted from video images, the next step of the fall detection process is classification. In broad terms, during this phase, the system classifies movement and or pose as a fall or a fallen state through an algorithm that is part of one of these two categories; generative or discriminative models.

Discriminative models are able to determine boundaries between classes; either by explicitly being given those boundaries or by setting them themselves using sets of pre-classified descriptors.

Generative models approach the classification problem in a totally different way, as they explicitly model the distribution of each class and then use the Bayes theorem to link descriptors to the most likely class, which, in this case, can only be a fall or a not fall state.

2.3.3.2.1 Discriminative Models

The final goal of any classifier is assigning a class to a given set of descriptors. The discriminative models are able to establish the boundaries separating classes, so the probability of a descriptor belonging to a specific class can be given. In other terms, given α as a class, and [A] as the matrix of descriptor values associated with a pose or movement, this family of classifiers is able to determine the probability $P(\alpha|[A])$.

Feature-Threshold-Based

Feature-threshold-based classification models are broadly used in the studied systems. This approach is easy and intuitive, as the researcher establishes threshold values for the descriptors, so their associated events can be assigned to a specific class in case those thresholds are exceeded.

This is the case of the system proposed in [159]. It classifies the action as a fall or a non-fall in accordance with a double rationale. On one hand, it establishes thresholds of ellipse features to estimate whether the pose fits a fallen state; on the other, an MHI feature exceeding a certain value indicates a fast movement and, therefore, a potential fall. The system proposed in [142] adds acceleration to the former features and, in [168], head speed over a certain threshold and COG position out of the segment defined by ankles are indicatives of a fall.

Similar approaches, where threshold values are determined by system developers based on previous experimentation, are implemented in a good number of the studied systems, as they are simple, intuitive and computationally inexpensive.

Multivariate Exponentially Weighted Moving Average

Multivariate exponentially weighted moving average (MEWMA) is a statistical process control to monitor variables that use the entire history of values of a set of variables. This technique allows designers to give a weighting value to all recorded variable outputs, so the most recent ones are given higher weight values, and the older ones are weighted lighter. This way, the last value is weighted λ (being λ a number between 0 and 1) and previous β values are weighted λ^{β}. Limits to the value of that weighted output are established, taking as a basis the expected mean and standard deviation of the process. Certain systems, like [156], use this technique for classification purposes. However, as it is unable to distinguish between falling events and other similar ones, events tagged as fall by the MEWMA classifier need to go through an ulterior support vector machine classifier.

Support Vector Machines

Support vector machines (SVM) are a set of supervised learning algorithms first introduced by Vapnik et al. [303].

SVMs are used for regression and classification problems. They create hyperplanes in high dimension spaces that separate classes nonlinearly. To fulfill this task, SVMs, similar to artificial neural networks, use kernel functions of different types.

A standard SVM boundary definition is shown in figure 4.

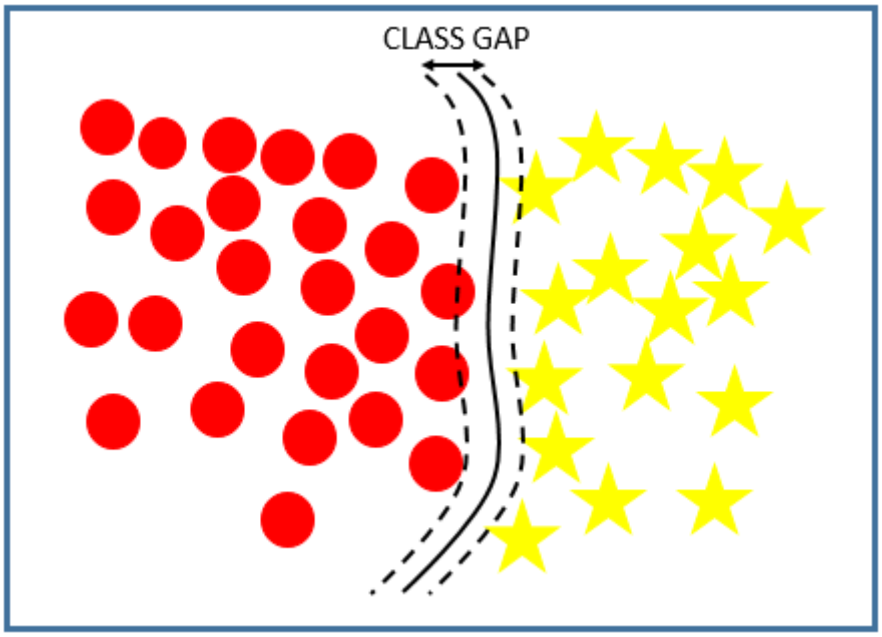


Figure 4. Support vector machine boundary definition.

In [202], linear, polynomial, and radial kernels are used to obtain the hyperplanes; in [195], radial ones are implemented, and in [176], polynomial kernels are used to achieve nonlinear classifications.

The support vector data description (SVDD), introduced by Tax et al. [304], is a classifying algorithm inspired by the support vector machine classifier, able to obtain a spherically shaped boundary around a dataset and, analogously to SVMs, it can use different kernel functions. The method is made robust against outliers in the training set and is capable of tightening classification by using negative examples. SVDDs classifying algorithms are used in [218].

SVMs have been very used in the studied systems as they have proofed to be very effective; however, they require high computational loads, something inappropriate for edge computing systems.

K-Nearest Neighbor

K-nearest neighbor (KNN) is an algorithm able to model the conditional probability of a sample belonging to a specific class. It is used for classification purposes in [144], [145], [176] and [202] among others.

KNNs assume that classification can be successfully made based on the class of the nearest neighbors. This way, if for a specific feature, all μ closest sample neighbors are part of a determined class, the probability of the sample being part of that class will be assessed as very high. This study is repeated for every feature contained in the descriptor, so a final assessment based on all features can be made. The algorithm usually gives different weights to the neighbors, and heavier weights are assigned to the closest ones. On top of that, it also assigns different weights to every feature. This way, the ones assessed as most relevant get heavier weights.

Decision Tree

Decision trees (DT) are algorithms used both in regression and classification. It is an intuitive tool to make decisions and explicitly represents decision-making. Classification DTs use categorical variables associated with classes. Trees are built by using leaves, which represent class labels, and branches, which represent characteristic features of those classes. DTs built process is iterative, with a selection of features correctly ordered to determine the split points that minimize a cost function that measures the computational requirements of the algorithm. These algorithms are prone to overfitting, as setting the correct number of branches per leaf is usually very challenging. To reduce the complexity of the trees, and therefore, their computational cost, branches are pruned when the relation cost-saving/accuracy loss is satisfactory. This type of classifier is used in [215] and [217].

Random forest (RF), like the one used in [182] and [215], is an aggregation technique of DT, introduced by Braiman [305], which main objective is avoiding overfitting. To accomplish this task, the training dataset is divided into subgroups, and therefore, a final number of DTs, equal to the number of dataset subgroups, is obtained. All of them are used in the process, so the final classification decision is actually a combination of the classification of all DTs.

Gradient boosting decision trees (GBDT) is another DT aggregation technique whose algorithm was first introduced by Friedman [306] where simple DTs are built and, for each one of them, a classification error in training time is determined. An error function based on calculated individual errors is determined, and its gradient is minimized by combining individual DT classifications in a proper way. This aggregation technique, specifically developed for DTs, is actually part of a broader family that will be more extensively presented in the next section.

Both techniques, RF and GBDT, are used in [215].

Boost Classifier

Boost classifier algorithms are a family of classifier building techniques that create strong classifiers by grouping weak ones. It is done by adding up models created from the training data until the system is perfectly predicted or a maximum number of models is reached.

This is done by building a model from the training data. Then, a second model is created to correct the errors from the first one. Models are added until the training set is well predicted or a maximum number of them is added. During the boosting process, the first model is trained on the entire database while the rest are fitted to the residuals of the previous ones.

Adaboost, used in [151], can be utilized to increase performances with any classification technique, but it is most commonly used with one-level decision trees.

In [192], boosting techniques are used on a J48 algorithm, a tree-based technique, similar to random forest, which is used to create univariate decision trees.

Sparse Representation Classifier

Sparse representations classification (SRC) is a technique used for image classification with a very good degree of performance.

Natural images are usually rich in texture and other structures that tend to be recurrent. For this reason, sparse representation can be successfully applied to image processing. This

phenomenon is known as patch recurrence and, because of it, real-world digital images can be recognized by properly trained dictionaries.

SRCs are able to recognize those patches, as they can be expressed as a linear combination of a limited number of elements that are contained in the classifier dictionaries.

This is the case of the SRC presented in [152].

Logistic Regression

Logistic regression is a statistical model used for classification. It is able to implement a binary classifier, as the one needed to decide whether a fall event has taken place. For such a purpose, a logistic function is used. It can be adjusted by using classifying features associated with events tagged as fall or not fall.

This method is used in systems like [221], where a logistic classifying algorithm is employed to classify events as fall or not a fall, based on a vector that encodes the temporal series of rotation energy and generalized force.

Some artificial neural networks implement a logistic regression function for classification, as the one described in [234], where a CNN uses this function to determine the detection probability of each defined class.

Deep Learning Models

In [211], the last layers of the ANN implement a Softmax function, a generalization of the logistic function used for multinomial logistic regression. This function is used as the activation function of the nodes of the last layer of a neural network, so its output is normalized to a probability distribution over the different output classes. Softmax is also implemented in the last layers of the artificial neural networks used in [203] and [231], among other studied systems.

Multilayer perceptron (MLP) is a type of multilayered ANN with hidden layers between the entrance and the exit ones able to sort out classes non-linearly separable. Each node of this network is a neuron that uses a nonlinear activation function, and it is used for classification purposes in [176] and [215].

Radial basis function neural networks (RBFNN) are used in the last layer of [217] to classify the feature vectors coming from previous CNN layers. This ANN is characterized by using radial basis functions as activation functions and yields better generalization capabilities than other architectures, such as Softmax, as it is trained via minimizing the generalized error estimated by a localized-generalization error model (L-GEM).

Often, the last layers of ANN architectures are fully connected ones, as in [186], [204] and [214], where all nodes of a layer are connected to all nodes in the next one. In these structures, the input layer is used to flatten outputs from previous layers and transform them into a single vector, while subsequent layers apply weights to determine a proper tagging and, therefore, successfully classify events.

Finally, another ANN structure useful for classification is the autoencoder one, used in [198]. Autoencoders are ANNs trained to generate outputs equal to inputs. Its internal structure includes a hidden layer where all neurons are connected to every input and output node. This way, autoencoders get high dimensional vectors and encode their features. Then, these features are decoded back. As the number of dimensions of the output vector may be

reduced, this kind of ANNs can be used for classification purposes by reducing the number of output dimensions to the number of final expected classes.

2.3.3.2.2 Generative Models

The approach of generative models to the classification problem is completely different from the one followed by the discriminative ones.

Generative models explicitly model the distribution of each class. This way, given α as a class, and [A] as the matrix of descriptor values associated with a pose or movement, if both P ([A]|α) and P (α) can be determined, it will be possible, by direct application of the Bayes theorem, to obtain P (α|[A]), which will solve the classification problem.

Hidden Markov Model

Classification using the hidden Markov model (HMM) algorithm is one of the three typical problems that can be solved through this procedure. It was first proposed with this purpose by Rabiner et al. [307] to solve the speech recognition problem, and it is used in [228] to classify the feature vectors associated with a silhouette.

HMMs are stochastic models used to represent systems whose state variables change randomly over time. Unlike other statistical procedures, like Markov chains, which deal with fully observable systems, HMMs tackle partially observable systems. This way, the final objective of the HMM classifying problem resolution will be decided, based on the observable data (feature vector), whether a fall has occurred (hidden system state).

The system proposed in [228] determines, using an HMM as a classifier, on the basis of silhouette surface, centroid position and bounding box aspect ratio, whether a fall takes place or not. To do it, and to take as a reference recorded falls, a probability is assigned to the two possible system states (fall/not fall) based on value and variation along the event timeframe period of the feature vector. This classifying technique is used with success in this system, though in [308], a brief summary of the numerous limitations of this basic HMM approach is presented, and several more efficient extensions of the algorithm, such as variable transition HMM or the hidden semi-Markov model, are introduced. These algorithm variations are developed as the basic HMM process is considered ill suited for modeling systems where interacting elements are represented through a vector of single state variables.

A similar classification approach using an HMM classifier is used in [175], where future states predicted by an autoregressive-moving-average (ARMA) algorithm are classified as fall or not-fall events. ARMA models are able to predict future states of a system based on a previous time-series. The model integrates two modules, an autoregressive one, which uses a linear combination of weighted previous system state values, and a moving average one, which linearly combines weighted previous errors between system state real values and predicted ones. In the model, errors are assumed to be random values that fit a Gaussian distribution of mean 0 and variance σ^2.

2.3.3.3 Tracking

A good number of the reviewed systems identify objects through ANN or extract silhouettes from the background. Then, relevant features are associated with the already segmented objects. This assignment requires a constant update, and, therefore, object correlation needs to be established from frame-to-frame. This correlation is made through object tracking, and a good number of different techniques are used for such a purpose.

2.3.3.3.1 Moving Average Filter

The double moving average filter used in [193] smooths vertical distance from joints to the ground plane. This filter determines twice the mean value of the last n samples, acting this way as a low pass filter, eliminating high-frequency signal components associated with noise.

2.3.3.3.2 PID Filter

The system proposed in [170] uses a proportional-integral-differential (PID) filter to maintain tracking on silhouettes segmented from the background. Constants of the filter to guarantee smooth tracking, reducing overshoots and steady-state errors, are calculated through a genetic algorithm. This algorithm, inspired by the theory of natural evolution, is a heuristic search where sets of values are selected or discarded based on its ability to reduce to a minimum the absolute error function and, therefore, minimize overshoots and steady errors.

2.3.3.3.3 Kalman Filter

Kalman filter, first introduced by R. E. Kalman in [309], is a recursive algorithm that allows improvements in the determination of system variable values by combining several sets of indirect system variable observations containing inaccuracies. The resulting estimation is more precise than any of the ones, which could be inferred from a single indirect observation set.

This way, in [168], the tracking of joints, followed by three independent Kinect® systems, is fused by a Kalman filter. The resulting joint position is estimated by integrating information from the three systems and is more accurate than one of any of the individual systems.

A particular variation in the use of Kalman filtering is the one in [225], where a procedure call deep-sort, presented in [295], is used. In this process, a Kalman algorithm is used to estimate the next location of the tracked person, and then the Mahalanobis distance is calculated between the detected person in the following frame and its estimated position. By measuring this distance, uncertainty in the track correlation can be quantified. This filter performance is deeply affected by occlusion. To mitigate this problem, the uncertainty value is associated with the track descriptor and, to keep tracks after long occlusion periods, the process saves those descriptors for 100 frames.

Although this filtering algorithm works very well to maintain tracks in linear systems, human bodies involved in a fall tend to behave nonlinearly, substantially degrading its ability to maintain tracking.

2.3.3.3.4 Particle Filter

This method, used in [143], is a Monte Carlo algorithm used for object tracking in video signals. Introduced in 1993 by Gordon [310] as a Bayesian recursive filter, it is able to determine future system states, in this case, future positions of the tracked object.

The filter algorithm follows an iterative approach. This way, after a cloud of particles, image pixels, in this case, have been selected, weights are assigned to them. Those weigh values are a function of the probability of being part of the tracked object. Then, the initial particle cloud is updated by using the weight values. Based on object cinematic, its movement is propagated to the particle cloud, predicting, this way, the future object situation. The process continues with a new update phase to guarantee the predicted cloud matches the tracked object.

This algorithm, although affected by occlusion, has proven to be highly capable of maintaining tracks on objects moving nonlinearly and, therefore, the result is adequate to track human bodies during fall events.

Rao–Blackwellized particle filter (RBPF), as the one used in [191], is a type of particle filter tracking algorithm used in linear/nonlinear scenarios where a purely Gaussian approach is inadequate.

This algorithm divides particles into two sets. Those that can be analytically evaluated and those that cannot. This way, the filtering equations are separated into two sets, so two different approaches can be used to calculate them. The first set, which includes linear moving particles, is solved by using a Kalman filter approach, while the second one, whose particles move nonlinearly, is solved by employing a Monte Carlo sampling method.

2.3.3.3.5 Fused Images

In [137], a fusing center fuses images taken from orthogonal views, and the obtained object is tagged with a number. Objects identified in the next frame are correlated to previous ones if they meet the minimum distance established threshold. This way, the tracking is maintained.

2.3.3.3.6 Camshift

This algorithm, integrated into OpenCV and used in [187], first converts images RGB to hue-saturation-value (HSV) and, starting with frames where a CNN has created a bounding box (BB) around a detected person, it determines the hue histogram in each BB. Then, morphological operations are applied to reduce noise associated with illumination. In the consecutive frame, the area that better fits the recorded Hue histogram is established and compared with detected BBs. That way, a correlation can be established and, therefore, a track on a person.

2.3.3.3.7 Deep Learning Architectures

DeepSORT is a CNN used to track multiple objects at the same time, as shown in [215].

The system presented in [199] tracks images using an algorithm as follows: First, in every new frame, a YoLO convolutional architecture is used to identify people. Once all people in the frame have been identified, a Siamese CNN is used to first determine the characteristic features of every person identified in the frame and then compare them with the ones associated with people identified in previous frames, looking for similarities. At the same time, an LSTM ANN is used to predict people's motion, so associations to maintain track of people from frame-to-frame can be made. Based on feature similarity and movement association, a track can be established on people present in consecutive video frames or can be started when a new person appears for the first time in a video sequence. An almost equal process is used in [225] to keep track of people with two CNNs working in parallel, a first one to identify people and a second one to extract characteristic features out of them. That way, tracks can be established.

In [169], a CNN is used to detect people in every frame. A BB is established around, and distances from central point BBs of consecutive frames are determined. Boxes meeting minimum distance criteria in consecutive frames are correlated and, this way, tracking is established.

2.3.3.4 Classifying Algorithms Performances

A number of the reviewed systems establish comparisons with other ones. Many of them base that comparison on performance figures obtained on different datasets, while some others establish a system-to-system comparison based on the same database. However, systems are, in broad terms, an aggregation of two main blocks, the first one whose mission is inferring descriptors from images and a second one that classifies those features. This way, system

comparison, even on the same dataset, compares two aggregated blocks so, comparisons on performances of a specific block is difficult to assess, as it is influenced by the other one.

To avoid these problems, these comparisons have been ignored. The only ones taken into consideration have been those that compare one of the blocks and are based on the same dataset. The results are shown in Table 2. In global terms, SVMs and deep learning classifiers are the ones with better performances. The best working classifying deep learning architectures are MLP, autoencoders and those implementing Softmax algorithms like GoogLeNet. It is also relevant that in accordance with C.J. Chong et al. [134], systems whose descriptors are dynamic and, therefore, include references to the time variable, have better performances than those other ones whose descriptors do not incorporate that variable.

2.3.3.5 Validation datasets

The systems included in this research have been tested by using datasets. On many occasions, those datasets have been specifically developed by the researchers to test and validate their systems, so their performances can be determined. These datasets, although briefly discussed in the articles presenting the systems, are not usually publicly accessible.

However, there is also a group of datasets used in the system validation and performance determination phases that are public. Most of them are also accessible through the Internet, so developers can download and use them for research purposes. All the datasets belonging to this category used in the development of the systems contained in this review are collected in Table 3.

Datasets associated with the reviewed systems, both the publicly accessible ones and the ones that are not, are recorded by either volunteers or actors young and fit enough to guarantee that a simulated fall will not harm them. In some of them, therapists advise actors, so they can imitate how an elderly person moves or falls. Finally, none of the databases includes elderly real falls or daily life activities performed by elderly people.

The datasets are grouped by collected signal type, so five big groups are identified.

- The first group is integrated by a single dataset. It collects falls and activities of daily life (ADL) executed by volunteers whose results are recorded using different sensors, included RGB and IR cameras. It is used by a single system for validation purposes;
- The second group, which includes three datasets, incorporates depth and accelerometric data. By its relevance and number of reviewed systems using it in their performance evaluation, one dataset is especially important, UR fall detection [157]. This dataset, employed by over a third of all studied systems, includes 30 falls and 40 ADLs recorded by two depth systems, one providing frontal images and one recording vertical ones. This information is accompanied by accelerometric data and was released in 2015.
- The third group is composed of nine datasets. They all mix ADLs and falls recorded in different scenarios by RGB cameras, either conventional or fish eye ones, from different perspectives and at different heights. Two of them exceed the mark of 10% users, LE2I [151] and the Multicam Fall Dataset [138].

  LE2I, published in 2013, is a dataset that includes 143 different types of falls performed by actors and 48 ADLs. These events were recorded in environments simulating the ones that could be found in an elderly home.

Multicam includes 24 scenarios recorded with 8 IP cameras, so events can be analyzed from multiple perspectives. Twenty-two of the scenarios contains falls, while the other two only include confounding actions. Volunteers simulate events, and this dataset was released in 2010;

- The fourth group includes six datasets. Different activities, falls included, are recorded by depth systems. The two most used ones are the Fall Detection Dataset [158] and SDUFall [140], though both of them fall below the 10% users mark.

  Fall Detection Dataset, used by almost 10% of the systems, was published in 2017. The images in this dataset are recorded in five different rooms from eight different view angles, and five different volunteers take part in it.

  SDUFall, published in 2014, is another dataset that gathers depth information associated with six types of actions, being a fall one of them. Actions are repeated 30 times by 10 volunteers and are recorded by a depth system;

- The fifth group, composed of a single dataset, collects synthetic information. CMU Graphics Lab—motion capture library [183] is a dataset that contains biomechanical information related to human body movement captured using motion capture (MoCap) technology. To generate that information, groups of volunteers, wearing sensors in different parts of their bodies, execute diverse activities. The information collected by the sensors is integrated through a human body model and stored in the dataset, so it can be used for development purposes. This approach to system development and validation has numerous advantages over conventional methods, as it gives developers the possibility of training their systems under any possible image perspective or occlusion situation. However, clutter and noise, the other important problems for optimal system performance, are not included in the information recorded in this database.

2.3.3.6 Conclusions

The systems based on artificial vision have deeply evolved over the course of the last five years. The amount of effort devoted to the development of this technology applied to this field has been huge and fully in line with the one seen in artificial vision in general. This important research effort, proofed by the number of reviewed published papers, the highest among all considered technologies, has allowed vison-basedfall detection systems to reach a degree of maturity high enough to start being implemented in commercial applications.

These systems examine human pose, human movement or a mix of both and categorize them as fall in case the established criteria are met. All of them have a common structure of two blocks, a first one that assigns abstract descriptors to input video signals, and a second one that classifies them. In some of the reviewed systems, these two blocks are preceded by another one. Its objective is improving the quality of the incoming signal by reducing noise or adapting its format to the needs of the blocks downstream it.

Almost all reviewed systems work with RGB, near infrared or depth video inputs. Systems working with RGB video signals have evolved from the use of global descriptors to the use of local ones. Global descriptors extract information from the foreground, once it has segmented, and encode it as a whole, while local ones focus on area patches from where relevant features, characteristic of human movement or pose, can be derived. This evolution has made systems more resilient to perspective changes and noise due to illumination and occlusion.

Depth information is also used either solely or complementing RGB images. The systems using it have proofed to be very reliable in high noise conditions due to illumination. However, higher prices and an effectiveness limitation up to distances where depth data can be inferred from stereoscopic vision remain relevant limitations to this technology.

The second block of these systems approaches the classifying problem from two possible perspectives, discriminative or generative. Discriminative models establish boundaries between classes, while generative ones model each class probability distribution.

Although an extensive array of techniques has been used to implement both blocks, the use of ANNs is becoming increasingly popular, as their ability to learn to give them a matchless advantage. This is the case of [233] , a system that uses images that have raised false alarms for retraining. Among all possible ANN architectures, two families have proven to offer good performances in the field of artificial vision, convolutional (CNN) and recurrent ones (RNN). Convolutional networks are able to create feature maps out of images that express what can be seen in them. Recurrent architectures, and specially LSTMs, are able to grasp the dynamics associated with video clips, as the cycles in their structure allow them to remember passed features and link them along time. New architectures fusing layers of both networks, CNNs and LSTMs, being able to identify objects and abstract their movement, show promising results in the area.

After object identification, movement capture is needed, so its dynamics can be abstracted. To do it, object tracking is required. This activity can be done through a number of techniques that can be grouped into two blocks, linear and nonlinear. Due to the nonlinear nature of the movement of the human body during falls, the last block of techniques has proven to be more suitable for this purpose.

A number of datasets are used for system validation and performance determination purposes. However, their fragmentation and the total absence of a common reference framework for system performance evaluation make comparison very difficult. In addition, all datasets are recorded by actors or volunteers clearly younger than the elderly community. The significant differences between simulated and real falls and between falls of elderly and young people are documented by Kangas [311], and Klenk [261], so reasonable doubts on the performances of all reviewed systems in the real world are raised. In any case, the clash between privacy protection and real-world datasets makes it difficult to get good quality data for system training and validation.

No articles mentioning the orientation of system design towards their potential users have been found during this research. The only articles found in the area of fall detection systems regarding this aspect are the ones of Thilo et al. [15], and Demiris et al. [16], where the elderly community needs are described, and recommendations to developers are given. This way, there is evidence of a disconnection between developers and users, which, eventually, leads to low acceptance rates. However, and although acceptance is low, the only technologies commercially used for automatic fall detection are the wearable and video based ones.

The implementation of vision-based fall detection systems has traditionally fallen in the field of ambient systems. However, robots are offering the possibility of making them mobile, and the potential future incorporation of smart glasses or contacts gives the chance to make this system wearable. In these cases, cloud computing may not be an option, so the computational cost will need to be taken into consideration, and low-power consumption will be a key factor.

Finally, in accordance with L. Ren et al. [259], optimal detection performance comes from fusion-based systems that complement vision-based technologies with alternative ones.

2.4 Conclusions

Most taxonomies of fall detection systems classify them into three blocks: wearables, ambient, and vision-based systems. All systems share 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 usually pre-processed to reduce its noise as much as possible, and then it is analyzed to infer movement descriptors. Finally, those descriptors defining movement are classified using a number of techniques to determine whether a fall has taken place or if a specific gait meets the requirements associated with a high fall probability.

Wearable systems use sensors carried by the monitored person to evaluate their movements. The vast majority of system sensors are either accelerometers or gyroscopes, although some other kinds, such as microphones, pressure sensors, ECG, and EMG, are also used. Despite the disadvantages associated with reduced connectivity capabilities and limited edge processing power, these devices have reached a high degree of maturity. The high number of papers related to wearable systems collected during this review proves the interest of the researching community in these technologies, especially in the inertial one.

Fall detection ambient systems are based on contact, passive infrared, acoustic, radar, or Wi-Fi technologies. Their main difference, compared to wearable devices, is sensor position, as while in wearable systems they are carried by the monitored person, in the case of the ambient ones, sensors are placed around him.

Although these systems present an important advantage over wearable ones, as they are not battery-dependent, their development over the last few years has not been as impressive as that experienced by the wearable ones. A good indicator of this weaker research interest is the number of published papers associated with these technologies, which is around half the number of articles associated with wearable or vision-based devices. This lower researching interest makes ambient systems stay in a position of low maturity, and a lot of effort remains to be done to change this situation.

Vision-based systems work with RGB, near-infrared, or depth video inputs. In line with the development of artificial vision technologies, mainly using ANN's, a huge amount of effort has been made over the last few years to develop this kind of devices. This important research effort, proven by the number of published papers, which is the highest one of all considered technologies, has allowed vision-based fall detection systems to reach an important degree of maturity.

In spite of the overall reduced acceptance of these systems, most commercial automatic fall detection systems are based on wearable technologies, especially the inertial one, while a good number of the most recent ones are based on vision-based technologies. This is probably an effect of the degree of technology maturity and researching interest, which, in turn, may be an indication of how suitable a certain technology is to solve the problem of fall determination.

However, as it will be shown in the next chapter, the use of these systems may be perceived as acceptable by the elderly community and their caregivers in certain situations, provided they are really adapted to the user's needs. These situations are mostly associated with times when human supervision is low or nonexistent. One of the most common scenarios of this situation is the sleep time during the night. In this period, elderly people tend to wake up feeling the need to go to the toilet. Disorientation and low illumination are common in this

situation, and therefore, the chances of fall are higher than under other circumstances. Additionally, supervision absence could substantially delay any needed medical response.

The above-mentioned circumstances are very clearly identified in the next chapter as a likely fall scenario where an automatic fall detection system would be well accepted. However, commercial systems based on wearable technologies would not be an answer to this problem, as the elderly community tends to remove body sensors when they go to bed to create comfortable sleep environments. Ambient technologies have proven to be immature, and visual-based systems' performances would be severely degraded by low light conditions, as they work in the visual or near-infrared spectrum.

In this situation, developing a visual-based fall detection system working in the far-infrared spectrum might be a good solution. Additionally, and to get the best possible acceptance, the other two main problems associated with today's vision-based fall detection systems would be solved: insufficient amount of human falls real-world data and privacy protection.

3 User's needs

Prior to system definition an extensive research to identify user's needs was conducted in an attempt to establish a proper connection with the different groups related to the world of elderly care. User's needs were obtained following the recommendations established by 29148-2018 - ISO/IEC/IEEE [312] and 830-1998 – IEEE [313].

During the problem analysis and solution characterization phase six main groups, directly or indirectly involved in the elderly care tasks, were identified:

- Elderly.
- Friends and family.
- Home-care givers.
- Nursing home managers.
- Nursing home care givers.
- Paramedics and emergency medical personnel.

The following process aims to determine user's needs and requirements.

3.1 Methods

Between January and May 2021, 36 qualitative, open-ended, semi-structured interviews were conducted. Additionally, as a secondary source of information, 146 forms containing qualitative open and close-ended questions were distributed and answered by individuals belonging to all identified stakeholder groups.

Before starting the sampling process, initial contacts were established with five nursing homes, and the research project was presented to their managers. Through these institutions, further contacts with individuals belonging to all stakeholder groups could be established. The sampling process was purposive to obtain a diverse selection of individuals representing a range of age, dependency, and gender for the elderly community. It also covered a range of closeness in relation, age, and gender for the family and friends group, and a range of professional experience, hierarchical position, and age for the rest of the groups.

The inclusion criterion for individuals belonging to the elderly group was presenting a dependency. For the family and friends group, it implied a relation with a dependent person, while for the rest of the groups, the criterion was having a professional experience of at least six months in contact with dependent individuals, except for the emergency medical personnel, whose inclusion requirement was just having a professional experience of at least six months. This is because there are no subgroups in this professional community specifically devoted to elderly attention. Besides, friends and family often play the role of home-care givers when dependents do not live in nursing homes. When that was the case, for the purpose of this research, these individuals were regarded as part of the home-care givers group, instead of the friends and family group.

The semi-structured interviews started with an overall presentation of the automatic human fall detection systems and focused on four thematic aspects:

(1) Degree of confidence in these systems.

(2) User's needs and requirements

(3) Privacy protection

(4) Usage environment.

A guideline based on the state-of-the-art revision was used to conduct the interviews. This guideline was revised after each interview, as interviewees were given the opportunity to add new questions. However, no new questions were proposed after the ninth interview.

Given the pandemic situation due to COVID-19 at that time, all interviews were conducted through Video Tele Conference. The length of the interviews ranged from 12 to 43 minutes with an average of 21 minutes. Audios of the interviews were transcribed and anonymized. Prior to analysis, transcripts were read at least twice to guarantee familiarity with the data. Transcripts inspection followed the principles of qualitative content analysis [314] and [315], a research method for the subjective interpretation of text data through the systematic classification process of coding and identifying patterns. These patterns lead to the relevant concepts for the system user in each thematic area.

The forms included a mix of close and open-ended questions based on the guideline developed for the interviews. The questions were preceded by an introductory video presenting the world of automatic human fall detection systems. The analysis process followed to code the information contained in the open-ended questions followed the same identical principles of qualitative content analysis used for the interviews.

3.2 Results

The identified relevant concepts behind each thematic aspect are the following ones.

3.2.1 Degree of confidence

The elderly community shows a moderate degree of confidence in these systems, accepting them when human supervision is not an option, as long as their caregivers find their use acceptable. The use of these systems is better perceived when human relations cannot be established (e.g., at nighttime during the sleeping period).

For the groups of family and friends and home-caregivers, although human supervision is generally preferred, the use of these systems is accepted in some cases, especially in the case of home-caregivers, when it implies a reduction in care burden or, in the case of family and friends, who often pay, at least partially, for the care service, when it implies a cost reduction. The home-caregiver group believes that, overall, direct human supervision provides reassurance to the dependent, an added value element not supplied by the automatic systems.

The nursing home managers prefer human supervision in general, though they accept the system when it can imply a cost reduction for the care service as long as its performance is reasonable.

Finally, the group of paramedics and emergency medical personnel consider that the degree of confidence in these systems should be based on performance and reaction time, which they regard as a key factor, as often survival after a fall depends on immediate medical attention.

All groups have low knowledge of these systems and their performances.

3.2.2 User's needs and requirements

Reliability is the key requirement of a detection system for all groups. This feature is especially relevant for the friends and family group, which requires exceptionally good performance under all circumstances.

For the caregivers' groups, these systems should be easy to use and designed to complement human supervision. In the case of home-caregivers, they should also reduce the care burden. According to the opinion of nursing home caregivers, the systems should be fine-tuned to the needs of semi-supervised dependents.

Nursing home managers require systems that are cheap, easy to install/reinstall, and easy to deploy.

Paramedics and emergency medical personnel consider reduced system reaction time a critical requirement. While reliability is assessed as essential, false alarms should be reduced as much as possible.

3.2.3 Privacy protection

Privacy protection is paramount for the elderly group, and although they confess unfamiliarity with the technicalities of data breaches and cybersecurity, they also express concern regarding the unlawful use of images captured by the system.

The friends and family group share this concern and express certain mistrust in the cybersecurity measures that could be taken to protect the system information.

The groups of caregivers show concern regarding data breaches, especially in the case of images, although they think cybersecurity measures could mitigate this risk. They believe the threat could not only come from cyberattacks but also from people with physical access to the system, as they could unlawfully retrieve data. Finally, for image-based systems, they prefer the use of infrared or very low-resolution images.

The group of managers is worried about the impact of cyberattacks and system hacking on the institution. They express their concern about legal liabilities and economic repercussions of data breaches. In addition, they perceive non-imaged based systems as less vulnerable.

Finally, the emergency personnel consider privacy protection a relevant issue, especially in the case of image-based systems.

3.2.4 Usage environment

Both the elderly and the friends and family group believe fall detection systems should be adapted to all types of environments.

Home-caregivers think the system should be optimized for environments where it could help diminish the care burden, while the nursing home caregivers believe it should be optimized for environments where humans do not monitor semi-supervised dependents. Both groups assume those contexts are linked to night environments, especially during bedtime. A common scenario for that context is the one associated with dependents getting up after waking up feeling the urgency to go to the toilet.

Managers believe system environmental optimization should be limited to the most likely operational environments to reduce development costs. The emergency personnel express a similar approach.

3.3 Discussion

The degree of confidence in the automatic fall detection systems varies across groups, depending on their familiarity with the system. However, in broad terms, these systems and their performances are relatively unknown, which makes it very difficult to have high degrees of confidence in them. Nevertheless, the systems could be accepted when their use implies

cost or care burden reductions. However, direct human supervision remains the preferred option for all groups because human contact is always a reassuring factor.

System reliability is the key requirement for all groups. In the case of the groups that could potentially operate the system, the caregivers, further requirements are issued to make the system friendly, easy to use, and adapted to the needs of semi-supervised dependents, who could be the main beneficiaries of the system. Furthermore, the system is also required to diminish the care burden as much as possible, and additional requests are made to make it low-cost, easy to install/reinstall, and easy to deploy. Additionally, emergency personnel request system reaction times as low as possible to minimize the time from the event happening to medical intervention, as this time is often related to the survival rate.

Privacy protection is a very relevant issue, and all groups express concern for data breaches. This concern is especially acute in the case of image-based systems. Most groups consider cybersecurity measures a reasonable risk mitigation factor, although they should be accompanied by other measures, as data could also be unlawfully retrieved by personnel with physical access to the system. Managers show concern for legal liabilities and economic repercussions of data breaches. Finally, a preference for the use of infrared or very low-resolution images is expressed, in the case of vision-based systems, as these types of data preserve privacy.

The environments where these systems are more likely to be deployed, according to the opinions of the caregiver groups, and in line with the main ideas expressed in the first paragraph of the discussion section, are those where direct human supervision is not a clear option. These contexts are linked to patients who are not continuously supervised, as their degree of dependency is not extreme yet, and to periods when supervision is not scheduled. The common situation described by several interviewees is the one associated with semi-supervised patients who get up at nighttime to go to the toilet. This is an especially delicate situation, as several hours can go by prior to the next scheduled human contact with the caregivers at wake-up time.

3.4 Conclusions

Automatic fall detection systems are relatively unknown among dependents and their caregivers. This, combined with the value provided by human contact, makes direct human supervision the preferred option.

However, there are certain situations, usually related to patients whose degree of dependency is not extreme, where the use of these systems could be well-accepted by patients, their families, and caregivers. These situations are linked to times when either the patients do not have human supervision or it is infrequent. Under these circumstances, the use of fall detection systems could be a useful aid.

A common scenario described by numerous interviewees is related to semi-supervised patients getting up at nighttime. In this situation, the person is often disoriented, increasing the likelihood of a fall. Additionally, in the event of a fall, it may go unnoticed until the next day, substantially delaying potentially needed medical intervention.

Under these circumstances, a fall detection image-based system could satisfy this identified need better than other options. Patients in bed during nighttime dress light outfits and get rid of any accessories, disqualifying the use of wearable systems. Ambient systems might seem like the optimal choice under these circumstances, but the low maturity state of these technologies currently discourages the use of such systems. Vision-based systems combine good technological maturity and optimal operation 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 nighttime.

4 Human Pose Estimation from Far Infrared Images

4.1 Dataset

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 engaging in different activities. The dataset contains far infrared images (FIR) and 3-dimensional and 2-dimensional annotations associated with their joint positions.

The potential uses of this dataset include training for systems in the fields of 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.

4.1.1 Related work

Although there are not major, public datasets containing FIR video clips and their associated annotations some works, like [316], create small FIR imagery datasets of around 700 images with manual 2-dimensional annotations in order to verify the performance of certain pose estimation networks.

However, in the field of RGB datasets the situation is different, as four major datasets; FLIC [317], LSP [318], MPII Human Pose [319] and COCO [320] are systematically used to train and validate pose estimation architectures. Additionally, Human 3.6M [321], MPI-INF-D-HP [322], NTU RGB+D [227] and Deepcap [323] have also been used with this purpose.

LSP (Leeds Sports Pose) dataset contains 12000 images from sport activities. Individuals in the imagery are practicing sports, which make specially challenging their pose determination. The annotations associated to the images label 14 joints of the human body and this dataset has been used for single person pose estimation models.

FLIC (Frames Labeled in Cinema) dataset contains 5000 images. It collects images from 30 very popular Hollywood movies by taking the first frame of every block of 10. The annotations reflect the position of 10 body joints and this collection of images has been used to train and evaluate both single and multi-person systems.

MPIII (Max Planck Institute of Informatics) Human Pose dataset collects the annotations of 40000 people in 25000 images. These annotations determine the position of 15 joints of these individuals and, as FLIC, it has been used to train and validate single and multi-person models. This dataset is one of the main benchmarks for evaluation of articulated human pose estimation models and it is widely accepted, together with COCO, as the main standard for system comparison.

COCO (Common Objects in Context) is an image database issued by Microsoft. It is a large-scale object detection, segmentation, and captioning dataset. The annotations are contained in a JSON file, which collects information of images of 80 types of elements. One of the elements is people and, for the images containing persons, 17 joints are labeled. The dataset contains annotated images of 250000 people in 200000 images and since 2014, when it was first presented, it is yearly revised and improved.

4.1.2 Data modalities

A FIR camera is used to record our dataset, which is synchronized with a 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 of each joint for 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 x 640 pixels. The joint information consists of 3-dimensional positions of 19 major body joints, defined with an error of less than 5 millimeters for each of the recorded frames. Additionally, the 2-dimensional projection of those coordinates onto the recording plane is also provided.

4.1.3 Action classes

The dataset contains 27 action classes in total. 26 of them are daily life activities while the other one includes different types of falls. The different actions are repeated by the volunteers in four different positions so frontal, rear and side views of the same actions are 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. These actions include:

1. Giving directions.
2. Discussing.
3. Eating.
4. Taking photos.
5. Exercising on the ground.
6. Running in place.
7. Walking.
8. Sitting and standing up.
9. Coughing.
10. Exercising.
11. Playing basketball.
12. Picking up objects.
13. Limping.

The second block includes a single person who executes a different set of actions with validation purposes. These activities include:

1. Brushing teeth.
2. Encouraging your team.
3. Toasting.
4. Taking a selfie.
5. Crouching for meditation.
6. Walking a dog.
7. Throwing a stone.
8. Talking on the phone.

9. Stretching yourself.
10. Hopping.
11. Kicking a ball.
12. Tying shoelaces.
13. Rotating your trunk.

Finally, the third block includes four volunteers who are recorded from different perspectives falling forward, falling backwards and side falling. The falls start from static or dynamic situations and a number of them are slow falls, a common type of fall in the elderly community.

A few examples of annotated images belonging to video clips of volunteers performing different activities and fallings can be seen in Figure 5.

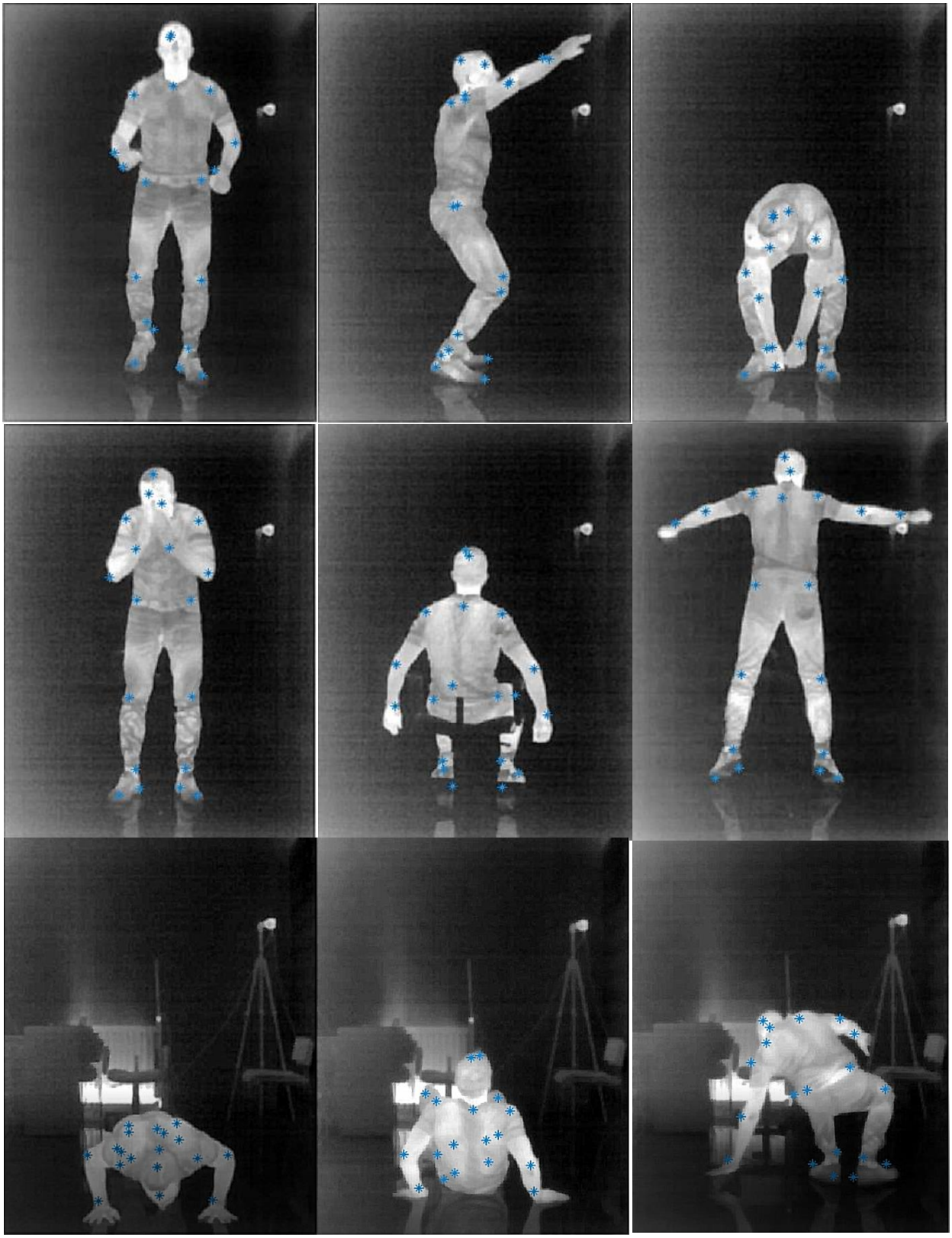


Figure 5. Volunteer (a) running, (b) playing basketball, (c) picking up an object, (d) coughing, (e) sitting, (f) exercising, (g) falling forward, (h) falling backwards, (i) side falling.

4.2 2D human pose estimation networks

4.2.1 State-of-the-art

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

This process involves image pose determination 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 [324], the generative and the discriminative ones.

The discriminative approach requires an initial extraction of image characterization elements. These elements, called features, are then abstracted, often in a statistical way, in a subsequent process step called feature description. Then, abstracted features are assembled together using a human body model for such a purpose, and finally, this assembly is fed to a classification or regression model in order to estimate body pose.

On the other hand, the alternative approach, the generative one, includes top-down and bottom-up methods. The former integrates image elements to compose descriptive features, which are sometimes directly used to estimate a human pose and other times are employed to localize body parts. These localized body parts are then used to compose a human structure, adopting a specific pose. The top-down methods follow an opposite flow, as they use high-level semantic descriptors to guide low-level element recognition.

4.2.1.1 Feature extraction and characterization

Depending on its level of abstraction, image features can be classified as low, mid or high-level abstraction features.

The low-level ones are associated to shape or appearance of the human body as a whole or any of its parts. Low-level image features traditionally include silhouette, as in the systems described in [325]-[327], contours, as in [328], [329], and edges [330], [331]. This way, all items define limits but while silhouettes define human body limits, contours outline body parts and edges mark lines of sudden variation in an image.

The specific processes followed to extract these lines vary from system to system but, in overall terms, contours are obtained by removing image background, so foreground silhouette lines defining body limits can be inferred. Contours can be extracted from silhouettes, once they have gone through a body segmentation process and edges are directly determined once a filtering process based on differential kernels has been applied.

Mid-level image features try to capture the composition and distribution of image elements. These features include Fourier descriptors as in [332], Poisson features [333] and shape context [334]-[337]. The last features, which are the most common ones, capture every relevant body element position related to the position of the rest, usually by using polar representations.

Additional local mid-level features include gradient edge encoding, as the histogram of oriented gradients (HOG) [338], [339], scale Invariant Feature Transform (SIFT) [340], [341], shapelet features encoding [342] and edgelet feature encoding [343]. Among them, the two first methods were widely applied to extract features in computer vision tasks before the use of artificial neural networks became the norm for automatic image feature extraction.

Global mid-level features have also been used in a number of systems. This way, the system described in [344] uses foreground maps while systems like the ones described in [345], [346]

use grid features to encode HOG and SIFT features respectively, outperforming standard HOG and SIFT feature encoding.

Finally, multilevel hierarchical encoding is used in a number of systems [347]-[350] proving to be more resilient to geometric transformation than previous methods.

High-level features encode not only information related to body part identification but also data referred to location and orientation [351]-[353] or to spatiotemporal correlations [354]. Unlike mid-level features, which encode data following rigid predefined patterns, high-level ones adapt patterns to every situation, obtaining, this way, a better adaptation.

Motion features are able to capture the spatial correlations of objects along time, as they move in relation to the camera. Optical flow and dense optical flow [355]-[357] encode the pattern established by an object, its silhouette, contours and edges, as it moves along time. Optical flow gradient contains relevant information regarding object movement and, therefore, it can be successfully used to track it, including human poses [358], [359]. Alternative techniques able to determine local motion similarities, such as motionlet [360] or motion and appearance patches [361], which can determine image difference, are also used for the same purpose.

4.2.1.2 Human body models

Once features at any level have been extracted from images, discriminative methods assemble them together using a human body model. These models can be kinematic, planar, or volumetric, depending on their characteristics.

Kinematic models adopt a skeletal structure with segments linking joints. This way, the information associated with joint positions and the one that determines body part orientation configure a set that fully defines human pose. Kinematic models can be divided into two main groups. The first one, called predefined, associates features to a model whose body parts' dimensions are fixed. The second group includes all models associated with learned graph structures, and among these structures, the most popular ones are called pictorial structure models (PSM). These tree-structured models have been successfully applied to human pose estimation tasks, both in two and three dimensions, in systems like the ones described in [362]-[365]. They model the human body as a collection of elements gathered in a deformable configuration. Each of the body parts is represented as a simple element that is linked by spring-like connections to the adjacent elements. This way, human body is treated as an assembly of parts connected at the height of the joints by deformable links. The approach, which requires low computational power, is very successful at estimating human pose as long as all body parts are clearly visible. However, when occlusion problems appear this approach fails to solve successfully the problem.

Improvements of PSM models have been tried by including relations between non-connected body parts [366], by adding multiple tree models [367], [368] or by using Bayesian networks [369], [370].

Planar models add to the information associated to joints and body part orientation its appearance. The two main types of planar models are Active Shape Models (ASM) and cardboard models. The former ones present a basic human silhouette deformed accordingly to the adopted pose. To do it systems like the ones describe in [371]-[373] use principal component analysis techniques. The cardboard models capture information associated to the color of body part image patches. This way, systems based on this model [374] code information associated to the color histogram and average color of every body part image patch.

Volumetric models represent the human body as a three-dimensional structure, either by using an array of simple geometric forms such as cylinders or cones, or by adapting a mesh to the human body surface. In the first case, the different body parts are represented by simple elements, while joints connecting those parts are given a number of degrees of freedom from 1 to 3. This way, once joint positions are estimated on the two-dimensional image, the three-dimensional pose can be inferred by solving the surface projection problem using least-square methods as in [375].

The alternative method to model a three-dimensional human body is adapting a mesh to its surface. Meshes are deformable triangular surfaces that define the limiting surface of a body. The most popular method to adapt a mesh to a human body is known as Shape Completion and Animation of People (SCAPE) and is extensively used in different systems [376]-[379]. Enhanced methods for mesh adaptation that take into account shadows in images with a single light source are tried in [380], [381].

All models are subject to constraints associated to joints limitations and behavioral patterns of motion. This way, the information contained in an image, once it has been adapted to a human body model subject to constraints, is enough for pose estimation tasks [382]. Two main techniques have been used to infer movement constraints from collected motion data. The first one, based on joints limits, is used in [383] while the second one, a physics-based model which accounts for dynamics effects in joints and on the ground, is the base to design the systems described in [384], [385].

4.2.1.3 Generative methods

Generative human pose estimation methods imply a geometric projection of a volumetric human body model over the image plane so it matches 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. This way, the process, which tries to find an absolute minimum, requires a complex search over the model state space in order to find it. A good number of systems have been developed following these methods [386]-[389] and, as expected, they are susceptible to local minima errors, requiring good initial pose estimations in order to avoid it. The most common methodologies to obtain the searched minimum are local optimization [390]-[392] and stochastic search [393], [394].

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

4.2.1.4 Discriminative methods

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 the following ones.

Support Vector Machines

Support Vector Machines are discriminative algorithms used for classification or regression able to determine hyperplanes used for class discrimination. A good number of systems use these algorithms to estimate human pose [395]-[397].

Relevance Vector Machines

Systems using Relevance Vector Machines [398], [399] provide probabilistic regression of human pose using Bayesian inference. Relevance Vector Machines have a functional form equivalent to a Gaussian model although it incorporates a covariance function.

Mixture of Experts

These algorithms mix together Bayesian Mixture of Experts algorithms and some others in order to enhance system performance. Systems based on them [400], [401] require higher amounts of data than previous ones but offer better results.

Manifold learning methods

These methods try to generalize linear frameworks like the Linear Discriminant Analysis or the Principal Component Analysis so they become sensitive to nonlinear data structures. This way, they try to reduce the number of dimensions of nonlinear high order data structures by projecting them onto lower order bases, usually following non-supervised methods. Several systems, as the ones described in [402]-[406], use these methods to determine human pose.

Pose embedding methods

Embedding methods are learning algorithms able to identify images of humans in poses similar to a given one following a direct method of image comparison. This way, pose determination systems using embedding methods [407], [408], like the ones based on manifold ones, are able to diminish the number of system dimensions and determine the most similar pose of a dataset to the one observed in a specific picture.

Locality-constrained Linear Coding

Locality-constrained Linear Coding applies constraints at local level to select descriptors of the image from a codebook and is able to determine the linear combination of those descriptors that best define the image. This way, systems like [409], [410] can estimate human pose.

Bag-of-words based methods

Bag-of-words was the most used computer vision algorithm before the deep learning neural networks became the dominant pipeline. As its name shows, systems based on these methods [411] isolate the most relevant image features creating, this way, a vocabulary or a set of words. This way, a histogram reflecting word occurrence in the image is built and it is used as its final representation. Finally, this representation is fed to either a classifier, in order to compare the pose in the image with a set of pre-defined human poses, or to a regression model, in order to determine the pose in the image.

Random forest

Random forest methods aggregate decision trees to execute regression or classification tasks. Decision trees are built using leaves, which represent class labels, and branches, which are associated to features of the classes. This way, by following an iterative process decision trees are able to perform its assigned task. The computational burden of trees is low, as the split points are selected, so the cost function that measures processing power is minimized. However, trees tend to overfit. To reduce this tendency the set of training data is split into a number of groups and one tree is trained using each group, then the resulting trees are grouped together obtaining this way a random forest [412]-[414] used with regression or classification purposes.

Certain techniques [415] have been used to improve random forest performances by using two-layered random forests. This way, the first layer is used as the classifying part and the second one regresses joints positions. Alternative enhancing techniques have been used in [416], where the authors use a Hough Forest, which is a random forest adapted to execute a generalized Hough transform. Hough transform is a technique used to identify simple figures generalized to recognize more complex features present in an image.

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 ANN architectures [417]-[419]. The second approach [420]-[422], 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 a number of advantages over the top-down option, as it is able to better overcome the early commitment problems associated to a faulty detection of individuals in the image. Furthermore, although the computational cost of the top-down approaches is lower than the bottom-up one when the number of individuals is low, when it grows, the top-down philosophy cost becomes higher.

4.2.2 Materials and methods

The area of interest of this work is centered on fall detections systems used in real life. In the user's need chapter, the communities of users manifest that they consider the use of these

systems only in situations where human supervision is not possible. This way, fall detection systems used in reality will process images of a single person and because of it, due to the advantages presented by the top-down approaches under these circumstances, the review of ANN architectures will be restricted to this group.

Although the number of 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 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 [423] 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 [424] 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 [425].

4.2.2.1 Convolutional architectures

The most representative convolutional architectures have been trained with images from the FIR-Human dataset in order to evaluate their performances when working with this kind of images.

DeepPose

DeepPose [423] is a convolutional architecture whose backbone is Alexnet [426]. The network uses a cascade of regresssor to refine joint position determination as shown in figure 6. 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.

Figure 6. DeepPose structure. [426]

ConvNet POSE

ConvNet POSE is a convolutional architecture presented in [427]. It represents the first approach to heat-map generation leaving the previous regression philosophy used by DeepPose.

The architecture integrates three modules as shown in figure 7. 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.

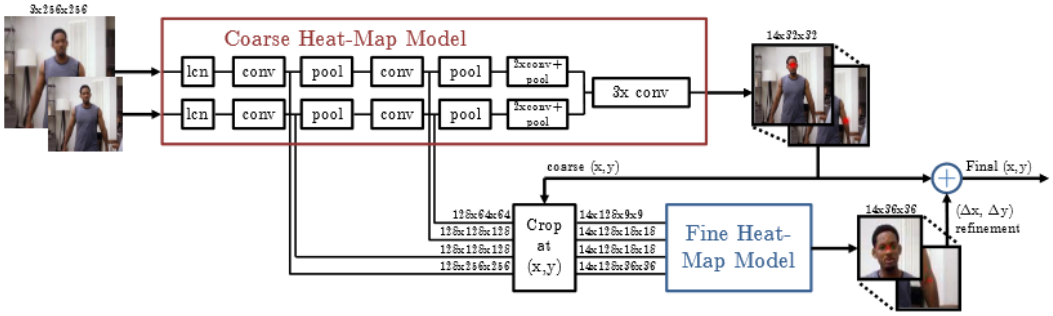


Figure 7. ConvNet structure. [427]

Convolutional pose machines

Convolutional pose machines [428] is an architecture able to produce an array of 2D heat-maps that represent the space probability distribution for the location of each key-point. The architecture is multi-stage and end-to-end trainable as shown in figure 8. This way, at the first stage the input data is the original image that, after being processed by a standard Visual Geometry Group structure, produces heat-maps for every joint. Subsequent stages use 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 in order to learn spatial relationships that, together with the combined input of the original image and the heat-maps generated by the previous stage, improve the accuracy of the network output.

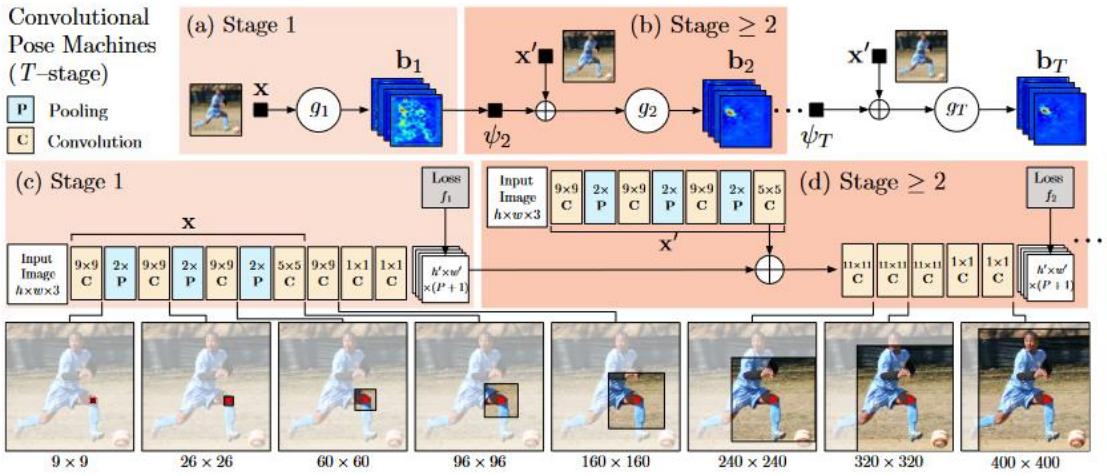


Figure 8. Convolutional Pose Machines structure. [428]

Stacked hourglass networks for human pose estimation

The stacked hourglass architecture [429] takes its name from its look, which resembles an hourglass as shown in figure 9. 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.

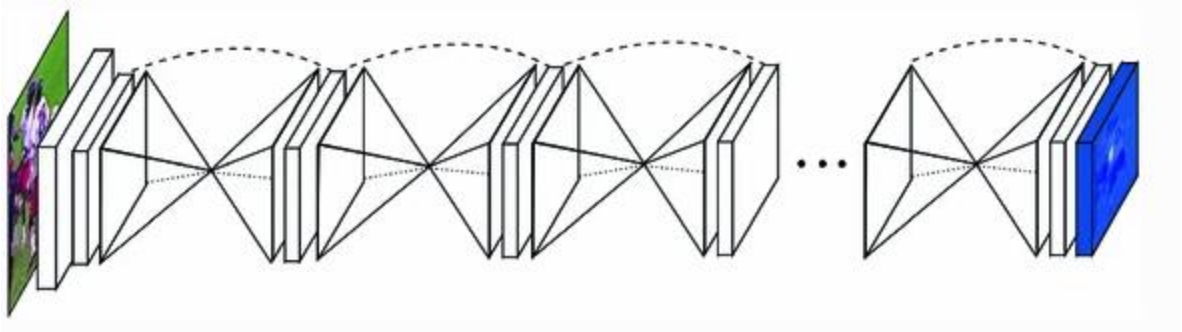


Figure 9. Stacked hourglass structure. [429]

Human pose estimation with iterative error feedback

Iterative error feedback [430] 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. This architecture is shown in figure 10.

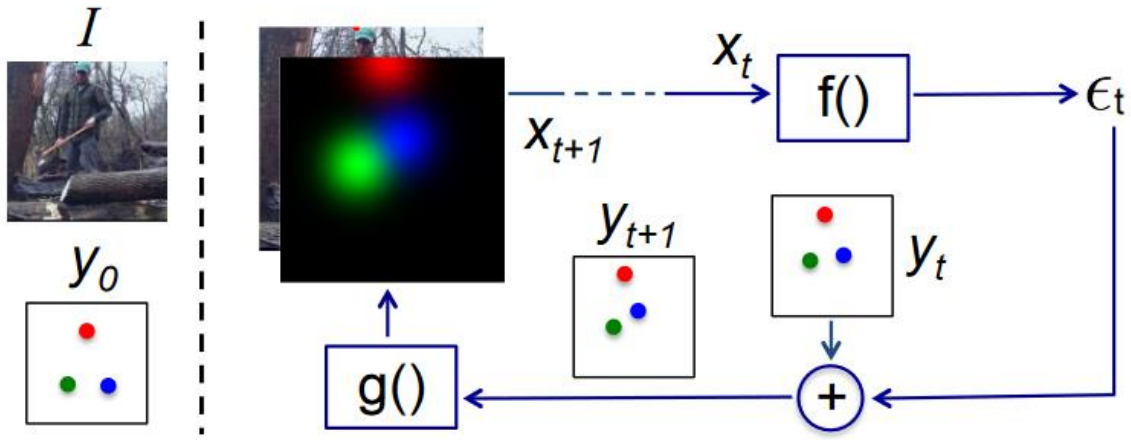


Figure 10. Iterative error feedback structure. [430]

Cascade feature aggregation for human pose estimation

This architecture [419], shown in figure 11, 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.

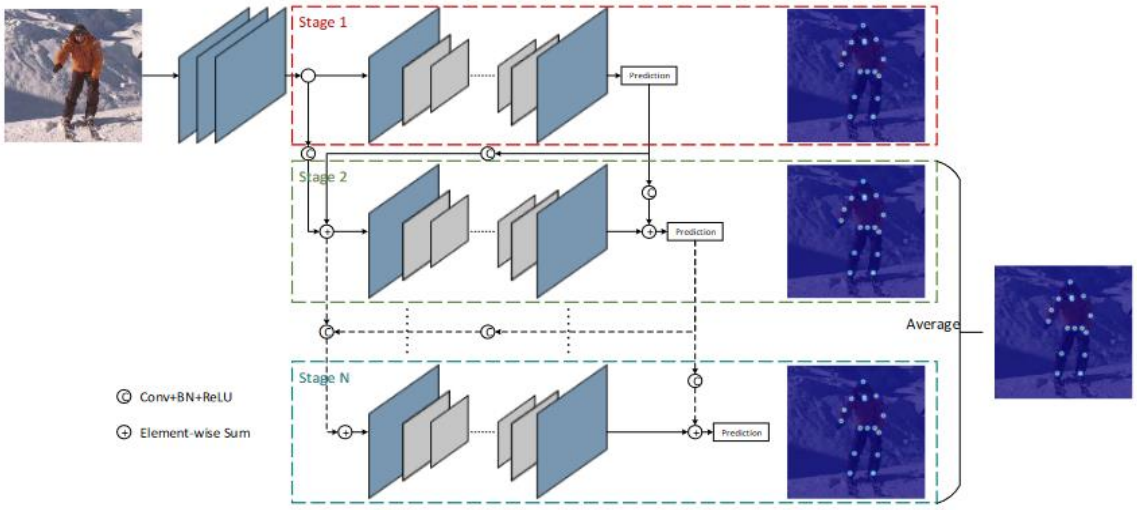


Figure 11. Cascade feature aggregation structure. [419]

4.2.2.2 *Transformers architectures*

TFPose

TFPose [431] is an architecture shown in figure 12 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 [432] 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.

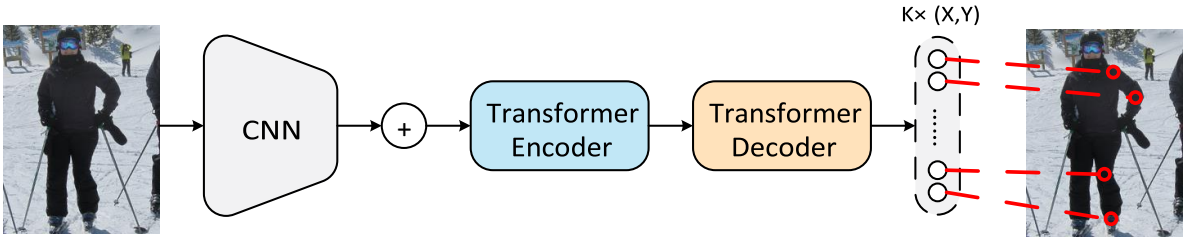


Figure 12. TFPose structure. [431]

ViTPose

This architecture [433], shown in figure 13, 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.

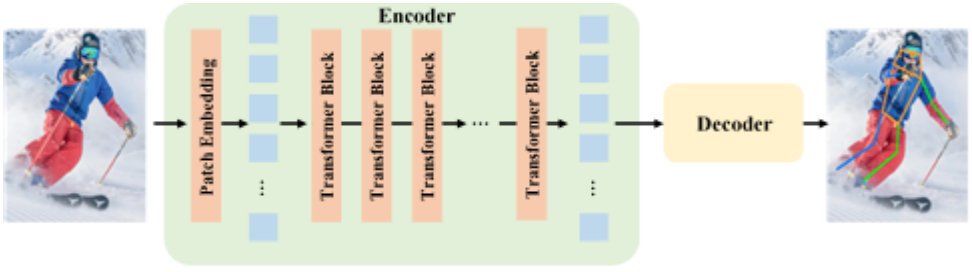


Figure 13. ViTPose structure. [433]

4.2.2.3 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.

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

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.

4.2.2.4 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 [423].
- 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 [434]. PCKh is a variation of PCK, whose reference distance is defined as 50% of the ground-truth head segment length [319]. 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.

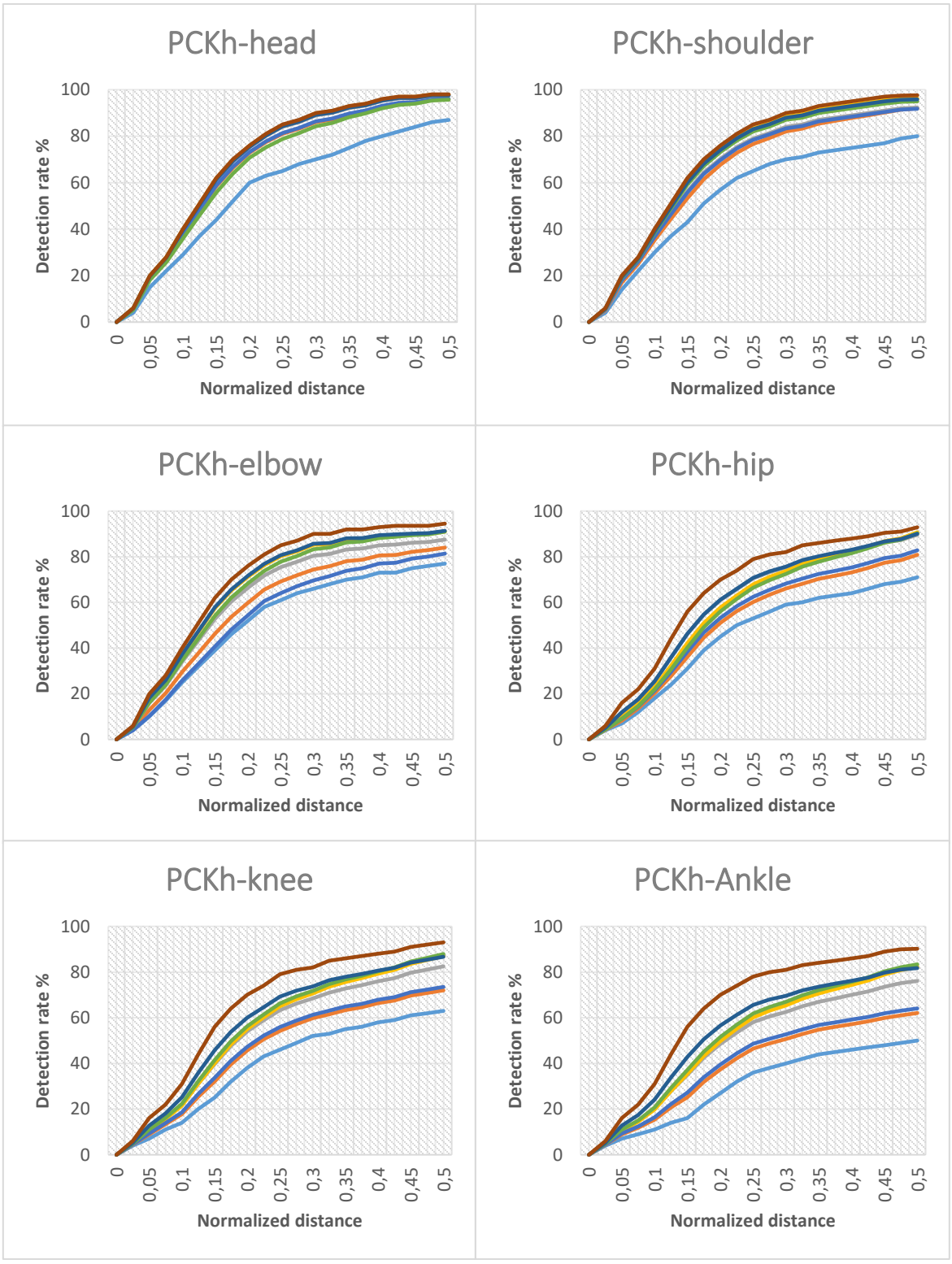
Due to the generalized use of PCKh, a common metric used in all the papers that present and evaluate the architectures described in the previous paragraphs, PCKh will be used in this work as the common metric for comparison.

4.2.3 Results and discussion

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 networks 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 [435].

The block one of the FIR-Human dataset was used for network training while the block two was employed for validation and network comparison. To enrich both blocks as much as possible the data augmentation strategy proposed in [436] was adopted. It includes random rotations (45º, -45º), random scaling (0.65, 1.35), flipping and half body data augmentation. This way, the total number of images was multiplied by four.

The results of the system performance comparison are summarized in Figure 14.


PCKh-head
PCKh-shoulder
PCKh-elbow
PCKh-hip
PCKh-knee
PCKh-Ankle
Detection rate %
Normalized distance
0
20
40
60
80
100
0
0,05
0,1
0,15
0,2
0,25
0,3
0,35
0,4
0,45
0,5

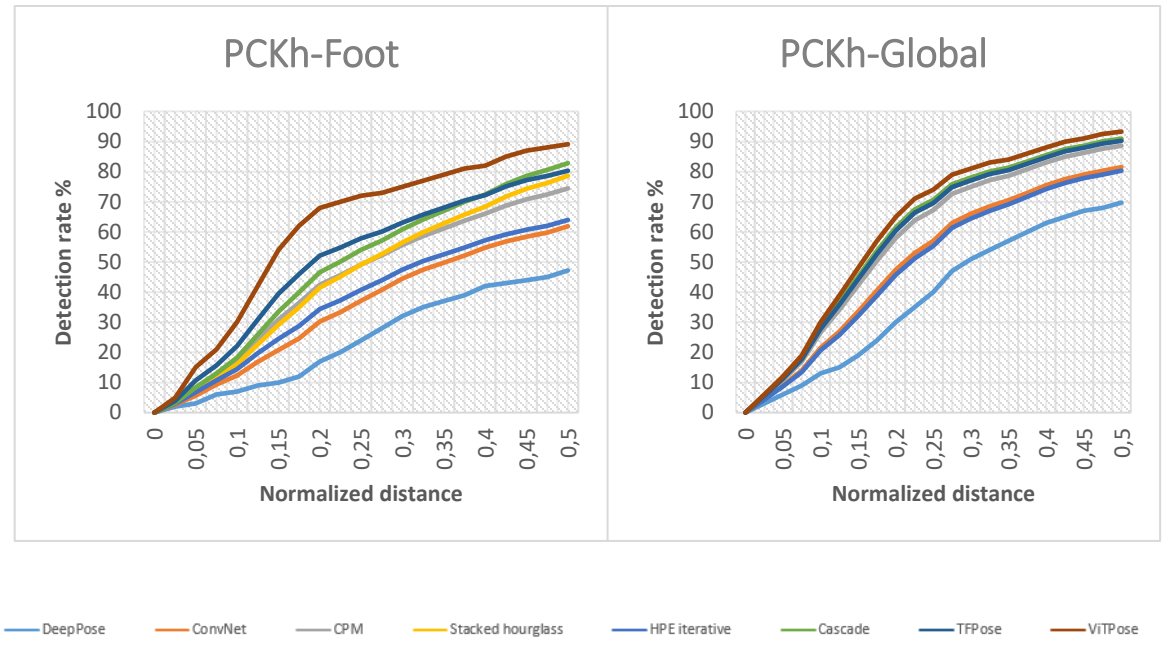


Figure 14. System performance comparison.

Table 11 collects the performances of the different networks as a function of joint.

Table 11. PCK@0.5 for the different human body joints.

MODEL	PCKh@0.5								
	Head	Shoul.	Elb.	Wrist	Hip	Knee	Ank.	Foot	Total
DeepPose (ResNet - 101) [423]	87.9	79.3	76.8	75.2	71.1	70.7	49.8	47.2	69.7
ConvNet Pose [427]	96.7	91.2	83.7	78.0	81.0	80.6	64.6	61.5	79.7
CPM [428]	98.6	91.3	86.8	87.2	89.1	88.7	78.1	74.0	86.7
Stacked hourglass (3 Stages) [429]	98.8	95.6	91.0	87.3	90.2	89.8	83.4	78.6	89.3
HPE IF [430]	96.3	90.9	81.3	72.6	82.8	82.4	66.2	63.9	79.6
Cascade (ResNet-101 Cascaded with 2 ResNet-50) [419]	96.5	94.7	90.8	87.1	89.9	89.5	83.7	82.9	89.4
TFPose (Resnet -50; Nd=6) [431]	98.6	95.2	90.8	86.2	89.9	89.5	82.4	80.4	89.1
ViTPose (ViTAE-G) [433]	98.6	96.9	94.3	92.1	93.0	92.6	90.0	89.1	93.3

Table 12 presents the computational cost required by the different models.

Table 12. Computational cost.

MODEL	Output	Input image size	Flops (GFLOPs)
DeepPose (ResNet - 101) [423]	Regression	(3, 192, 256)	7.69
ConvNet Pose [427]	Heat-map	(3, 192, 256)	28.56
CPM [428]	Heat-map	(3, 288, 384)	143.57
Stacked hourglass (3 Stages) [429]	Heat-map	(3, 384, 384)	64.5
HPE IF [430]	Heat-map	(3, 192, 256)	36.58
Cascade (ResNet-101 Cascaded with 2 ResNet-50) [419]	Heat-map	(3, 288, 384)	61.3
TFPose (Resnet -50; Nd=6) [431]	Regression	(3, 288, 384)	20.4
ViTPose (ViTAE-G) [433]	Heat-map	(3, 432, 576)	76.59

Figure 15 illustrates the ground truth heat-maps of a FIR image and the predictions made by the different systems.

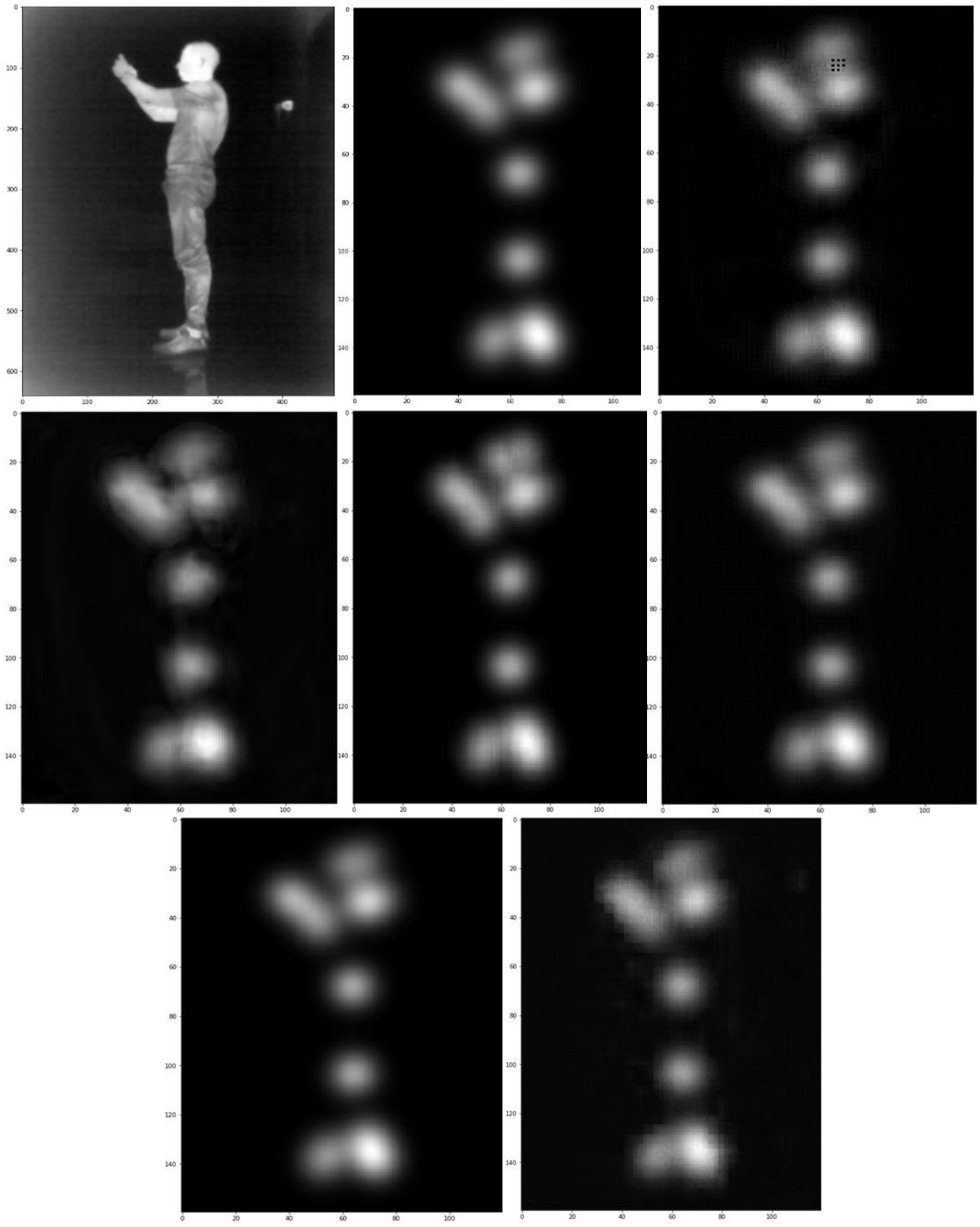


Figure 15. (a) Base image, (b) Ground truth heat-map, (c) ConvNet Pose prediction, (d) CPM prediction, (e) Stacked hourglass prediction, (f) HPE IF prediction, (g) Cascade prediction, (h) ViTPose prediction.

As expected, and in line with the results obtained in the different papers that present each considered system, there is a significant difference 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 performance than the ones based on the classic use of CNNs.

All models demonstrate exceptional performance at identifying the head, as can be easily inferred from Figure 12, where the obtained PCKh is quite similar for most of them. A similar result is observed in the case of the shoulders, the joint closest to the head. However, as the joints get farther from the head, the model's ability to determine joint positions degrades significantly, especially in the case of the ankle and the foot. Additionally, the performances of the different systems, which are similar for the least challenging key-points, vary widely for the most challenging ones.

Finally, 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 exceptions like CPM, better image input resolutions lead to better outcomes, especially for the most challenging joints, albeit at a higher computational cost.

4.2.4 Conclusions

A wide set of models, representing most of the state-of-the-art human pose recognition networks, have been trained using the FIR-Pose dataset.

The results obtained using the FIR-Pose dataset are very much in line with those obtained using RGB datasets. These results show that architectures based on transformers outperform convolutional-based ones, as the performances of the former exceed the results of the latter for the same input image resolution.

In general terms, 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, it is observed that higher input resolutions allow models to yield better performances but come at the cost of higher computational requirements.

5 Dynamic Descriptors for Fall Characterization

As already presented in [12], although vison-basedfall detection system's performances are very satisfactory, the significant differences between simulated and real falls, and between falls of elderly and young people, documented in [13], [14], as well as the difficulty to access real-world data as a consequence of privacy protection, yield reasonable doubts about the performances of these systems operating in real circumstances.

These doubts are a direct consequence of the use of kinematic descriptors [12] 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, the obtained descriptors could be inaccurate or incorrect and the system performances in the real-world could be poorer than expected.

To solve this problem, two alternative approaches could be adopted. The first one is based on the correction of the modeling errors associated to a training not based on accurate information. This approach, considered in [437] for the field of data-driven fault diagnosis and in [438], [439] for the field of automatic control, implies access to real data after initial system training in order to correct the modeling errors caused by inexact training information. Unfortunately, the absence of any database containing real-world data makes this approach impossible in the field of automatic fall detection.

The second alternative approach considers the use of dynamic descriptor instead of the classical kinematic ones. These descriptors approach the human body in terms of balance and stability, this way, differences between real and simulated falls become irrelevant, as all falls are a direct result of a failure in the continuous effort of the body to keep balance, regardless of other considerations.

This work implements this alternative approach in the field of automatic fall detection systems for the very first time. To do it, an ANN able to regress, from a sequence of 2-dymensional (2D) poses, the projected position onto the ground plane of both feet and body center of mass (COM) is proposed. This ANN also determines the feet contact status with the ground. This way, the body base of support (BoS) and ground COM projection can be established. Finally, with this information, a simple algorithm is able to assess, from a dynamic perspective, whether a fall has taken place.

5.1 Material and methods

5.1.1 Human balance

Humans are biped beings whose COM is placed over the support area determined by their feet during the activities that they develop erected. These activities include standing, which involves both feet in contact with the ground, walking, which mostly implies one foot touching the ground, and running or jumping, activities associated to phases with no ground contact.

It is widely assumed [440] that the model of the inverted pendulum is a reasonable approach to the study of human balance in a quiet standing posture. In this model two main forces are considered, the body weight, applied at the center of gravity (COG), which is the projection of the COM onto the ground, and the ground reaction, applied at the center of pressure (COP). This way, both forces are equal and their combined torque is in constant variation to guarantee balance as a result of continuous displacements in the COG and COP positions, either voluntary or not.

The same paper also presents that, regarding anteroposterior (A/P) directional balance in standing positions, the COM acceleration projection onto the horizontal plane is proportional to the distance that separates COG and COP.

$$p\text{-}g = \frac{I\,\ddot{x}}{W\,d} = K\,\ddot{x}$$

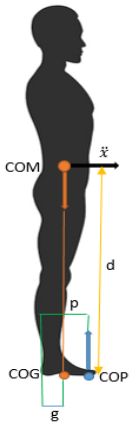


Figure 16. A/P human balance in standing posture.

Being p and g the positions of the COP and COG, I the body moment of inertia, W the weight, d the vertical distance from the ankle to the COM and $\ddot{x}$ the COM acceleration projection onto the horizontal plane.

On the other hand, the stability in the mediolateral (M/L) direction in a standing posture adopts a different approach, as it is based on a load/unload strategy developed by the hips.

During locomotion activities, as shown in figure 17, these two mechanisms control the trajectory of the COP to ensure desired COM acceleration/decelerations are obtained in a constant effort to obtain and regain balance as the body moves.

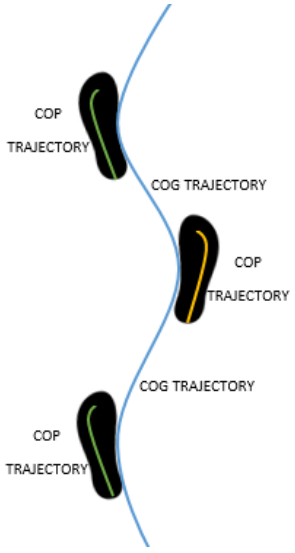


Figure 17. COP and COG trajectories during locomotion.

5.1.2 Human fall problem definition

When for whatever reason the human balance process fails and the continuous activities of the body to regain stability are not successful a fall happens.

In the elderly community, as shown in [441], the 4 most common causes of fall are walking forward (24%), standing quietly (13%), sitting down or lowering (13%) and initiation of walking (11%). Therefore, assessing stability when walking or standing should be the primary target of any fall detection system whose goal is elderly fall detection.

The human standing stability is fundamentally modeled by an inverted pendulum [442] and, under this assumption, the COG position referred to the BoS defined by feet position and status is widely adopted as the main human standing stability indicator. This way, a number of human static stability studies, as the ones in [443]-[446], are based on different derivations of this concept, establishing regions never to be left by the COG in order to keep standing balance.

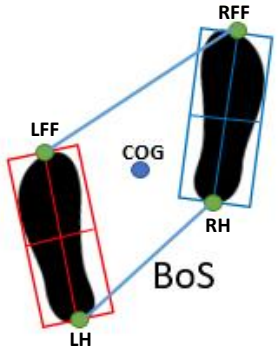


Figure 18. Standing stability diagram where BoS is defined by the Left Forefoot (LFF), Left Heel (LH), Right Forefoot (RFF) and Right Heel (RH).

On the other hand, human locomotion is an inherently unstable activity. This way, during the single support phase, when walking, or during the jumping phase, when running, the COG moves forward, away from the support provided by the standing foot, towards the future position where the swinging foot will land, putting the body in an unstable situation.

The main gait stability indexes are described in [447] by Bruijn. This way, a stable gait is defined as the one that does not lead to falls in spite of perturbations. These indexes assess the ability of individuals to handle perturbations during displacements.

Due to their biomechanics significance, their sound mechanical basis and their capability to predict individual's ability to handle perturbations, the extrapolated center of mass (XCoM) [448] and the foot placement estimator (FPE) [449] will be considered to assess gait stability in this work.

The validity of XCoM to assess both A/P and M/L stability is evaluated in [447], [450] with good results. In addition, this index is also used in [451] to determine M/L stability of above-knee amputees with excellent outcomes.

The foot placement estimator is introduced in [449] and [452] to assess A/P stability and it is then extended to M/L stability in [453].

Extrapolated center of mass

Hof [448] defines three potential states during human locomotion regarding A/P stability. A first one with XCoM behind COP, being both of them behind the forward edge of the BoS (BoS_{xmax}). A second state where both XCoM and COP continue behind BoS_{xmax} but, in this case, COP is behind XCoM and a third case where XCoM is ahead of BoS_{xmax}.

The first case is clearly stable, while the second one, although temporarily stable, will lead to an unstable situation unless action is taken. The third state is unstable and it will be followed by a COG replacement through trunk or extremities movement, so COG is sent back behind BoS_{xmax}, by a step, which will take the situation back to the first phase, or by a fall.

This way, and once the third state is reached during forward movement, a step to place the swinging foot heel ahead of the XCoM must be taken before an unacceptable body sway angle, which widely vary from person to person [454], is reached.

A reverse logic is used to assess A/P stability during backwards movements.

In [448] XCoM is defined, for A/P stability, as

$$XCoM = x + \frac{\dot{x}}{w_0} \quad (1)$$

with $w_0 = \sqrt{\frac{l}{g}}$

Being l the vertical distance between ankle and COM and x the COG position in the A/P axis.

While M/L stability and its relationship with YCoM is extensively described in [450].

$$YCoM = y + \frac{\dot{y}}{w_0} \quad (2)$$

With y being the COG position in the M/L axis.

Foot placement estimator

This index, introduced by Wight [449], tries to determine forward foot placing assuming that the human body behaves as an inverted pendulum, that the angular moment is conserved at heel ground contact and that energy is conserved from that time on until the body maximum potential energy point is reached. Although the model is a simplification of what happens in reality, violations of these assumptions have little effect on the final outcome, as proved in [452] and [455].

This way, when walking, the model, shown in figure 19, assumes a total standstill at body maximum potential energy point, so movement needs to be restarted then. However, this is not what happens when the person does not stop walking at that step and, therefore, foot should actually fall short of the calculated FPE position. This assumption is backed by the results obtained in [456].

In [452], the angular velocity after heel impact assuming an inverted pendulum model $\dot{\theta}_2$ is calculated as

$$\dot{\theta_2} = \frac{mh(v_x cos\Phi + v_y sin\Phi)cos\Phi + I\dot{\theta_1}cos^2\Phi}{mh^2 + Icos^2\Phi} \quad (3)$$

$$\frac{1}{2}(I + m\ \frac{h^2}{cos^2\phi}\ )\ \dot{\theta_2^2} + mgh = mg\ \frac{h}{cos\phi} \quad (4)$$

Being v_x and v_y the components of the COM speed, I the moment of inertia, h the height of COM above ground level and ϕ the angle between the swinging leg and the vertical at heel contact.

As φ is the angle of the rear leg with the vertical at heel contact, equation (3) can be rewritten as

$$\dot{\theta_2} = \frac{mhv_x(cos\Phi + \tan\varphi sin\Phi)cos\Phi + I\dot{\theta_1}cos^2\Phi}{mh^2 + Icos^2\Phi} \quad (5)$$

Equations (4) and (5), combined, lead to ϕ determination that, in turn, can be used to calculate FPE, as leg length is known.

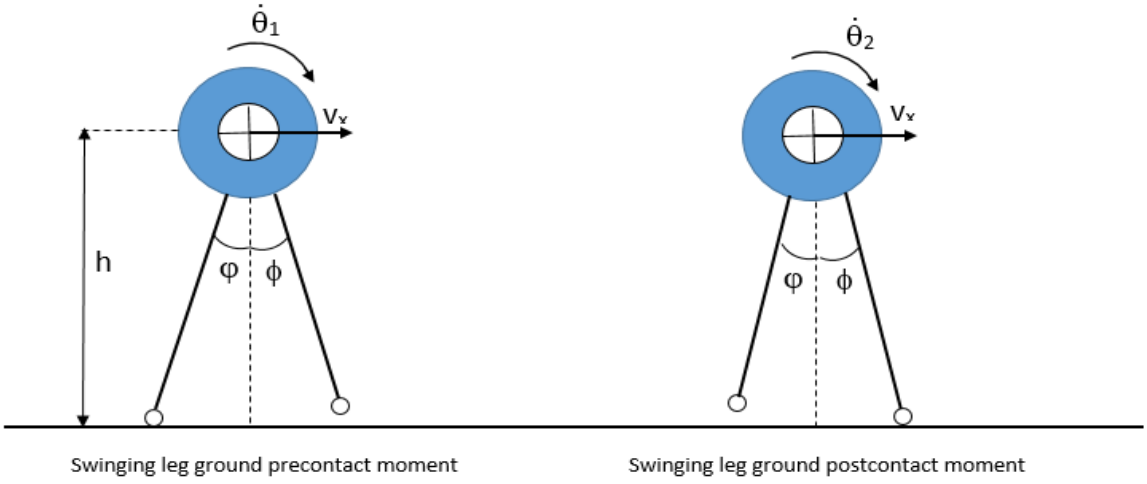


Figure 19. Simplified biped walker before and after swinging leg ground contact.

5.1.3 Dynamic approach

The approach to human fall from a balance and stability perspective represents an alternative to the classical use of kinematic descriptors. To implement this approach, the COG and feet position, together with feet status; either in contact with the ground or not, need to be determined.

To do it, in this work, an end-to-end solution by using an ANN is proposed. The network, illustrated in figure 20, takes as input a series of 2D joint key-points positions (e.g., determined by an off-the-shelf 2D joint detector) and returns the projection of COM and feet joints onto

the horizontal plane, as well as their contact status, allowing, this way, the stability indexes determination.

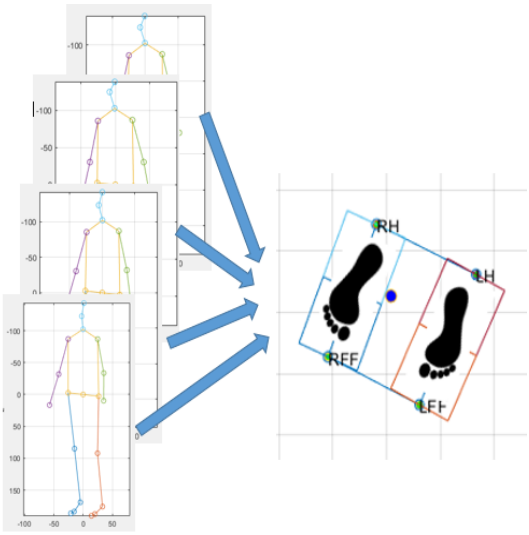


Figure 20. The network is able to provide the projection of COM (blue dot) and feet joints position (Right Heel RH, Left Heel LH, Right Forefoot RFF and Left Forefoot LFF) onto the horizontal plane as well as their status (feet joints in contact with the ground)

5.1.3.1 Data pre-processing

As the only relevant information for our purpose is related to relative joint placement, the array containing the coordinates of the body joints is referred to the root (midpoint between hips). Then, and in order to get temporal coherence, the series of arrays is filtered using a 1-€ filter [457].

Then, every array of the time series, representing the 2D joints position at a specific time, is flattened to a 1D vector to reduce input complexity as much as possible. Finally, the 1D vectors reflecting instantaneous joints position are stacked along a time window in a 2D tensor. This tensor will be the input feeding the ANN.

5.1.3.2 CoGNet Network

CoGNet is an end-to-end deep learning architecture aiming to regress COM projection onto the ground plane from 2D pose series. Furthermore, BoS is also determined, obtaining, this way, the crucial elements to determine human equilibrium and assess stability.

Given the optimal results of the network developed by Pavllo et al. [458] estimating 3D poses from 2D key-points series and the good performance of the one implemented by Zou et al. [459], which is able to determine whether feet are in contact with the ground, we propose an architecture inspired by both networks with a common backbone and an output block which delivers a matrix of 2D positions over the ground plane, $B \in \mathbb{R}^{2n}$, and ground contact labels for both heels and forefeet. The matrix collects the heels, forefeet and CoG position projected onto the horizontal plane.

The model is a temporal convolutional architecture, a structure first introduced by Lea [460]. In this type of structure, the gradient path, between output and input, has a fixed length that highly mitigates vanishing and exploiting gradients, a common problem in other architectures such as RNN's. Convolutional structures also offer precise control over the temporal receptive field, a determinant characteristic to incorporate time evolution in the model, something of extreme importance in a network designed to provide the COG movement. To do it, dilated convolutions [461] are used, so dependencies from previous states are incorporated to the present COG determination. This type of architecture, which models time dependencies, has been successfully applied to different fields, such as semantic segmentation [462], sound classification [463] or image extraction [464].

The embedding block takes as input the 2D tensor. Replicate padding is then used to deal with boundaries and a 1D convolution, with kernel size 3, is applied to get the block output.

The following residual blocks apply 1D dilated convolutions and use ReLU activation functions [465] to reduce added non-linearity as much as possible. In addition, residual connections are used to reduce network-training time [466]. Furthermore, batch normalization methods [467] improve network performance in the environment associated to the noisy inputs coming from 2D pose estimation systems by reducing the internal covariate shift. Finally, Dropout [468] methods allow overfitting prevention during the network-training phase and greatly improve generalization.

The output block applies a convolution covering the entire time window input and, this way, the block input tensor is flattened to a 1D vector. This vector feeds two branches, one providing forefeet, heels and COG projections on the horizontal plane and another one providing ground contact probability to forefeet and heels. Both branches use, for different purposes, a fully connected layer with a sigmoid activation function.

The entire network architecture is shown in figure 21.

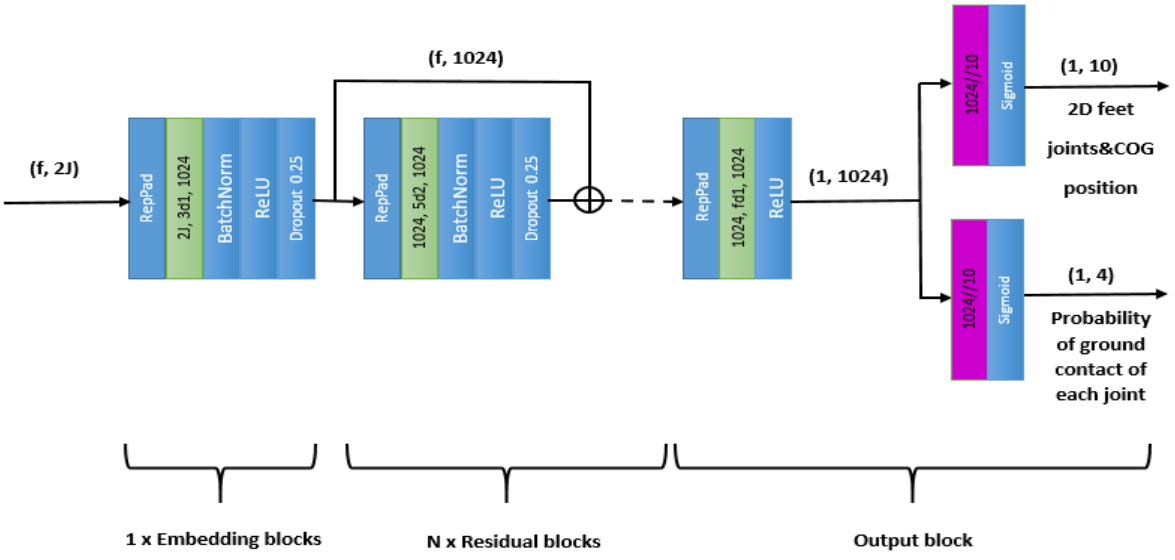


Figure 21. Network structure where f is the number of accepted frames and J the number of joints. Blocks in green represent convolutional layers where the number of input channels, size and dilation of the kernel and number of output channels is indicated.

Final processing of the output time series includes a 1-€ filtering step.

5.1.3.3 Network training

The network is trained using, mainly, the Human 3.6M [321] database, but a number of samples coming from MPI-INF-D-HP [322], NTU RGB+D [227] and Deepcap [323] were added to make the final training set as rich as possible, favoring generalization this way. The ground truth forefeet and heels projection onto the ground plane are inferred from the 3D databases annotations, while the GOG determination is made by projecting onto the same plane the COM position which, in turn, is determined through segmental analysis. Finally, the ground contact status labels are established through the process described below, as there is no publicly available database providing that information.

The final dataset includes 169444 samples that are split into two groups, one for training, containing 136665 samples, and one for testing, which groups the rest of them.

Ground-truth COM determination

COM position can be determined by using either forward or inverse dynamic techniques [469]. The first block of methods aims to measure the ground reaction forces by using multiple force plates, instrumented force treadmills or pressure insoles. This way, a first integration is able to determine COM velocity and, through a second integration, COM displacement can be obtained. The second group of techniques, the inverse dynamic ones, captures the three-dimensional kinematics of the entire body and estimates body COM position by considering all body parts. This second array of methods is also known as segmental analysis and it is based on the study by de Leva et al. [470].

In this work COM position is determined through segmentation by defining body segments from anatomical landmarks, which, in turn, can be obtained from body joints. Different publications approximate standard values for the segments COM and moment of inertia [470], [471]. Based on them, body COM can be obtained as a weighted sum of the different segments' contribution.

Thus, three-dimensional COM position can be determined as:

$$\boldsymbol{X_{COM}} = \frac{\Sigma\, m_i\, x_i}{\Sigma\, x_i}$$
$$\boldsymbol{Y_{COM}} = \frac{\Sigma\, m_i\, y_i}{\Sigma\, y_i} \qquad (6)$$
$$\boldsymbol{Z_{COM}} = \frac{\Sigma\, m_i\, z_i}{\Sigma\, z_i}$$

Being x_i, y_i and z_i the coordinates of the i-th body segment and m_i its mass.

In accordance with this procedure, the COG position is inferred from the three-dimensional joints position provided by the 3D database annotations. To do it, the estimation of body segment mass and its COM placement is obtained by using the data provided by De Leva [470].

Ground-truth ground contact status labeling

Ground-truth labeling should be obtained from a public database. Regretfully, no database of this kind exists, therefore, labels had to be determined through the use of a proper algorithm specifically designed for this purpose.

Following a process similar to the one used by Zou [459] contact labels are automatically created for each forefoot and heel. The joint is declared in contact with the ground when it

meets a double condition. On the one hand, its horizontal average speed during the last 0.1 second needs to be under the 5 cm/sec mark. On the other hand, the joint mean position during the last 0.1 second must maintain a height of 5 cm or lower over the mean height of the joint in the first two seconds of the clip.

Loss function

The objective loss function used for network training is:

$$\zeta_c = \zeta_{2D}(\hat{o}) + w\ \zeta_b(l) \quad (7)$$

The first component is a L2 loss function that compares ground-truth joints and COG projections, o, with network outputs, ô.

$$\zeta_{2D}(\hat{o}) = \|o \quad - \quad \hat{o}\|_2^2 \quad (8)$$

The second component is a binary cross entropy function weighted by a factor w

$$\zeta_b(l) = -\frac{1}{n}\sum_{i=1}^{n} l_i \ln(l_i') + (1 - l_i)\ln(1 - l_i') \quad (9)$$

Where l_i' and l_i are the predicted contact probability and the ground-truth label of the i-th joint.

5.1.4 Fall detection algorithm

In order to develop a fall detection algorithm based on the information provided by CoGNet, stability will be assessed in both the A/P and M/L directions. A/P axis will be defined as the mean feet direction during the last half second and M/L axis as the normal direction to A/P axis.

5.1.4.1 Use of extrapolated center of mass indexes

In the A/P axis, the system will be assessed as stable as long as XCoM is behind BoS_{max} and ahead of BoS_{min}, following a similar rationale to determine stability in the M/L axis using YCoM.

Once the system has been declared unstable, an action is required to take it back to a stable condition. In the A/P axis, one foot must overtake XCoM. This way, when the foot of the swinging leg lands, a new BoS will be defined and the stable condition will be re-declared. M/L stability regain after YCoM has over exceeded the limits of the BoS will require similar actions.

XCoM and YCoM will be determined in accordance with (4) and (5). The vertical component of the COM, l, will be approached by the sacral marker height over ground, as this assumption is accepted as a reasonable one even in activities with important limbs excursions [472], [473], [474]. This way, according to the anthropomorphic information contained in [475] and [476], the mean l for the European population over 70 years old is 97.4 cm for men and 93 cm for women.

Assuming an inverted pendulum model, actions trying to regain stability should be completed prior to complete fall time, which, according to [477], is as follows

$$T(\vartheta_0, w_0) = \sqrt{\frac{l}{2\,g}} \int_{\vartheta_0}^{\pi/2} \frac{d\vartheta}{\sqrt{\frac{w_0^2}{2\,g}\frac{l}{} + \cos\vartheta_0 - \cos\vartheta}} \tag{10}$$

For the moderate sway angles associated to human balance, as ϑ_0 ranges from 2.5º to 7º [446], a person falling time varies between 1 and 1.48 seconds.

This way, the maximum time an unstable condition will be maintained before declaring a fall will be 1.48 seconds if stability is not regained before.

5.1.4.2 Use of the foot position estimator index

The FPE will be calculated at heel landing and compared with the real step length. For this estimation, the human body will be considered a rigid simplified biped walker (fig. 4) whose segment masses, lengths, COM positions and inertial moments are determined using the information provided by De Leva et al. [470].

As in the previous section, the sacral height over ground is assumed to be a good approximation to COM vertical position. The angle φ is calculated by using the information provided by CoGNet and the leg length. Then, through (5) and (6) ϕ can be determined and, once this last angle is known, FPE determination is immediate. Finally, this distance is compared with the real step length given by CoGNet.

5.1.5 Performance evaluation

5.1.5.1 Network evaluation

Error determination most popular metric in the field of 3D human pose estimation is Mean Per Joint Position Error (MPJPE) [478]-[480]. It determines the mean of all Euclidean distances separating the predicted and ground-truth positions of each joint. In this case, a specific evaluation of the network performance to predict each output point is also desired, as the COG placement depends on the position of a good number of joints while the projection of the feet joints onto the ground plane depends on the position of a single joint.

This way, as there are reasons to think that performances could vary depending on the specific output point, the Mean Absolute Deviation (MAD) of each one of them will be used with evaluation purposes. Additionally, the Mean Squared Error (MSE) and median error will be employed to evaluate network quality outputs. These indexes are defined as follows:

$$\text{MAD} = \sum_1^n \frac{|e_i - \bar{e}|}{n}$$

$$\text{MSE} = \sqrt{\frac{1}{\text{n}} \sum_1^n (e_i - \bar{e})^2}$$

Where e_i is the forecasted determination error of the i-th sample and $\bar{e}$ is the mean error.

On the other hand, Accuracy, the most used evaluation metric for binary classification [481], is used to evaluate the network performance labelling contact status with the ground of every foot joint.

$$\text{Accuracy - AC} = \frac{TP+TN}{TP+TN+FN+FP} \text{ x } 100$$

With TP, true positive; TN, true negative; FP, false positive and FN, false negative.

5.1.5.2 Fall detection algorithm evaluation

The indexes used to evaluate the fall detection algorithm, as they are the most common ones in this area [12], are the following ones:

$$\text{Sensitivity – SE} = \frac{TP}{TP+FN} \text{ x } 100$$

$$\text{Specificity – SP} = \frac{TN}{TN+FP} \text{ x } 100$$

$$\text{Accuracy – AC} = \frac{TP+TN}{TP+TN+FN+FP}$$

5.2 Results and discussion

5.2.1 Network implementation

The network is implemented in PyTorch with an Adam function used as optimizer and a learning rate of 10^{-6}. The chosen batch size is 256 and the network is trained for 80 epochs.

The selection of 1024 channels as the number of working channels for this network responds to a documented [458] reasonable trade-off between network performances and required computational processing power for similar networks.

The definition of w, the weighting value of the loss function, is critical, as it will determine how balanced is output accuracy. Inadequate w selections will lead to output accuracy biases in favor of labeling or position determination.

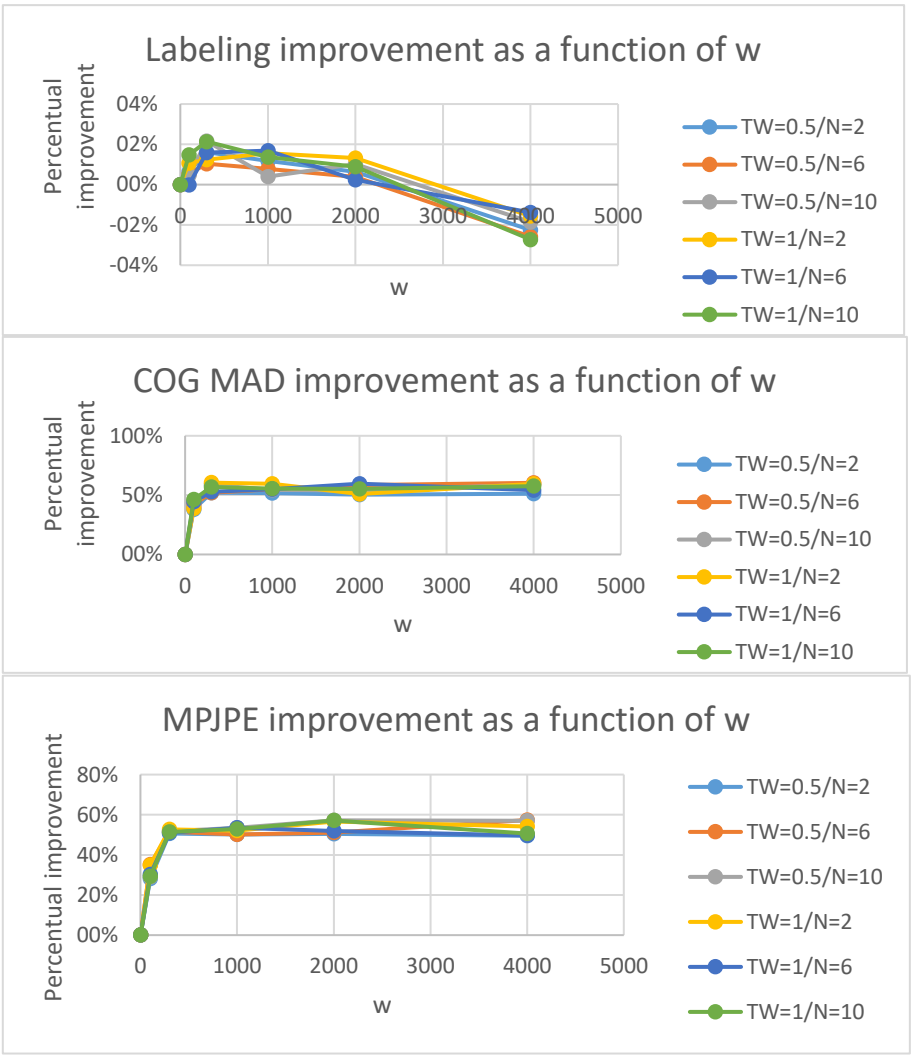


Figure 22. (a) Ground Contact labeling improvement referred to a benchmark error of w=1 as a function of w value. (b) COG MAD. (c) Mean Per Joint Position Error (MPJPE)

This way, as shown in figure 22, MPJPE and COG position determination improvements saturate very quickly with values of w over 500, while increasing values of that parameter decrease labeling performance at a very slow rate for all selected network configurations. Bearing those ideas in mind w is assigned a value of 2500, as that figure yields a reasonable trade-off between joint position placement and joint contact status accuracies.

To determine properly the number of network residual blocks (N) an optimal trade-off between computational cost, performance and considered number of input frames (W) should be reached.

A good approximation to the computational cost per predicted frame could be the required number of floating-point operations to be executed or FLOPs (Floating-point operations). This

way, the needed operations to complete the dot product of two vectors of size B would be B + (B – 1), as B multiplications and B-1 additions are required to do it.

Following that rationale, the processing power required by a fully connected layer, given all input neurons are connected to all the output ones, equals (2 I – 1) O FLOPs, where I represents the number of input neurons and O the number of output ones. For the particular case of this architecture, given the high number of input neurons, the cost can be approached by 2 I O.

For the case of the 1D convolutions of the network the needed computational power in FLOPs is given by

$[C_{in}\ k + (C_{in}\ k - 1)]\ Co\ (W + Pa - Re + 1)$

Which, given that $2\ C_{in}\ k$ is substantially bigger than 1 in all cases, the previous equation can be approximated by

$2\ C_{in}\ k\ Co\ (W + Pa - Re + 1)$

Where C_{in} and C_o are the number of input and output channels, k is the kernel size, W is the width of the input matrix, Pa is the padding contribution and Re is the receptive field of the dilated convolution.

Therefore, the total computational cost of the network is:

Table 13. Computational cost due to convolution layers.

Block	Input channels	Output channels	Kernel width	MFLOPS
Embedding	Joints x 2 (J x 2)	1024	3	0.0123J W
Residual	1024	1024	5	10.48576 W
Output	1024	1024	W	1.04857 W

Table 14. Computational cost due to fully connected layers.

Block	Input neurons	Output neurons	MFLOPS
Output	1024	10	0.02048
	1024	4	0.008192

Which, in a single expression, yields:

MFLOPs = 0.028672 + W (1.04857+0.0123 J + 10.48576 N)

Where J is the number of considered joints and W can be obtained from time window duration and sample rate.

Figure 23 shows the relaions between time window, number of residual blocks and network output accuracy. Additionally, the required processing power is also depicted.

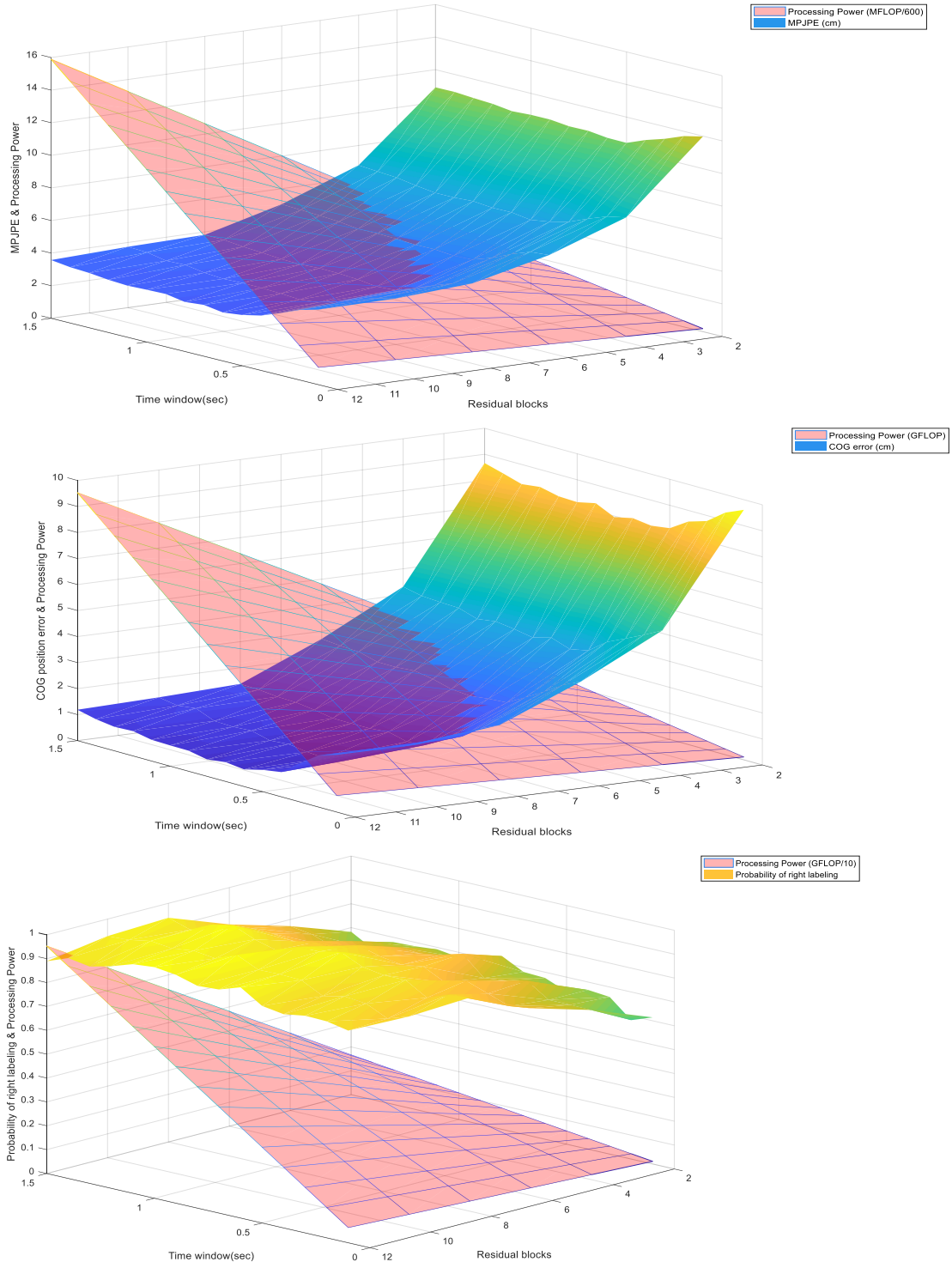


Figure 23. (a) MPJPE, (b) COG position error (c) Probability of correct labeling (Accuracy) plus required processing power for all three cases as a function of the number of implemented residual blocks and the accepted time window.

As depicted in Figure 23, system MPJPE saturates when the number of residual blocks over exceeds 8 and it shows a minimum with time frames within the bracket from 0.4 to 0.8 seconds. Additionally, COG determination error saturates with six or more residual blocks and reaches its lower values within the time frame from 0.5 to 1.0. Finally, the probability of correct labeling is over 0.9 when the number of residual blocks exceeds eight and the time frame is between 0.4 and 1.3 seconds. For those reasons and trying to keep the required computational complexity as low as possible, the final network architecture is set to accept time frames of 0.5 seconds and implements 8 residual blocks.

This architecture, for the conventional number of joints provided by 2D networks, 17, and a selected frame rate of 30 Hz requires, according to (11), a computational power of 1.22 GFLOPS. After its training, the network was speed tested on two platforms, a first one with a NVIDIA RTX 3080 GPU, yielding a speed of 527 frames per second (FPS), and a second non-GPU one with an Intel Core i9-12900 processor, whose tested speed was 105 FPS. In both cases, speeds are well over the one required to guarantee real time network operation. Furthermore, and given the performances of some of the chipsets mounted on new mobile devices (Table 15), network operations in real time, even on this kind of platforms, should not be problematic.

Table 15. 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

5.2.2 Network evaluation

With the above-described architecture, the error distribution per joint on the horizontal plane for a batch of 256 samples of the database has the following spatial distribution.

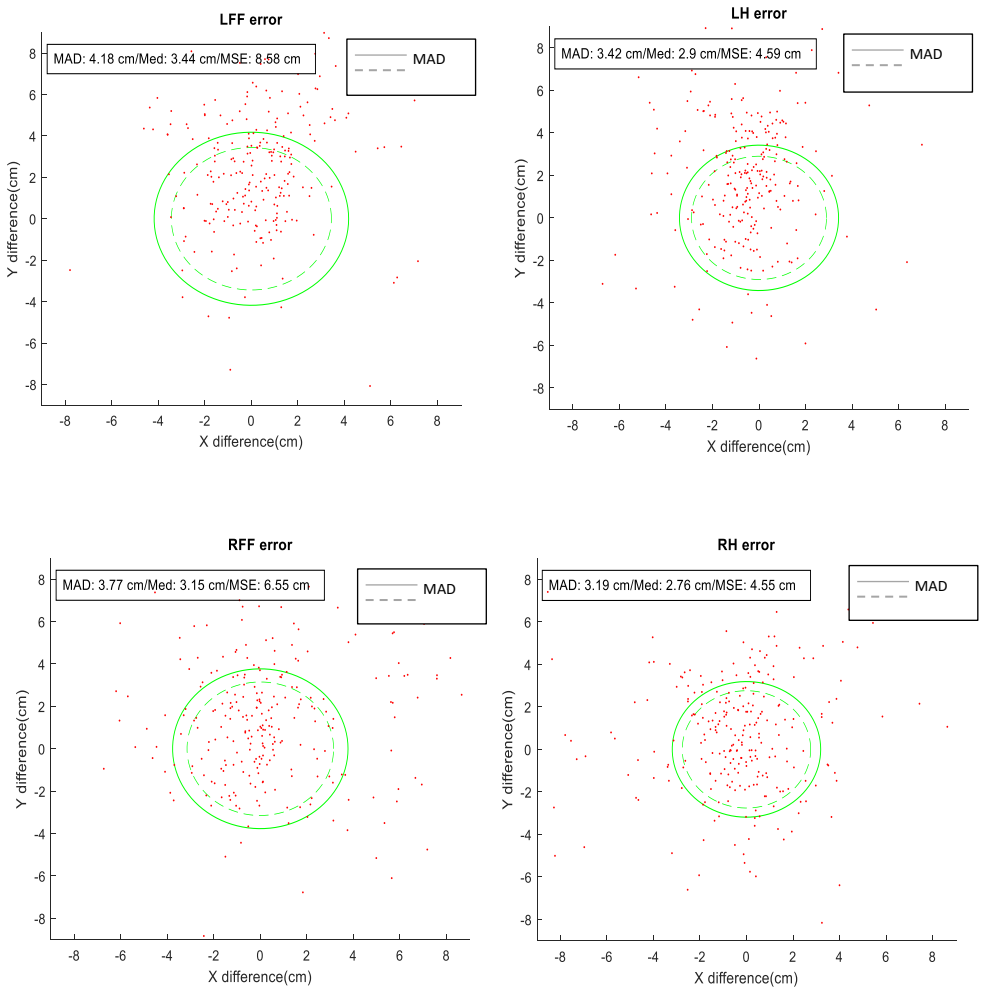


Figure 24. MAD, MSE and sample median in cm per forecasted joint plus sample distribution. (a) Left Forefoot (LFF). (b) Left Heel (LH). (c) Right Forefoot (RFF). (d) Right Heel (RH).

As shown in figure 24, the forecasted MAD remains in the bracket between 3.19 and 4.18 cm with a MSE between 4.55 and 8.58 cm and a slight directional bias (mean forecasted joint bias 1.15 cm. Table 16).

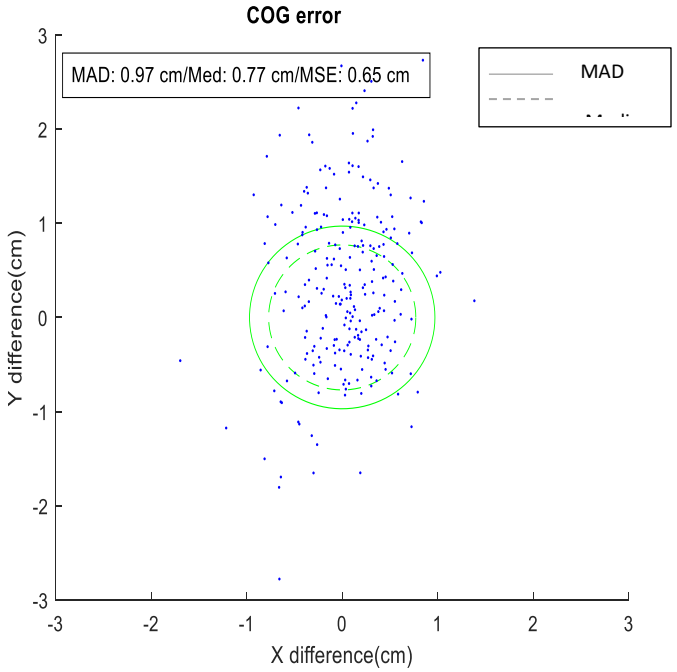


Figure 25. MAD, MSE and sample median in cm and sample distribution of the forecasted COG.

The forecasted COG error, as shown in figure 25, is quite slim, with a MAD below 1 cm, a MSE of 6.5 mm and a reduced directional bias (3.8 mm. Table 16)

Table 16. Mean ΔX, ΔY, MAD, MSE and median of all forecasted joints and COG. All values are in cm.

Joint	Mean ΔX	Mean ΔY	MAD	MSE	Median
LFF	0.43	2.80	4.18	8.59	3.45
LF	-0.55	1.88	3.42	4.59	2.90
RFF	0.29	1.17	3.77	6.55	3.15
RF	-0.44	0.50	3.19	4.55	2.76
COG	0.05	0.43	0.97	0.65	0.77

Additionally, for the implemented architecture, labeling Accuracy reaches 95.8%.

The outcome of the network is shown in figure 26.

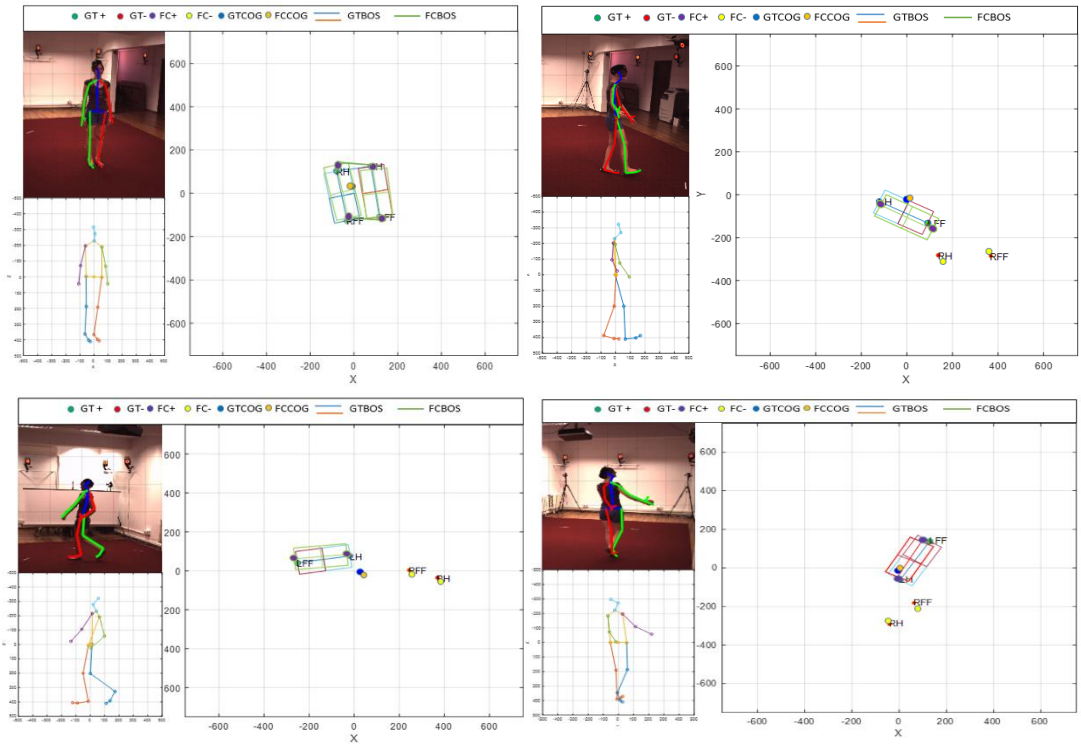


Figure 26. Different network presentations including joint position and indications of positive and negative ground contact, both ground-truth (GT), and forecasted (FC); COG position, both ground-truth and forecasted (GTCOG & FCCOG) and base of support.

5.2.2.1 *Network sensitivity to detected number of joints*

Simplified determinations of the body COM are made in a number of studies, starting by Inman [482], who proposed the use of sacrum as an approximation to its position. This proposal was revised in different studies [473], [474], concluding that the sacral marker can indicate COM position reasonably well in the vertical axis. In a later study, Yang [483] affirms that sacral marker can characterize COM in all axis for walking and, to a lesser extent, upon slips. Other human activities, like running, tilt pelvis forward and induce greater excursions of the upper and lower limbs, changing, this way, body mass distribution and displacing COM from the sacrum, especially in the M/L axis. This conclusion is reached by Napier [472], whose study suggests that a single sacral marker could be valid to estimate COM in the vertical and anteroposterior directions during the stance phase of running.

Other studies, like the one conducted by Gill [484], try to determine body COM during running by disregarding head and arms, concluding that, although its position can be reasonably well inferred in the A/P and vertical axis, in the M/L axis COM trajectory is poorly predicted.

Therefore, human COM three-axial position determination can be inferred from sacral marker in activities not implying high upper or lower limb displacements. However, activities associated to important extremities movements require limb movement consideration if M/L COM position needs to be determined.

On the other hand, both feet joints projections onto the horizontal plane and their contact labelling could be very sensitive to a correct detection of both feet and leg joints by the 2D network.

To check the network sensitivity to passed 2D joints an error determination is carried out for the four following cases (Table 17).

Table 17. Network performances with decreasing number of joints. In 1 the head joints are disregarded and in the three following lines the omissions of the previous lines are maintained and extra ones are added. All distances are in cm.

	Forecasted feet joints projection		COG		
Considered joints	MAD	MSE	MAD	MSE	**Labelling Accuracy**
All	3.64	6.29	0.97	0.65	95.8%
1- Previous - head joints	3.95	7.18	1.13	0.92	94.2%
2- Previous - arms joints	4.52	9.16	1.98	1.06	92.2%
3- Previous - feet joints	7.63	11.55	2.15	1.26	76.8%
4- Previous - legs joints	28.46	25.23	4.87	4.42	19.7%

The network performances decrease as the number of joints descends. In the first case, the head joints are omitted, revealing their discreet contribution. When arms joints are disregarded, the contribution to performance degradation slightly increases and, when feet joints are also eliminated, although the COG determination remains quite stable, both forecasted feet joints projections and percentage of correct labelling substantially decrease. Finally, when leg joints are also omitted performances severely decrease.

5.2.3 Fall detection algorithm evaluation

The NTU RGB+D 120 [485] was the dataset used for fall detection algorithm evaluation. This comprehensive database includes the data of the NTU RGB+D one and adds some extra information. The total set contains 114480 video clips and their associated skeleton data. This skeleton information is inferred from the raw data captured by three Kinect V2 cameras. All these actions are grouped in 120 classes, each one of them trying to illustrate normal human daily life activities. One of these classes is falling down, which includes 316 falls taken from 3 different perspectives.

The 26 actions whose video clips include more than one person were removed from the validation dataset, as CoGNet was trained in scenarios of a single individual. This way, the total number of initially considered video clips was cut down to 89652, grouped in 134 classes of actions.

Finally, all falls were manually checked to assure that the selected ones were complete falls and none of them finished in stable positions such as crouching. Additionally, all non-fall selected actions were also manually checked to assure no unstable permanent situations (COG out of BoS determined by feet in standing situations) were maintained (e.g., being sat) or purposefully searched (e.g., sitting down).

The validation dataset obtained this way includes 238 falls and 2128 daily live activities coming from the 133 activity classes. All the considered actions are recorded from three different perspectives and meet the criteria described in previous paragraphs.

5.2.3.1 Validation based on the use of skeleton data

The great advantage of the use of skeleton data over other methods is that, as three different cameras are used to record the action, no joint is occluded at any time and, therefore, all of them can be correctly projected on a 2D view.

All falls in the dataset start from a standing position and, therefore, no FPE index can be calculated, as this index calculation requires coming from a situation of movement that must include taking, at least, one step prior to the fall event.

Table 18 includes the performance evaluation indexes of our algorithm, the extrapolated center of mass algorithm, applied to the described dataset. Additionally, table 19 compares accuracy indexes of different methods on NTU RGB+D dataset. Finally, the confusion matrix is presented in table 20 and figure 27 presents the outcome of the network.

Table 18. X/YCoM algorithm performance indexes.

SE	SP	AC
99.16%	99.25%	99.24%

Table 19. Accuracy comparison of different methods on NTU RGB+D dataset.

Method	Data	AC
Xu et al. [486]	RGB	91.70%
Anahita et al. [487]	Depth	96.12%
Han et al. [488]	Depth	99.20%
Ours (X/YCoM algorithm)	**Pose**	**99.24%**

Table 20. Confussion matrix.

		Forecasted	
		Fall	Not a fall
Real	Fall	708	6
	Not a fall	48	6336

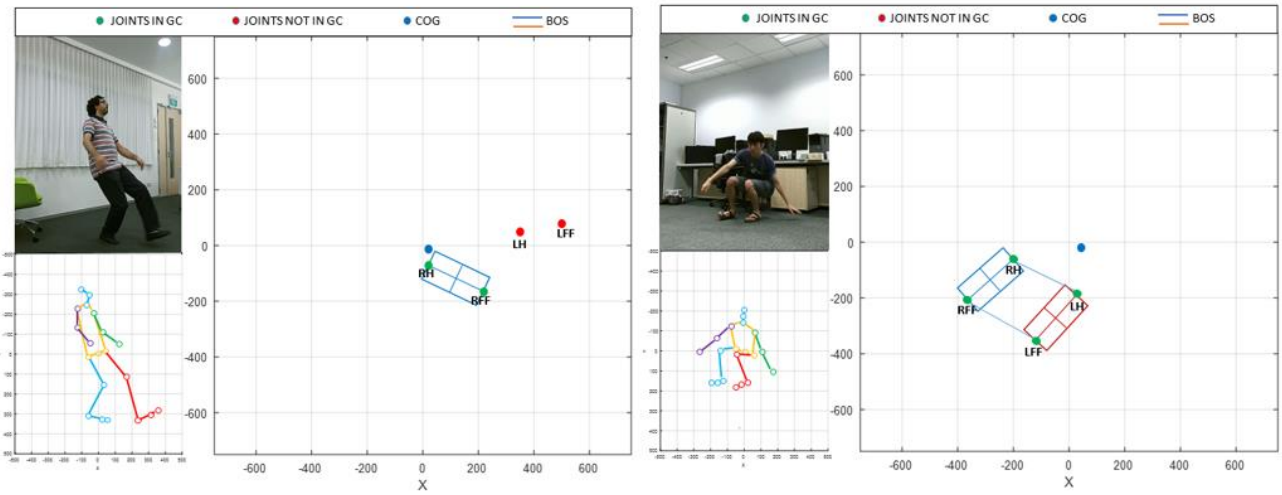


Figure 27. Falls presentations from the dataset including joints position, CoG and BoS.

5.2.3.2 *Validation based on monocular images*

This second validation block intends to evaluate our algorithm performance when monocular images, the cheapest and most common source of information in real-world situations, are used and no skeleton information is available. In this case, the joints positions will be provided by a 2D pose network that estimates them from the images. There is a number of neural networks able to do it and, in this case, Google's MediaPipe, a very light network based on BlazePose [489], will be used due to its accuracy and simplicity.

The used dataset is the same one than in the previous section but 30 extra falls have been added. Half of them come from Multiple Cameras Fall [490] dataset, a database that includes several falls in different environments taken from 8 different perspectives. The other half can be found in UR fall [491] dataset and, in this case, all of them are taken from a single perspective.

All the added falls, unlike the original ones contained in NTU RGB+D, start in a walking situation and, therefore, are very relevant, as, for the elderly community, the majority of falls start in a walk situation [441]. However, they come from databases with no skeleton data associated, as no publicly accessible dataset including this type of falls and their associated skeleton annotations has been identified [12]. For that reason, these falls were not included in the previous validation block.

These new falls, which include previous steps to the fall event initiation, allow FPE determination. However, as previously explained, the real foot position falls short of the calculated FPE most of the times. This documented tendency is experimentally verified by using the dataset, finding substantial differences between step length deficiencies in falls and in any other activity. Figure 28 presents the obtained data showing maximum step length reductions as a function of the movement angle referred to the A/P axis in falls and in the rest of activities.

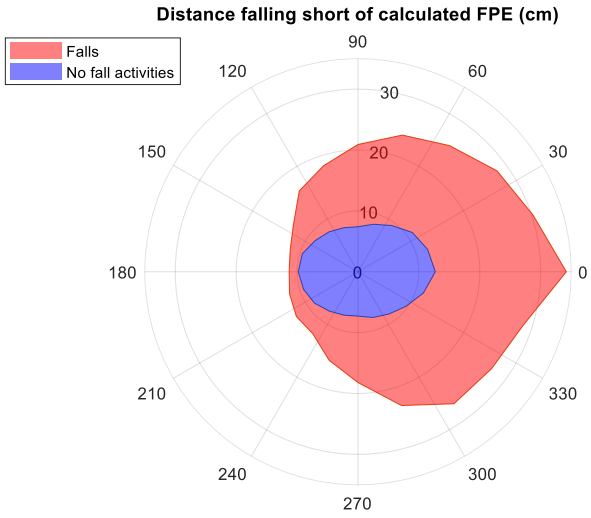


Figure 28. Maximum step reduction referred to calculated FPE.

These findings are in line with Millard's work [452], whose results place this difference, for natural walking speeds, around 10 cm. This way, an alternative algorithm for walk forward situations can be established based on step distance reduction referred to the calculated FPE. In these situations, either a positive fall declared by the extrapolated center of mass algorithm, or one determined by the FPE criterion, whose threshold is set at 12 cm, will be considered a valid fall declaration.

Table 21 presents the performance evaluation indexes for both the X/YCoM algorithm and the one including the FPE criterion support. Additionally, table 22 includes the confusion matrixes for both cases.

Table 21. System performance indexes.

	SE	**SP**	**AC**
No FPE criterion implemented	90.46%	98.34%	97.41%
FPE criterion implemented	97.64%	98.09%	98.04%

Table 22. Confussion matrixes.

		Forecasted			
		No FPE		FPE	
		Fall	Not a fall	Fall	Not a fall
Real	Fall	768	81	829	20
	Not a fall	106	6278	122	6262

As expected, the performance indexes have degraded because of the diminished precision in the joint position determination, mainly as a consequence of occlusion and illumination phenomena. MediaPipe default visibility parameter to consider a joint well positioned enough to be presented to the user is 0.5. This variable reflects the probability the network gives to correct joint identification and its precise positioning, being 0.5 a reasonable threshold for the joint to be considered. In the used dataset, the mean number of joints with a visibility value over 0.5 is 84.9%. That is the reason behind performance degradation.

The implementation of the FPE criterion has had a substantial positive impact on system sensitivity, with an improvement of over 7%, as the number of false negatives has been drastically reduced. The number of false positives has very slightly increased with a tiny impact on specificity and the system accuracy has increased to over 98%. These figures leave few doubts about the positive impact of FPE criterion inclusion to categorize fall events starting in a walking situation.

Figure 29 illustrates the network's outcome.

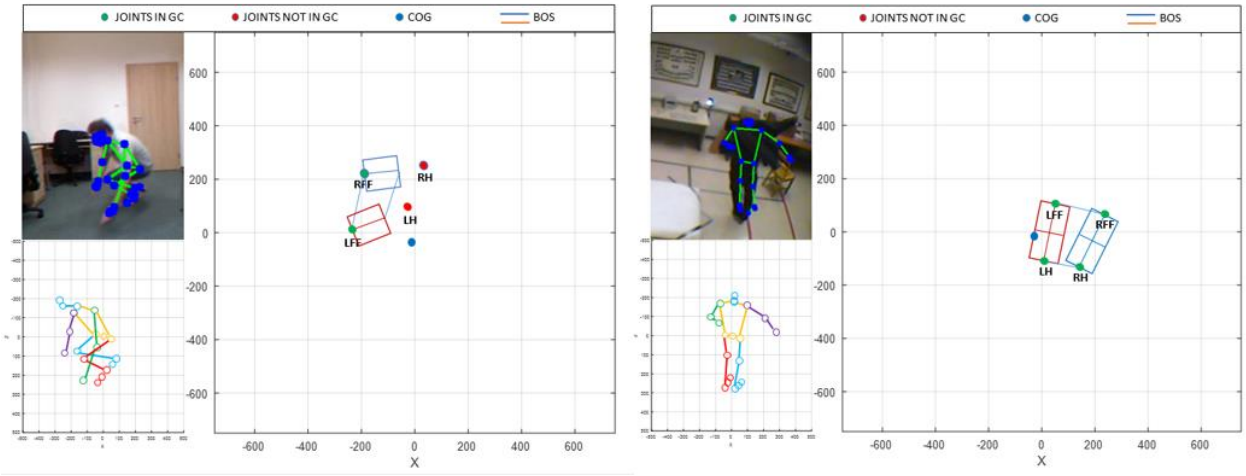


Figure 29. Falls presentations from the UR fall and Multiple cameras fall datasets. Joints position, CoG and BoS are presented.

Finally, table 23 compares our algorithm accuracy with the ones offered by other state-of-the-art methods using RGB monocular images on UR fall dataset, the most popular dataset in the area of monocular fall detection system development [492]. The results prove that it slightly outperforms other state-of-the-art fall detection methods in laboratory conditions and given the independence of this approach from the conditions of the fall, real or simulated, it may behave much better than the rest of the systems in a real situation.

Table 23. Accuracy comparison of different methods on UR fall dataset.

Method	Data	AC
Qingzhen Xu et al. [226]	RGB monocular images	91.7%
X. Cai et al. [231]		96.2%
X. Wang et al. [234]		97%
S. Kalita et al. [209]		94.28%
C. Menacho et al. [493]		88.55%
D. Kumar et al. [199]		98.1%
S. Kasturi et al. [176]		96.34%
Ours (X/YCoM algorithm)		**98.57%**

5.3 Conclusions

This work successfully implements a new approach to automatic fall detection. Previous systems extract and classify kinematic features from video clips of public fall datasets, which record movements and simulated falls of young actors and volunteers. However, the differences in the way young and elderly people move and fall are well documented. Therefore, these systems' performances in the real world may be poorer than documented.

Our method approaches fall detection from the perspective of human balance and extracts features from images to assess whether a person maintains equilibrium. By extracting dynamic features from images, the problem caused by the differences in the way young and elderly move becomes irrelevant. This is because all falls are always a consequence of a failure in the continuous effort of the body to keep balance, regardless of any other consideration.

The dynamic descriptors employed in this work are used to calculate stability indexes, which assess whether a fall has taken place. These descriptors are provided by CoGNet, a temporal convolutional architecture able to regress the COG and feet joint positions projected onto the ground plane. Additionally, the network determines the ground contact status of the projected feet joints, providing all the information needed to assess human stability from a dynamic perspective. CoGNet takes a sequence of 2D joint positions over time as input and is light enough to be run in real-time.

Using the information provided by CoGNet, a simple algorithm like the one proposed in this work can determine whether a fall has taken place. This algorithm is extremely precise when the joint coordinates passed to CoGNet are accurate and complete, as is the case when working with skeleton data. Under these circumstances, the algorithm's results match or even exceed other state-of-the-art systems based on kinematic descriptors in laboratory conditions, as shown in Table 19. These performances slightly decrease when accuracy diminishes due to occlusion or illumination phenomena, common problems when using monocular images. However, even under these conditions, Table 23 shows that our algorithm's outcomes still exceed those given by other state-of-the-art automatic fall detection methods.

Finally, a number of stability indexes based on COG movement are used to evaluate gaits with a higher probability of falling, allowing interventions to improve walking and prevent falls. The information provided by CoGNet could be used to easily calculate those indexes

using monocular images of the patient instead of the traditional, more complex, and expensive systems based on force plates used in the Computerized Dynamic Posturography systems.

6 System validation

This chapter is focused on assessing the performances of an automatic fall detection system based on FIR imagery as a solution to the problem described in chapter five.

6.1 Methods and materials

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

For 2D human pose estimation, the networks considered in chapter six are utilized as they represent most of the state-of-the-art human pose recognition networks ever developed. These networks have been trained using the FIR-Human dataset following the procedures detailed in chapter six. 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 chapter 7, whether a fall has taken place.

The dataset used for validation consists of the 72 falls from the 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.

The performance evaluation indexes used to assess the system are SP, SE, and AC, which are the same as those employed in chapter 7 to evaluate the fall detection algorithm, as explained there.

Finally, the FPE criterion is active at all times in the fall detection algorithm, although its effects will only be noticeable when steps prior to the fall have been taken.

Figure 30 illustrates the network's outcome.

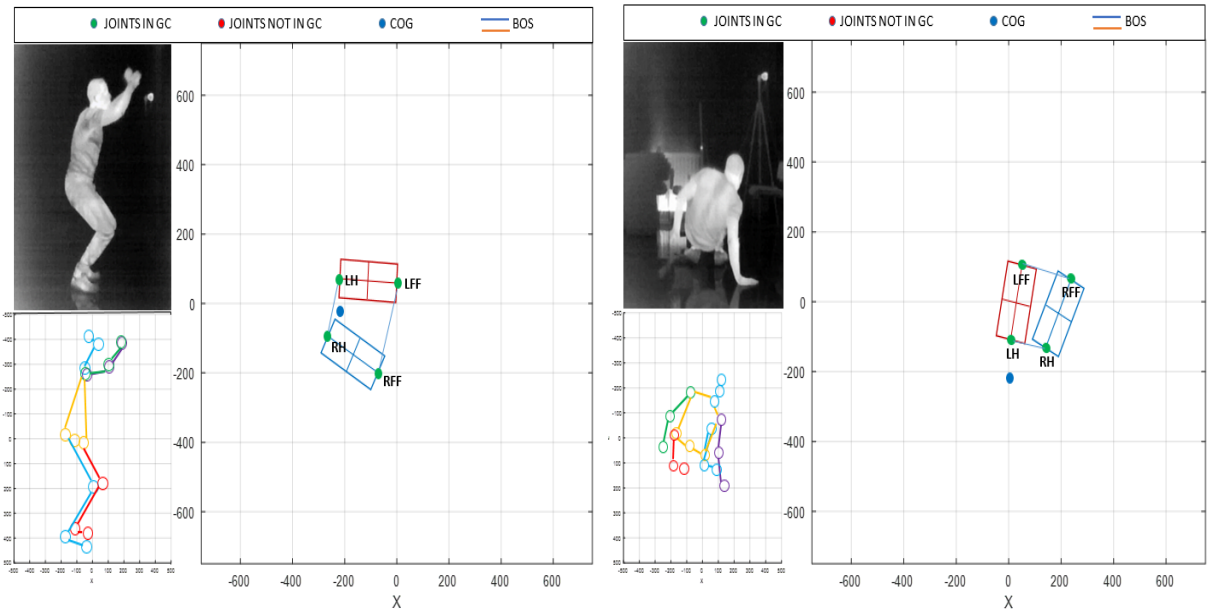


Figure 30. Presentations of (a) volunteer playing basketball (b) volunteer falling backwards from the FIR-Human dataset. Joints position, COG and BoS are presented.

6.2 Results and discussion

Table 24 presents the performance evaluation indexes of the system described in the previous section, and its results vary widely depending on the performances of the used 2D network. Additionally, Table 25 includes the confusion matrices for all the considered cases.

Table 24. System accuracy comparison on FIR-Human dataset by type of 2D network employed.

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

Table 25. Confussion matrixes.

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

These results align well with the findings shown in Table 17, where CoGNet's sensitivity to lower joints fading becomes evident. 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

11. As the networks become more performant, they yield better results in the lower rows of the table.

Moreover, the less performant networks show higher sensitivities to occlusion. The combined effect of both phenomena leads to inaccurate joint coordinates being passed to CoGNet, especially for the joints that are farthest from the head, such as knees, ankles, and feet. These inaccuracies result in incorrect Base of Support (BoS) definitions, causing the system to misjudge situations where the Center of Gravity (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.

Therefore, all false negatives are related to incomplete falls or falls recorded from angles that induce occlusion effects in the lower joints during the last phases of the fall. In such cases, the COG is extremely close to the limits of the BoS, which is observed in all false negatives associated with the most performant networks. On the other hand, situations with common occlusion phenomena are responsible for the majority of false negatives in the case of less capable networks.

Finally, 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, or to occlusion phenomena. Once again, misjudgments by CoGNet regarding the COG's position in relation to the BoS are responsible for the false positives in the most performant networks. The less accurate networks, in addition to this, also suffer from the effects of occlusion phenomena, further contributing to their mistakes in this area.

6.3 Conclusions

This work implements, for the very first time, an automatic fall detection system based on FIR imagery as a solution to the problem described in chapter five, particularly focusing on poorly illuminated environments. The system integrates a 2D human pose estimation network, CoGNet, and the fall detection algorithm described in the previous chapter.

The selected 2D pose estimation networks represent the majority of state-of-the-art networks ever developed for this purpose. They were trained using the FIR-Human dataset following the methods explained in chapter six.

The process involves passing the output of the 2D pose estimation networks, which contains the positions of the main body joints, to CoGNet. CoGNet then delivers the position of the body's COG and BoS. Finally, the fall detection algorithm described in chapter seven utilizes the information provided by CoGNet to determine whether a fall has taken place.

The accuracy of fall determination varies depending on the precision of the joint coordinates delivered by the pose estimation network. Inaccuracies in this determination result from two main phenomena: failures in detecting joints that are distant from the head, leading to inappropriate BoS limits establishment, and joint occlusion due to the angle of record. The most performant pose estimation networks' failures are primarily a result of the first phenomenon, while less capable networks also encounter occlusion issues, contributing to output inaccuracies.

The accuracy indexes presented in Table 24 for the most performant network, ViTPose, show values very close to the ones obtained by CoGNet and its associated fall detection algorithm when they work with visible light spectrum images. 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 sector of elderly care.

7 Conclusions

This thesis addresses the ecosystem of automatic fall detection systems, an area with a high number of published papers, aiming to solve some of the real problems faced by the community of dependants and those contributing to their quality-of-life improvements.

The work begins with a thorough review of the state-of-the-art in the area to establish a common understanding that all systems developed for fall detection purposes fall into one of three categories: wearables, ambient, and vision-based systems.

Wearable systems use sensors carried by the monitored person to evaluate their movements and offer good performance due to the proximity of sensors to the person. However, this also leads to their main weakness - dependency on batteries and discomfort for the user.

On the other hand, ambient and vision-based systems distribute sensors in the environment surrounding the monitored person, offering more comfort for users but at the expense of lower system accuracies and movement restrictions.

Both types of systems face low user acceptance, and the majority of datasets used for system training and validation do not contain real fall data but instead rely on falls performed by actors or volunteers, which poses a challenge for generalization to the dependent community.

The next step after reviewing the state-of-the-art was to identify real problems that could be solved by such systems. Extensive interviews and online polls were conducted to identify scenarios where users would accept a fall detection system. One such scenario is associated with semi-supervised patients getting up at night when they may be disoriented, increasing the likelihood of a fall going unnoticed until the next day.

In this situation, ambient systems are preferable, as wearing sensors in bed is uncomfortable. Among ambient systems, a vision-based system is a desirable option due to its accuracy and maturity. However, there is a limitation - no existing vision-based system can properly work in the low illumination conditions associated with this scenario.

A vision-based system working on FIR imagery would be a suitable solution to this problem, as it functions well in low illumination conditions and preserves privacy, addressing concerns expressed by various groups related to elderly care.

7.1 Research questions

After the previous introductory section, the three main research questions that arise are as follows.

7.1.1 Research question 1: What are the real problems of the dependant community that could be solved by an automatic fall detection system?

This question, extensively reviewed in chapter five, is answered by conducting the most comprehensive set of interviews and online polls ever conducted on the subject of automatic fall detection systems. A total of 36 open-ended interviews and 146 polls were conducted to extract conclusions regarding these systems. The interviews and polls were answered by individuals belonging to all identified stakeholder groups, including the elderly, their friends and family, their home-care givers, nursing home managers, nursing home care givers, and paramedics and emergency medical personnel who are in contact with the elderly community.

The polls and interviews focused on four thematic aspects: the degree of confidence in automatic fall detection systems, user needs and requirements, privacy protection, and usage environment.

Confidence in these systems varied significantly across the considered groups, but it became clear that they are relatively unknown, leading to relatively low levels of confidence. Nevertheless, when the use of these systems can reduce costs or care burden, they could be accepted, despite human supervision always being the preferred choice.

The main user requirement across all groups is reliability. Operators also require user-friendly and easy-to-use systems, while other groups express concerns regarding cost and low reaction times to give a prompt response to a fall event, maximizing survival rates.

Privacy protection is a concern for all groups, with worries about cyber attacks and unlawful information retrieval by personnel with physical access to the systems, especially in the case of vision-based systems. For this particular category of fall detection systems, vision-based ones, a preference for the use of infrared or very low-resolution images is expressed, as these types of data preserve privacy.

The environments where these systems are most likely to be deployed are those where human supervision is not a viable option, and patients are not continuously supervised due to their moderate level of dependence. One common scenario, mentioned by many interviewees, is associated with semi-supervised patients getting up at night to go to the toilet.

This scenario is suitable for fall detection ambient systems, as wearing a sensor in bed is uncomfortable, making wearable sensors less desirable. Among ambient systems, vision-based ones, given their higher degree of maturity, seem to be the optimal choice. However, existing visual-based fall detection systems do not function properly in low or non-illuminated situations, as they typically work with visual or near-infrared spectrum imagery.

7.1.2 Research question 2: How can visual-based fall detection systems work properly in low or non-illuminated environments?

All visual-based fall detection systems presented in papers or commercially distributed work with visual or near infrared spectrum images and, therefore, they cannot work in non-illuminated environments.

According to the state-of-the-art presented in chapter 4, the most performant systems are based on neural networks. However, as it is detailed in chapter 7, all these systems present a generalization problem that will be touched in the next research question. In chapter 7 a solution to this generalization problem, based on the use of 2D human pose estimation neural networks, is also presented.

On the other hand, the use of FIR imagery represents an optimal approach to the problem posed at the end of research question 1, as they are independent of illumination conditions and, additionaly, preserve privacy, an issue several groups related to elderly care have expressed concern for.

This way, the visual-based fall detection system proposed in this thesis to work in non-illuminated environments is based on the use of 2D networks. These networks must be first identified and then properly trained to identify joints positions on FIR images.

Chapter six presents an extensive state-of-the-art study of these networks, identifying the most performant ones. This chapter also details the elaboration of FIR-Human, an extensive FIR dataset containing a large number of images of several volunteers, together with the

position of their joints, both in 2 and 3 dymensions. Additionally, the chapter covers the training of the networks using the FIR imagery provided by FIR-Human.

The results of the validation phase of all these networks are presented also in chapter six showing results very in line with the ones obtained using visual spectrum datasets. In overall terms, transformer-based architectures over-perform convolutional ones and regression philosophies yield poorer results than heat-maps ones.

Additionally, higher input resolutions allow models to yield more accurate results at the cost of highter computational requirements and, in all cases, the head is the easiest key-point to spot while accuracy to place joints grows with the distance to the head making ankles and feet the most difficult key-points to identify and place.

7.1.3 Research question 3: How can the problem of generalization be overcome in the case of visual-based fall detection systems?

Chapter Seven tries to explain this question. All systems reviewed in this thesis extract features that characterize falls. This feature extraction is made during the training phase of the system, and to do it, a dataset properly labeled must be used. However, all publicly available datasets have been recorded by young actors or volunteers—people who move in a different way than the elderly. Thus, the kinematic descriptors extracted, which define falls of young people, may not properly describe falls of elderly persons, posing a generalization problem difficult to solve, given the total absence of real data.

The method proposed in Chapter Seven approaches fall detection from the perspective of human balance, extracting features from images to assess whether a person maintains equilibrium. This way, the problem explained becomes irrelevant, as all falls are a consequence of a failure in the continuous effort of the body to keep balance, regardless of any other consideration.

The dynamic descriptors proposed in this thesis are used to determine stability indexes, which, in turn, will be used for fall detection. A temporal-convolutional neural network called CoGNet, which takes as input the temporal series of body joint positions and can regress the COG (center of gravity) and feet position projected onto the ground plane, provides the descriptors. This output is used to determine the mentioned stability indexes, making it possible to assess whether a fall has taken place by using a simple algorithm.

This method is able to accept temporal series coming from any 2D human pose estimation network, including the ones coming from the networks trained with FIR imagery proposed in Chapter 6. The accuracy of this method is very high when the joint coordinates are accurate and complete, and it diminishes as the precision of those coordinates decreases as a consequence of occlusion phenomena.

7.2 Main contributions

The main contribution of this thesis is the development and validation of an automatic vision-based fall detection system able to operate properly under no-light conditions. Additional contributions are a deep review of the state-of-the-art of the field of automatic fall detection systems, the biggest block of interviews and polls regarding this area, the first FIR labelled imagery dataset showing people engaged in different activities, the training of all major human pose estimation neural network architectures using FIR imagery and the implementation and validation of an alternative vision-based fall detection method able to solve the problem of data generalization of this field.

7.2.1 State-of-the-art

This thesis starts with a major revision of the field of automatic fall detection systems. To do so, all papers published from 2015 to 2020 in the main public databases of research documentation regarding this field were reviewed. After the proper screening process, 198 articles were selected, and the systems and methods described in them were studied.

All systems presented by the selected articles fell within one of the three categories considered by the vast majority of taxonomies in this field: wearable, ambient, and vision-based. Additionally, all of them approach the problem of fall detection following a common path. They process signals related to the monitored person's movements and infer descriptors that characterize those movements. Then, these descriptors are classified to determine whether a fall has taken place.

Wearable systems use sensors carried by the monitored person to evaluate their movements. This technology is mature and very accurate, but attaching a sensor to the body is uncomfortable and makes the system battery-dependent.

Ambient systems use immature technologies and restrict the patient's movements to the area where sensors are deployed. However, they present clear advantages, such as unlimited processing power, no battery dependency, and a more comfortable situation for the patient, as no sensor is attached to their body.

Vision-based systems are quite mature and present the same set of advantages and disadvantages than ambient sensors. At the time this part of the thesis was developed no state-of-the-art specifically devoted to the area of vision-based fall detection systems had ever been published, so the results of the study were presented in the article "Comprehensive review of vision-based fall detection systems" [12].

This study of the state-of-the-art is extensively presented in the chapter three of this thesis.

7.2.2 User's needs determination

At the time of developing this part of the thesis, no major study covering users' needs had ever been published. To fill this gap, the most comprehensive set of interviews and polls was conducted in 2021. The final objective of this study was to identify users' needs not covered by the present technology.

Four thematic areas were covered by the interviews and polls: the degree of confidence in the systems, privacy protection, system reliability, and scenarios where these systems are likely to be deployed. Individuals belonging to the elderly community and all groups responsible for taking care of them were interviewed to gain a vision as holistic as possible of the different perceptions these communities have towards automatic fall detection systems.

The conclusions of this study are clear. These systems are quite unknown, and therefore, the degree of confidence in them is low in all communities. However, although direct human supervision is always the preferred option, there are certain circumstances under which the use of these systems is accepted.

These situations are often related to patients who retain a certain degree of independence. One of the common scenarios frequently described in the interviews is the one associated with semi-supervised patients getting up at night. In this situation, the person is often disoriented, and therefore, the likelihood of a fall is higher than in other circumstances. Additionally, in the event of a fall, it would probably go unnoticed until the next day, substantially delaying a potentially needed medical intervention.

This scenario, where the use of automatic fall detection systems is perceived as acceptable, is unsuitable for wearable systems, given the discomfort associated with attaching a sensor to the body during bedtime. Therefore, the use of ambient systems should be the proper choice in this case. However, the low degree of maturity of these systems makes them unfit for this purpose as well.

Under these circumstances, visual-based systems could be the optimal choice. However, all visual-based systems work either with visual or near-infrared spectrum imagery, and therefore, they are also unsuitable to properly work in the non-illuminated environments associated with the described scenario. The development of a system working with FIR images could be the perfect answer to the problem, as it not only works properly in low-light environments but also preserves privacy, a major concern expressed in the interviews and polls.

Chapter five of this work covers this study. The results of the investigation were presented in the XV technologies applied to electronics teaching conference and were published in the article “Fall detection system based on far infrared images” [494]

7.2.3 FIR-Human

Developing any visual-based fall detection system working on FIR imagery requires proper datasets for training. Regrettably, the only public datasets used to train person detection systems in the fields of autonomous driving or security and surveillance do not contain images of people falling or include joint annotations.

Therefore, creating a FIR dataset specifically tailored to the training requirements of a visual-based fall detection system becomes necessary if achieving an operational system working on FIR imagery is the final goal.

The recorded dataset, FIR-Human, is the first of its kind and contains video clips recorded by five volunteers. It has a size similar to the main datasets used for training human pose estimation systems working on conventional imagery. The FIR-Human clips are recorded at 23.98 frames per second, and each frame has a resolution of 480 x 640 pixels. The joint information consists of 3-dimensional positions of 19 major body joints, defined with an error of less than 5 millimeters for each recorded frame. Additionally, the 2-dimensional projection of those coordinates onto the recording plane is also provided.

The dataset comprises 27 activity classes, 26 of which are daily life activities, while the other one includes different types of falls. The volunteers perform different actions from various perspectives to ensure a rich and diverse dataset.

For a more extensive description of this dataset, please refer to chapter six.

7.2.4 Training of human pose estimation neural networks on FIR imagery

The development of an automatic fall detection system based on the use of CoGNet requires a consistent input of a time series of matrices containing the positions of the most significant joints of the individual present in the image.

This series of position matrices in the two-dimensional space of the image is generated through human pose estimation neural networks. This work identifies the essential architectures capable of providing those joint coordinates.

As stated in chapter six of this thesis, there are two essential architecture blocks capable of approaching this problem; convolutional architectures and transformers. The first ones can identify elements in an image and establish relationships between these elements and those in

their vecinity. Transformers, on the other hand, are architectures capable of focusing the network's attention on specific elements of the image without fully decomposing all the elements present, as in the previous case.

This attention-oriented feature of transformers allows for the identification of specific elements in the image at a much lower processing cost than convolutional networks. However, convolutional architectures require a much smaller volume of data during their training phase compared to transformers. The accuracy in identifying joints is similar for both architectures, although the reduced computational cost for transformers compared to convolutional structures allows input images to have higher resolution for the same processing cost.

In practice, this means that transformer-based architectures offer slightly better precision for equal processing requirements.

On the other hand, there are two essential philosophies to approach the problem of two-dimensional joints position determination. The first philosophy is known as direct regression, aiming to directly extract the coordinates of the studied joints from the image. The second philosophy, more complex, generates a heat map for each joint from the input image. This heat map is a probabilistic representation of the position of each joint.

Direct regression has shown lower accuracies in determining joints compared to the heat map technique. However, as in the previous case, generating heat maps requires substantially higher computational costs than direct regression.

The training conducted with all the identified architectures using both approaches (direct regression and heat maps) on the FIR-Human dataset yields results very similar to those obtained with visible spectrum images, which can be found in the numerous papers presenting each of the architectures.

Just like in the case of networks trained on visible spectrum images, training on labeled FIR images demonstrates, for the same computational costs, the superiority of architectures based on Transformers. It also proves that the accuracy of the heat map techniques is substantially better than that of direct regression. Additionally, it becomes clear that the precision in positioning the head is nearly identical for all architectures, while it significantly degrades as the joint moves away from the head. This way, only the more capable architectures can provide accurate coordinates for feet, ankles, or knees.

7.2.5 Dynamic descriptors

All systems able to detect automatically falls infer descriptors that characterize human movement. For the particular case of visual-based, systems these descriptors are extracted from images or video clips and are used to determine whether a fall has taken place.

The boundaries used to classify descriptors associated to actions are determined during system training. There are no publicly available datasets containing video clips of real falls of elderly people, as all datasets identified for this purpose in chapter four are recorded by young actors or volunteers. However, given the differences in the way elderly and young people move and fall, there is evidence to affirm that the generalization of the boundaries established to detect falls of young people may be inadequate for fall determination of elderly individuals.

This way, solving this generalization problem requires an alternative approach as the one we propose in this thesis. We approach fall detection from the perspective of human balance and extract features from images in order to assess whether a person maintains equilibrium. This way, by extracting dynamic features from images, the problem caused by the differences in the way young and elderly move becomes irrelevant, as all falls are always a consequence

of a fail in the continuous effort of the body to keep balance, regardless of any other consideration.

We implement this alternative approach by using CoGNet, a temporal convolutional architecture able to regress the COG and feet joints positions projected onto the ground plane. Additionally, the network also determines the ground contact status of the projected feet joints giving, thus, all the information needed to assess human stability from a dynamic perspective. It uses a sequence of 2D joint positions along time as input and is light enough to be run in real time.

This way, using the information provided by CoGNet, simple algorithms could determine whether a fall has taken place. Additionally, the output of the net can also be used to determine different estability indexes in order to evaluate gaits and tell which ones have a higher probability of falling, allowing this way interventions to improve walking and prevent falls.

An extensive description of this generalization problem and our alternative approach to it is presented in chapter seven and was published in the article “human stability assessment and fall detection based on dynamic descriptors” [495]

7.3 Future work

During the development of this thesis, four main areas that could be the subject of future research have been identified.

The first area is related to more in depth research works, which allow a better understanding of the real needs of the community of dependents and all the other ones around it.

This would enable developers to direct their efforts towards solutions that are better suited to those needs, ultimately increasing user confidence in the developed systems. Furthermore, a higher adoption and usage of these systems would help gather more real-world data, improving the training and effectiveness of such systems.

The research conducted in this field by this thesis is focused on nursing homes, where a very specific need has been identified that could not be addressed with the current technology of these systems. However, it is highly likely that research centered on semi-dependent communities, who still live in their homes with support from family and caregivers, could reveal another set of needs. Lastly, studying other types of communities with even lower levels of dependence, yet maintaining nearly full independence, could uncover additional needs that are not currently addressed by existing technology.

The second area of potential future research is related to the use of FIR-Human for training conventional fall detection systems. These systems, as we have extensively described, extract kinematic descriptors associated with people's movements to classify and determine if a fall has occurred. The establishment of the characteristics that allow differentiation between falls and other events is achieved during the system's training phase.

The most reasonable strategy to approach this training, in the case of neural networks, would be transfer learning, where the initial training is conventionally performed using visible spectrum images. Then, the training of the network continues using FIR-Human data, both for this final stage of training and for the system validation phase.

The third axis of future research identified is the use of CoGNet in the areas of human balance and gait analysis. These two fields, which are currently studied using computerized dynamic posturography, could greatly benefit from the utilization of networks like CoGNet. This is because CoGNet is capable of providing the same data as posturographs but using

extremely inexpensive equipment, such as monocular cameras, as opposed to the exorbitant cost of acquiring a posturograph.

The area of balance and gait analysis is essential in preventive medicine for the elderly, as it allows the identification of gait patterns that have a higher probability of falling. This, in turn, enables the implementation of physiotherapy programs capable of improving gait style and thus reducing the likelihood of falling. This is crucial for the elderly community, as falls often have severe consequences ranging from loss of mobility and independence to fatalities, as already seen.

Finally, the last identified area of future research is the development of ambient systems and in particular, mixed systems, systems that combine different technologies in order to determine, with the highest possible accuracy, the occurrence of a fall.

Ambient systems, unlike wearable ones, rely on less mature technologies, as evidenced by the lower number of published articles addressing this type of systems compared to the more mature technologies associated to wearable or vision-based systems. The development of these promising technologies could enable the creation of systems that are currently not feasible, improving the accuracy and cost-effectiveness of the existing ones.

Furthermore, the combination of ambient technologies alongside the more mature ones in the form of mixed systems allows better performances, as the weaknesses of a specific technology are addressed by others. This leads to significantly more accurate systems that deserve further investigation.

8 Bibliography

[1] U. Nations, "World Population Ageing 2017: Highlights," New York: Department of Economic and Social Affairs, United Nations, 2017.

[2] D. A. Sterling, J. A. O'connor and J. Bonadies, "Geriatric falls: injury severity is high and disproportionate to mechanism," Journal of Trauma and Acute Care Surgery, vol. 50, (1), pp. 116-119, 2001.

[3] L. Ren and Y. Peng, "Research of Fall Detection and Fall Prevention Technologies: A Systematic Review," IEEE Access, vol. 7, pp. 77702-77722, 2019. . DOI: 10.1109/ACCESS.2019.2922708.

[4] R. Leirós-Rodríguez, J. L. García-Soidán and V. Romo-Pérez, "Analyzing the Use of Accelerometers as a Method of Early Diagnosis of Alterations in Balance in Elderly People: A Systematic Review," Sensors, vol. 19, (18), 2019. . DOI: 10.3390/s19183883.

[5] A. Ramachandran and A. Karuppiah, "A Survey on Recent Advances in Wearable Fall Detection Systems," BioMed Research International, vol. 2020, pp. 2167160, 2020. DOI: 10.1155/2020/2167160.

[6] H. Sadreazami, M. Bolic and S. Rajan, "Fall Detection Using Standoff Radar-Based Sensing and Deep Convolutional Neural Network," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 67, (1), pp. 197-201, 2020. . DOI: 10.1109/TCSII.2019.2904498.

[7] H. Sadreazami, M. Bolic and S. Rajan, "CapsFall: Fall Detection Using Ultra-Wideband Radar and Capsule Network," IEEE Access, vol. 7, pp. 55336-55343, 2019. . DOI: 10.1109/ACCESS.2019.2907925.

[8] A. Bhattacharya and R. Vaughan, "Deep Learning Radar Design for Breathing and Fall Detection," IEEE Sensors Journal, vol. 20, (9), pp. 5072-5085, 2020. . DOI: 10.1109/JSEN.2020.2967100.

[9] Y. Hu et al, "A wifi-based passive fall detection system," in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, .

[10] M. Keaton et al, "WiFi-based in-home fall-detection utility: Application of WiFi channel state information as a fall detection service," in 2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), 2020, .

[11] X. Wang, J. Ellul and G. Azzopardi, "Elderly Fall Detection Systems: A Literature Survey," Frontiers in Robotics and AI, vol. 7, 2020. Available: https://www.frontiersin.org/article/10.3389/frobt.2020.00071. Access 5/4/21.

[12] J. Gutiérrez, V. Rodríguez and S. Martin, "Comprehensive review of vision-based fall detection systems," Sensors, vol. 21, (3), pp. 947, 2021.

[13] M. Kangas et al, "Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects," Gait Posture, vol. 35, (3), pp. 500-505, 2012.

[14] J. Klenk et al, "Comparison of acceleration signals of simulated and real-world backward falls," Med. Eng. Phys., vol. 33, (3), pp. 368-373, 2011.

[15] F. J. S. Thilo et al, "Usability of a wearable fall detection prototype from the perspective of older people-A real field testing approach," J. Clin. Nurs., vol. 28, (1-2), pp. 310-320, 2019. DOI: 10.1111/jocn.14599.

[16] G. Demiris, S. Chaudhuri and H. J. Thompson, "Older Adults' Experience with a Novel Fall Detection Device," Telemed. J. E. Health., vol. 22, (9), pp. 726-732, 201. DOI: 10.1089/tmj.2015.0218.

[17] L. Kau and C. Chen, "A Smart Phone-Based Pocket Fall Accident Detection, Positioning, and Rescue System," IEEE Journal of Biomedical and Health Informatics, vol. 19, (1), pp. 44-56, 2015. DOI: 10.1109/JBHI.2014.2328593.

[18] P. Pierleoni et al, "A High Reliability Wearable Device for Elderly Fall Detection," IEEE Sensors Journal, vol. 15, (8), pp. 4544-4553, 2015. . DOI: 10.1109/JSEN.2015.2423562.

[19] P. Kostopoulos et al, Increased Fall Detection Accuracy in an Accelerometer-Based Algorithm Considering Residual Movement. 20152. DOI: 10.5220/0005179100300036.

[20] L. Palmerini et al, "A Wavelet-Based Approach to Fall Detection," Sensors (Basel, Switzerland), vol. 15, pp. 11575-86, 2015. . DOI: 10.3390/s150511575.

[21] A. Kurniawan, A. R. Hermawan and I. K. E. Purnama, "A wearable device for fall detection elderly people using tri dimensional accelerometer," in 2016 International Seminar on Intelligent Technology and its Applications (ISITIA), 2016, . DOI: 10.1109/ISITIA.2016.7828740.

[22] C. Wang et al, "Low-Power Fall Detector Using Triaxial Accelerometry and Barometric Pressure Sensing," IEEE Transactions on Industrial Informatics, vol. 12, (6), pp. 2302-2311, 2016. . DOI: 10.1109/TII.2016.2587761.

[23] T. N. Gia et al, "IoT-based fall detection system with energy efficient sensor nodes," in 2016 IEEE Nordic Circuits and Systems Conference (NORCAS), 2016, DOI: 10.1109/NORCHIP.2016.7792890.

[24] A. K. Bourke et al, "Fall detection algorithms for real-world falls harvested from lumbar sensors in the elderly population: A machine learning approach," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016. DOI: 10.1109/EMBC.2016.7591534.

[25] S. Abdelhedi et al, "Development of a two-threshold-based fall detection algorithm for elderly health monitoring," in 2016 IEEE Tenth International Conference on Research Challenges in Information Science (RCIS), 2016 . DOI: 10.1109/RCIS.2016.7549315.

[26] N. Otanasap, "Pre-impact fall detection based on wearable device using dynamic threshold model," in 2016 17th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2016 . DOI: 10.1109/PDCAT.2016.083.

[27] J. He, C. Hu and X. Wang, "A Smart Device Enabled System for Autonomous Fall Detection and Alert," International Journal of Distributed Sensor Networks, vol. 2016, pp. 1-10, 2016. DOI: 10.1155/2016/2308183.

[28] A. Sucerquia, J. D. López and F. Vargas, "Two-threshold energy based fall detection using a triaxial accelerometer," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016. DOI: 10.1109/EMBC.2016.7591385.

[29] Álvarez de la Concepción, Miguel Ángel et al, "Mobile activity recognition and fall detection system for elderly people using Ameva algorithm," Pervasive and Mobile Computing, vol. 34, pp. 3-13, 2017. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.pmcj.2016.05.002. Access 5/4/21.

[30] P. Jatesiktat and W. T. Ang, "An elderly fall detection using a wrist-worn accelerometer and barometer," in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017, . DOI: 10.1109/EMBC.2017.8036778.
[31] N. Pannurat, S. Thiemjarus and E. Nantajeewarawat, "A Hybrid Temporal Reasoning Framework for Fall Monitoring," IEEE Sensors Journal, vol. 17, (6), pp. 1749-1759, 2017. . DOI: 10.1109/JSEN.2017.2649542.
[32] Putra IPES, Brusey J, Gaura E,Vesilo R., "An Event-Triggered Machine Learning Approach for Accelerometer-Based Fall Detection." Sensors, Vol 18, pp. 20, 2017.
[33] C. Medrano et al, "Combining novelty detectors to improve accelerometer-based fall detection," Med. Biol. Eng. Comput., vol. 55, 2017. . DOI: 10.1007/s11517-017-1632-z.
[34] R. Shen et al, "A Novel Fall Prediction System on Smartphones," IEEE Sensors Journal, vol. 17, (6), pp. 1865-1871, 2017. . DOI: 10.1109/JSEN.2016.2598524.
[35] D. Yacchirema et al, "Fall detection system for elderly people using IoT and Big Data," Procedia Computer Science, vol. 130, pp. 603-610, 2018. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.procs.2018.04.110. Access 7/6/22.
[36] B. Kaudki and A. Surve, "IOT enabled human fall detection using accelerometer and RFID technology," in 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), 2018, . DOI: 10.1109/ICCONS.2018.8663092.
[37] A. Sucerquia, J. D. López and J. Vargas-Bonilla, "Real-Life/Real-Time Elderly Fall Detection with a Triaxial Accelerometer," Sensors, vol. 18, (4), 2018. . DOI: 10.3390/s18041101.
[38] W. Saadeh, S. A. Butt and M. A. B. Altaf, "A Patient-Specific Single Sensor IoT-Based Wearable Fall Prediction and Detection System," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 27, (5), pp. 995-1003, 2019. . DOI: 10.1109/TNSRE.2019.2911602.
[39] L. Chen et al, "Intelligent fall detection method based on accelerometer data from a wrist-worn smart watch," Measurement, vol. 140, pp. 215-226, 2019. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.measurement.2019.03.079. Access 7/7/21.
[40] A. Shahzad and K. Kim, "FallDroid: An Automated Smart-Phone-Based Fall Detection System Using Multiple Kernel Learning," IEEE Transactions on Industrial Informatics, vol. 15, (1), pp. 35-44, 2019. . DOI: 10.1109/TII.2018.2839749.
[41] H. W. Guo et al, "A threshold-based algorithm of fall detection using a wearable device with tri-axial accelerometer and gyroscope," in 2015 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), 2015, . DOI: 10.1109/ICIIBMS.2015.7439470.
[42] H. Jian and H. Chen, "A portable fall detection and alerting system based on k-NN algorithm and remote medicine," China Communications, vol. 12, (4), pp. 23-31, 2015. . DOI: 10.1109/CC.2015.7114066.
[43] M. I. Nari et al, "A simple design of wearable device for fall detection with accelerometer and gyroscope," in 2016 International Symposium on Electronics and Smart Devices (ISESD), 2016, . DOI: 10.1109/ISESD.2016.7886698.
[44] T. Sivaranjani et al, "Fall assessment and its injury prevention using a wearable airbag technology," in 2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), 2017. DOI: 10.1109/ICPCSI.2017.8392175.
[45] A. Jefiza et al, "Fall detection based on accelerometer and gyroscope using back propagation," in 2017 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017. DOI: 10.1109/EECSI.2017.8239149.
[46] Y. Su, D. Liu and Y. Wu, "A multi-sensor based pre-impact fall detection system with a hierarchical classifier," in 2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2016. DOI: 10.1109/CISP-BMEI.2016.7852995.
[47] A. M. Sabatini et al, "Prior-to- and Post-Impact Fall Detection Using Inertial and Barometric Altimeter Measurements," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 24, (7), pp. 774-783, 2016. . DOI: 10.1109/TNSRE.2015.2460373.
[48] W. Lu et al, "Smart Triggering of the Barometer in a Fall Detector Using a Semi-Permeable Membrane," IEEE Transactions on Biomedical Engineering, vol. 67, (1), pp. 146-157, 2020. DOI: 10.1109/TBME.2019.2909907.
[49] J. Light, S. Cha and M. Chowdhury, "Optimizing pressure sensor array data for a smart-shoe fall monitoring system," in 2015 Ieee Sensors, 2015. DOI: 10.1109/ICSENS.2015.7370271.
[50] W. Lu et al, "Low-power operation of a barometric pressure sensor for use in an automatic fall detector," in 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016. DOI: 10.1109/EMBC.2016.7591120.
[51] J. Sun et al, "Fall detection using plantar inclinometer sensor," in 2015 IEEE 12th Intl Conf on Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and its Associated Workshops (UIC-ATC-ScalCom), 2015. DOI: 10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.308.
[52] J. Sun et al, "A plantar inclinometer based approach to fall detection in open environments," in Anonymous 2016. DOI: 10.1007/978-3-319-33353-3_1.
[53] M. Cheffena, "Fall Detection Using Smartphone Audio Features," IEEE Journal of Biomedical and Health Informatics, vol. 20, (4), pp. 1073-1080, 2016. . DOI: 10.1109/JBHI.2015.2425932.
[54] A. Leone et al, "An EMG-based system for pre-impact fall detection," in 2015 Ieee Sensors, 2015. DOI: 10.1109/ICSENS.2015.7370314.
[55] A. Leone et al, "A Wearable EMG-based System Pre-fall Detector," Procedia Engineering, vol. 120, pp. 455-458, 2015. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.proeng.2015.08.667. Access 6/7/22.
[56] G. Rescio et al, "A preliminary study on fall risk evaluation through electromiography systems," in 2015 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL), 2015 . DOI: 10.1109/IMCTL.2015.7359590.
[57] V. F. Annese et al, "A digital processor architecture for combined EEG/EMG falling risk prediction," in 2016 Design, Automation & Test in Europe Conference & Exhibition (DATE), 2016, .

[58] X. xi et al, "Evaluation of Feature Extraction and Recognition for Activity Monitoring and Fall Detection Based on Wearable sEMG Sensors," Sensors, vol. 17, pp. 1229, 2017. DOI: 10.3390/s17061229.
[59] A. Leone, G. Rescio and P. Siciliano, "Fall risk evaluation by surface electromyography technology," in 2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC), 2017. DOI: 10.1109/ICE.2017.8280003.
[60] J. Xiao et al, "A surface electromyography-based pre-impact fall detection method," in 2018 Chinese Automation Congress (CAC), 2018. DOI: 10.1109/CAC.2018.8623336.
[61] G. Rescio, A. Leone and P. Siciliano, "Supervised machine learning scheme for electromyography-based pre-fall detection system," Expert Syst. Appl., vol. 100, pp. 95-105, 2018. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.eswa.2018.01.047. Access 7/6/22.
[62] G. Mezzina et al, "EEG/EMG based architecture for the early detection of slip-induced lack of balance," in 2019 IEEE 8th International Workshop on Advances in Sensors and Interfaces (IWASI), 2019 . DOI: 10.1109/IWASI.2019.8791252.
[63] A. Nadeem, A. Mehmood and K. Rizwan, "A dataset build using wearable inertial measurement and ECG sensors for activity recognition, fall detection and basic heart anomaly detection system," Data in Brief, vol. 27, pp. 104717, 2019. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.dib.2019.104717. Access 5/5/22.
[64] A. Shahzad et al, "Quantitative Assessment of Balance Impairment for Fall-Risk Estimation Using Wearable Triaxial Accelerometer," IEEE Sensors Journal, vol. 17, (20), pp. 6743-6751, 2017 . DOI: 10.1109/JSEN.2017.2749446.
[65] M. Hemmatpour et al, "Polynomial classification model for real-time fall prediction system," in 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017. DOI: 10.1109/COMPSAC.2017.189.
[66] M. Di Rosa et al, "Concurrent validation of an index to estimate fall risk in community dwelling seniors through a wireless sensor insole system: A pilot study," Gait Posture, vol. 55, pp. 6-11, 2017. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.gaitpost.2017.03.037. Access 7/8/22.
[67] K. Chaccour et al, "Sway analysis and fall prediction method based on spatio-temporal sliding window technique," in 2016 IEEE 18th International Conference on E-Health Networking, Applications and Services (Healthcom), 2016, . DOI: 10.1109/HealthCom.2016.7749488.
[68] M. A. Brodie et al, "Eight-Week Remote Monitoring Using a Freely Worn Device Reveals Unstable Gait Patterns in Older Fallers," IEEE Transactions on Biomedical Engineering, vol. 62, (11), pp. 2588-2594, 2015. . DOI: 10.1109/TBME.2015.2433935.
[69] R. Igual, C. Medrano and I. Plaza, "A comparison of public datasets for acceleration-based fall detection," Med. Eng. Phys., vol. 37, (9), pp. 870-878, 2015.. DOI: https://doi.org/10.1016/j.medengphy.2015.06.009. Access 7/2/22.
[70] M. Manikandan et al, "Unified framework for triaxial accelerometer-based fall event detection and classification using cumulants and hierarchical decision tree classifier," Healthcare Technology Letters, vol. 2, pp. 101-107, 2015. . DOI: 10.1049/htl.2015.0018.
[71] A. Lisowska et al, "An evaluation of supervised, novelty-based and hybrid approaches to fall detection using silmee accelerometer data," in 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), 2015. DOI: 10.1109/ICCVW.2015.60.
[72] M. Vidigal, M. Lima and A. De Almeida Neto, "Elder falls detection based on artificial neural networks," in 2015 Fourteenth Mexican International Conference on Artificial Intelligence (MICAI), 2015. DOI: 10.1109/MICAI.2015.41.
[73] R. M. Gibson et al, "Multiple comparator classifier framework for accelerometer-based fall detection and diagnostic," Applied Soft Computing, vol. 39, pp. 94-103, 2016. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.asoc.2015.10.062. Access 7/3/22.
[74] C. Medrano et al, "The Effect of Personalization on Smartphone-Based Fall Detectors." Sensors 16, pp. 11, 2016. Available: https://www.mdpi.com/1424-8220/16/1/117#cite. Access 7/3/22.
[75] A. T. Özdemir, "An Analysis on Sensor Locations of the Human Body for Wearable Fall Detection Devices: Principles and Practice." Sensors, 16., pp. 1161, 2016. Available: https://www.mdpi.com/1424-8220/16/8/1161#cite, Access 7/9/22.
[76] P. Vallabh et al, "Fall detection using machine learning algorithms," in 2016 24th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), 2016, . DOI: 10.1109/SOFTCOM.2016.7772142.
[77] M. Ahmed et al, "Fall Detection System for the Elderly Based on the Classification of Shimmer Sensor Prototype Data," Healthc Inform Res, vol. 23, (3), pp. 147-158, 2017. DOI: 10.4258/hir.2017.23.3.147.
[78] B. Ando et al, "A NeuroFuzzy approach for fall detection," in 2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC), 2017, . DOI: 10.1109/ICE.2017.8280032.
[79] O. Aziz et al, " Validation of accuracy of SVM-based fall detection system using real-world fall and non-fall datasets." PLoS ONE 12 (7), 2017. Available: https://doi.org/10.1371/journal.pone.0180318. Access 17/4/22.
[80] A. Hakim et al, "Smartphone Based Data Mining for Fall Detection: Analysis and Design," Procedia Computer Science, vol. 105, pp. 46-51, 2017. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.procs.2017.01.188. Access 7/11/22.
[81] A. Jahanjoo, M. N. Tahan and M. J. Rashti, "Accurate fall detection using 3-axis accelerometer sensor and MLF algorithm," in 2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA), 2017, . DOI: 10.1109/PRIA.2017.7983024.
[82] C. Vincenzo et al, A Smartphone-Based System for Detecting Falls using Anomaly Detection. 2017. DOI: 10.1007/978-3-319-68548-9_45.
[83] T. Xie et al, "A multistage collaborative filtering method for fall detection," in 2017 International Joint Conference on Neural Networks (IJCNN), 2017. DOI: 10.1109/IJCNN.2017.7966277.
[84] S. B. Khojasteh et al, "Improving Fall Detection Using an On-Wrist Wearable Accelerometer." Sensors, 18, pp. 1350, 2018. Available: https://www.mdpi.com/1424-8220/18/5/1350#cite. Access 27/8/22.
[85] A. Lisowska and A. a. P. O'Neil I., "Cross-cohort Evaluation of Machine Learning Approaches to Fall Detection from Accelerometer Data." Proceedings of the 11th International Joint Conference on Biomedical Engineering Systems and Technologies - Vol. 5, pp. 77-82, 2018. Available: https://www.scitepress.org/Link.aspx?doi=10.5220/0006554400770082. Access 7/4/22.
[86] T. R. Mauldin et al, "SmartFall: A Smartwatch-Based Fall Detection System Using Deep Learning." Sensors, 18, pp. 3363, 2018. Available: https://www.mdpi.com/1424-8220/18/10/3363. Access 10/2/22.
[87] M. Musci et al, "Online Fall Detection using Recurrent Neural Networks," 2018.
[88] L. Nguyen, M. Saleh and R. Jeannès, "An efficient design of a machine learning-based elderly fall detector," in Anonymous 2018. DOI: 10.1007/978-3-319-76213-5_5.

[89] J. Santoyo Ramón, E. Casilari-Pérez and J. Cano-García, "Analysis of a Smartphone-Based Architecture with Multiple Mobility Sensors for Fall Detection with Supervised Learning," Sensors, vol. 18, pp. 1155, 2018. . DOI: 10.3390/s18041155.
[90] A. Chelli and M. Pätzold, "A Machine Learning Approach for Fall Detection and Daily Living Activity Recognition," IEEE Access, vol. 7, pp. 38670-38687, 2019. . DOI: 10.1109/ACCESS.2019.2906693.
[91] D. Yacchirema et al, "Fall detection system for elderly people using IoT and ensemble machine learning algorithm," Personal and Ubiquitous Computing, vol. 23, 2019. . DOI: 10.1007/s00779-018-01196-8.
[92] M. Salman Khan et al, "An unsupervised acoustic fall detection system using source separation for sound interference suppression," Signal Process, vol. 110, pp. 199-210, 2015. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.sigpro.2014.08.021. Access 10/8/22.
[93] A. Díaz-Ramírez, E. Domínguez and L. Martínez-Alvarado, "A falls detection system for the elderly based on a WSN," in 2015 IEEE International Symposium on Technology and Society (ISTAS), 2015, . DOI: 10.1109/ISTAS.2015.7439426.
[94] E. Principi et al, "Acoustic cues from the floor: A new approach for fall classification," Expert Syst. Appl., vol. 60, pp. 51-61, 2016. DOI: https://doi.org/10.1016/j.eswa.2016.04.007. Access 7/2/22.
[95] D. Droghini et al, "A Combined One-Class SVM and Template-Matching Approach for User-Aided Human Fall Detection by Means of Floor Acoustic Features," Computational Intelligence and Neuroscience, vol. 2017, pp. 1512670, 2017. DOI: 10.1155/2017/1512670.
[96] A. Irtaza et al, "A framework for fall detection of elderly people by analyzing environmental sounds through acoustic local ternary patterns," in 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2017. DOI: 10.1109/SMC.2017.8122836.
[97] S. M. Adnan et al, "Fall detection through acoustic Local Ternary Patterns," Appl. Acoust., vol. 140, pp. 296-300, 2018. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.apacoust.2018.06.013. Access 10/8/22.
[98] D. Droghini et al, "Audio Metric Learning by Using Siamese Autoencoders for One-Shot Human Fall Detection," IEEE Transactions on Emerging Topics in Computational Intelligence, pp. 1-11, 2019. . DOI: 10.1109/TETCI.2019.2948151.
[99] K. Chaccour et al, "Smart carpet using differential piezoresistive pressure sensors for elderly fall detection," in 2015 IEEE 11th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), 2015. DOI: 10.1109/WiMOB.2015.7347965.
[100] S. Jeon et al, "Self-powered fall detection system using pressure sensing triboelectric nanogenerators," Nano Energy, vol. 41, pp. 139-147, 2017. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.nanoen.2017.09.028. Access 10/8/22.
[101] M. Daher et al, "Elder Tracking and Fall Detection System Using Smart Tiles," IEEE Sensors Journal, vol. 17, (2), pp. 469-479, 2017. . DOI: 10.1109/JSEN.2016.2625099.
[102] J. Haffner et al, "A smart capacitive measurement system for fall detection," J. Electrostatics, vol. 92, pp. 45-53, 2018. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.elstat.2018.01.005. Access 12/3/22.
[103] Wei-Han Chen and Hsi-Pin Ma, "A fall detection system based on infrared array sensors with tracking capability for the elderly at home," in 2015 17th International Conference on E-Health Networking, Application & Services (HealthCom), 2015, . DOI: 10.1109/HealthCom.2015.7454538.
[104] Q. Guan et al, "Infrared signal based elderly fall detection for in-home monitoring," in 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2017. DOI: 10.1109/IHMSC.2017.91.
[105] X. Fan et al, "Robust unobtrusive fall detection using infrared array sensors," in 2017 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), 2017. DOI: 10.1109/MFI.2017.8170428.
[106] A. Hayashida, V. Moshnyaga and K. Hashimoto, "New approach for indoor fall detection by infrared thermal array sensor," in 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), 2017. DOI: 10.1109/MWSCAS.2017.8053196.
[107] J. Adolf et al, "Deep neural network based body posture recognitions and fall detection from low resolution infrared array sensor," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018. DOI: 10.1109/BIBM.2018.8621582.
[108] Y. Ogawa and K. Naito, "Fall detection scheme based on temperature distribution with IR array sensor," in 2020 IEEE International Conference on Consumer Electronics (ICCE), 2020. DOI: 10.1109/ICCE46568.2020.9043000.
[109] Z. Liu et al, "Fall Detection and Personnel Tracking System Using Infrared Array Sensors," IEEE Sensors Journal, vol. 20, (16), pp. 9558-9566, 2020. DOI: 10.1109/JSEN.2020.2988070.
[110] B. Y. Su et al, "Doppler Radar Fall Activity Detection Using the Wavelet Transform," IEEE Transactions on Biomedical Engineering, vol. 62, (3), pp. 865-875, 2015. DOI: 10.1109/TBME.2014.2367038.
[111] C. Garripoli et al, "Embedded DSP-Based Telehealth Radar System for Remote In-Door Fall Detection," IEEE Journal of Biomedical and Health Informatics, vol. 19, (1), pp. 92-101, 2015. DOI: 10.1109/JBHI.2014.2361252.
[112] B. Erol, M. G. Amin and B. Boashash, "Range-doppler radar sensor fusion for fall detection," in 2017 IEEE Radar Conference (RadarConf), 2017. DOI: 10.1109/RADAR.2017.7944316.
[113] H. Sadreazami, M. Bolic and S. Rajan, "On the use of ultra wideband radar and stacked LSTM-RNN for at home fall detection," in 2018 IEEE Life Sciences Conference (LSC), 2018. DOI: 10.1109/LSC.2018.8572048.
[114] S. Chen et al, "Low PRF low frequency radar sensor for fall detection by using deep learning," in 2019 IEEE 4th International Conference on Signal and Image Processing (ICSIP), 2019. DOI: 10.1109/SIPROCESS.2019.8868580.
[115] Y. Sun et al, "Privacy-preserving fall detection with deep learning on mmWave radar signal," in 2019 IEEE Visual Communications and Image Processing (VCIP), 2019. DOI: 10.1109/VCIP47243.2019.8965661.
[116] H. Sadreazami, M. Bolic and S. Rajan, "CapsFall: Fall Detection Using Ultra-Wideband Radar and Capsule Network," IEEE Access, vol. 7, pp. 55336-55343, 2019. . DOI: 10.1109/ACCESS.2019.2907925.
[117] Y. Shankar, S. Hazra and A. Santra, "Radar-based non-intrusive fall motion recognition using deformable convolutional neural network," in 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), 2019. DOI: 10.1109/ICMLA.2019.00279.
[118] A. Chelli and M. Pätzold, "A Machine Learning Approach for Fall Detection Based on the Instantaneous Doppler Frequency," IEEE Access, vol. 7, pp. 166173-166189, 2019. DOI: 10.1109/ACCESS.2019.2947739.
[119] H. Sadreazami, M. Bolic and S. Rajan, "Residual network-based supervised learning of remotely sensed fall incidents using ultra-wideband radar," in 2019 IEEE International Symposium on Circuits and Systems (ISCAS), 2019. DOI: 10.1109/ISCAS.2019.8702446.

[120] C. Ding et al, "Fall detection with multi-domain features by a portable FMCW radar," in 2019 IEEE MTT-S International Wireless Symposium (IWS), 2019. DOI: 10.1109/IEEE-IWS.2019.8804036.
[121] H. Sadreazami, M. Bolic and S. Rajan, "TL-FALL: Contactless indoor fall detection using transfer learning from a pretrained model," in 2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA), 2019. DOI: 10.1109/MeMeA.2019.8802154.
[122] H. Sadreazami et al, "Compressed domain contactless fall incident detection using UWB radar signals," in 2020 18th IEEE International New Circuits and Systems Conference (NEWCAS), 2020. DOI: 10.1109/NEWCAS49341.2020.9159760.
[123] A. Bhattacharya and R. Vaughan, "Deep Learning Radar Design for Breathing and Fall Detection," IEEE Sensors Journal, vol. 20, (9), pp. 5072-5085, 2020. DOI: 10.1109/JSEN.2020.2967100.
[124] H. Sadreazami, M. Bolic and S. Rajan, "Fall Detection Using Standoff Radar-Based Sensing and Deep Convolutional Neural Network," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 67, (1), pp. 197-201, 2020. DOI: 10.1109/TCSII.2019.2904498.
[125] H. Wang et al, "RT-Fall: A Real-Time and Contactless Fall Detection System with Commodity WiFi Devices," IEEE Transactions on Mobile Computing, vol. 16, (2), pp. 511-526, 2017. DOI: 10.1109/TMC.2016.2557795.
[126] H. Cheng et al, "Deep learning wi-fi channel state information for fall detection," in 2019 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-TW), 2019. DOI: 10.1109/ICCE-TW46550.2019.8991919.
[127] M. Huang et al, "Your WiFi knows you fall: A channel data-driven device-free fall sensing system," in ICC 2019 - 2019 IEEE International Conference on Communications (ICC), 2019. DOI: 10.1109/ICC.2019.8762032.
[128] Y. Hu et al, "A WiFi-based passive fall detection system," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020. DOI: 10.1109/ICASSP40776.2020.9054753.
[129] M. Keaton et al, "WiFi-based in-home fall-detection utility: Application of WiFi channel state information as a fall detection service," in 2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), 2020. DOI: 10.1109/ICE/ITMC49519.2020.9198407.
[130] T. H. Nguyen and H. H. Nguyen, "Towards a robust WiFi-based fall detection with adversarial data augmentation," in 2020 54th Annual Conference on Information Sciences and Systems (CISS), 2020. DOI: 10.1109/CISS48834.2020.1570617398.
[131] J. Ding and Y. Wang, "A WiFi-based Smart Home Fall Detection System using Recurrent Neural Network," IEEE Transactions on Consumer Electronics, pp. 1, 2020. DOI: 10.1109/TCE.2020.3021398.
[132] A. Collado Villaverde et al, Machine Learning Approach to Detect Falls on Elderly People using Sound. 2017. DOI: 10.1007/978-3-319-60042-0_18.
[133] A. Yajai et al, "Fall detection using directional bounding box," in - 2015 12th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2015. DOI: 10.1109/JCSSE.2015.7219769.
[134] C. -J. Chong et al, "Visual based fall detection with reduced complexity horprasert segmentation using superpixel," in - 2015 IEEE 12th International Conference on Networking, Sensing and Control, 2015. DOI: 10.1109/ICNSC.2015.7116081.
[135] H. Rajabi and M. Nahvi, "An intelligent video surveillance system for fall and anesthesia detection for elderly and patients," in - 2015 2nd International Conference on Pattern Recognition and Image Analysis (IPRIA), 2015. DOI: 10.1109/PRIA.2015.7161644.
[136] L. H. Juang and M. N. Wu, "Fall Down Detection Under Smart Home System," J. Med. Syst., vol. 39, (10), pp. 107-3. Epub 2015 Aug 15, 2015. DOI: 10.1007/s10916-015-0286-3.
[137] M. A. Mousse, C. Motamed and E. C. Ezin, "Video-based people fall detection via homography mapping of foreground polygons from overlapping cameras," in - 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2015. DOI: 10.1109/SITIS.2015.56.
[138] E. Auvinet et al, "Multiple cameras fall data set," 2011.
[139] M. Aslan et al, "Shape feature encoding via Fisher Vector for efficient fall detection in depth-videos," Applied Soft Computing, vol. 37, pp. 1023-1028, 2015. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.asoc.2014.12.035. Access 12/8/22.
[140] X. Ma et al, "Depth-Based Human Fall Detection via Shape Features and Improved Extreme Learning Machine," IEEE Journal of Biomedical and Health Informatics, vol. 18, (6), pp. 1915-1922, 2014. . DOI: 10.1109/JBHI.2014.2304357.
[141] Z. Bian et al, "Fall Detection Based on Body Part Tracking Using a Depth Camera," IEEE Journal of Biomedical and Health Informatics, vol. 19, (2), pp. 430-439, 2015. DOI: 10.1109/JBHI.2014.2319372.
[142] C. Lin et al, "Vision-based fall detection through shape features," in - 2016 IEEE Second International Conference on Multimedia Big Data (BigMM), 2016. DOI: 10.1109/BigMM.2016.22.
[143] F. Merrouche and N. Baha, "Depth camera based fall detection using human shape and movement," in - 2016 IEEE International Conference on Signal and Image Processing (ICSIP), 2016. DOI: 10.1109/SIPROCESS.2016.7888330.
[144] K. G. Gunale and P. Mukherji, "Fall detection using k-nearest neighbor classification for patient monitoring," in - 2015 International Conference on Information Processing (ICIP), 2015. DOI: 10.1109/INFOP.2015.7489439.
[145] K. R. Bhavya et al, "Fall detection using motion estimation and accumulated image map," in - 2016 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), 2016. DOI: 10.1109/ICCE-Asia.2016.7927288.
[146] Kun Wang et al, "Automatic fall detection of human in video using combination of features," in - 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016. DOI: 10.1109/BIBM.2016.7822694.
[147] O. Barnich and M. Van Droogenbroeck, "ViBe: A Universal Background Subtraction Algorithm for Video Sequences," IEEE Transactions on Image Processing, vol. 20, (6), pp. 1709-1724, 2011. DOI: 10.1109/TIP.2010.2101613.
[148] J. Chua, Y. C. Chang and W. K. Lim, "A simple vision-based fall detection technique for indoor video surveillance," Signal, Image and Video Processing, vol. 9, (3), pp. 623-633, 2015. DOI: 10.1007/s11760-013-0493-7.
[149] U. Pratap, M. A. Khan and A. S. Jalai, "Human fall detection for video surveillance by handling partial occlusion scenario," in - 2016 11th International Conference on Industrial and Information Systems (ICIIS), 2016. DOI: 10.1109/ICIINFS.2016.8262951.
[150] X. Wang, H. Liu and M. Liu, "A novel multi-cue integration system for efficient human fall detection," in - 2016 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2016. DOI: 10.1109/ROBIO.2016.7866509.

[151] I. Charfi et al, "Optimized spatio-temporal descriptors for real-time fall detection: Comparison of support vector machine and Adaboost-based classification," Journal of Electronic Imaging, vol. 22, pp. 041106, 2013. DOI: 10.1117/1.JEI.22.4.041106.
[152] A. Y. Alaoui et al, "Video based human fall detection using von mises distribution of motion vectors," in - 2017 Intelligent Systems and Computer Vision (ISCV), 2017, DOI: 10.1109/ISACV.2017.8054942.
[153] I. Charfi et al, "Definition and performance evaluation of a robust SVM based fall detection solution," in - 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, 2012 . DOI: 10.1109/SITIS.2012.155.
[154] A. Yajai and S. Rasmequan, "Adaptive directional bounding box from RGB-D information for improving fall detection," Journal of Visual Communication and Image Representation, vol. 49, pp. 257-273, 2017. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.jvcir.2017.08.008. Access 22/8/22.
[155] B. Lewandowski et al, "I see you lying on the ground — can I help you? fast fallen person detection in 3D with a mobile robot," in - 2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), 2017. DOI: 10.1109/ROMAN.2017.8172283.
[156] F. Harrou et al, "Vision-based fall detection system for improving safety of elderly people," IEEE Instrumentation & Measurement Magazine, vol. 20, (6), pp. 49-55, 2017. DOI: 10.1109/MIM.2017.8121952.
[157] M. Kepski and B. Kwolek, Embedded System for Fall Detection using Body-Worn Accelerometer and Depth Sensor. 2015. DOI: 10.1109/IDAACS.2015.7341404.
[158] K. Adhikari, H. Bouchachia and H. Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network," in - 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA), 2017. DOI: 10.23919/MVA.2017.7986795.
[159] G. M. Basavaraj and A. Kusagur, "Vison-basedsurveillance system for detection of human fall," in - 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017. DOI: 10.1109/RTEICT.2017.8256851.
[160] K. De Miguel et al, "Home Camera-Based Fall Detection System for the Elderly," Sensors, vol. 17, (12), 2017. DOI: 10.3390/s17122864.
[161] L. Yao, W. Min and K. Lu, "A New Approach to Fall Detection Based on the Human Torso Motion Model," Applied Sciences, vol. 7, (10), 2017. DOI: 10.3390/app7100993.
[162] M. Antonello et al, "Fast and robust detection of fallen people from a mobile robot," in - 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017. DOI: 10.1109/IROS.2017.8206276.
[163] (). IASLAB-RGBD Fallen Person Dataset. Available: http://robotics.dei.unipd.it/index.php/spin-off/2-uncategorised/117-fall. Access 20/7/22.
[164] M. N. H. Mohd et al, "An optimized low computational algorithm for human fall detection from depth images based on support vector machine classification," in - 2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA), 2017. DOI: 10.1109/ICSIPA.2017.8120645.
[165] Enea Cippitelli et al. (). TST Fall detection dataset v2. Access 20/8/22.
[166] (). The Falling Detection Datase. Available: http://vlm1.uta.edu/~zhangzhong/fall_detection/ (ac. Access 20/8/22.
[167] N. B. Joshi and S. L. Nalbalwar, "A fall detection and alert system for an elderly using computer vision and internet of things," in - 2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017. DOI: 10.1109/RTEICT.2017.8256804.
[168] N. Otanasap and P. Boonbrahm, "Pre-impact fall detection approach using dynamic threshold based and center of gravity in multiple kinect viewpoints," in - 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2017. DOI: 10.1109/JCSSE.2017.8025955.
[169] Q. Feng et al, "Fall detection based on motion history image and histogram of oriented gradient feature," in 2017 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2017.
[170] S. Hernandez-Mendez et al, "Detecting falling people by autonomous service robots: A ROS module integration approach," in - 2017 International Conference on Electronics, Communications and Computers (CONIELECOMP), 2017. DOI: 10.1109/CONIELECOMP.2017.7891823.
[171] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," Comput. Methods Programs Biomed., vol. 117, (3), pp. 489-501, 2014. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.cmpb.2014.09.005. Access 21/8/22.
[172] S. Kasturi and K. Jo, "Human fall classification system for ceiling-mounted kinect depth images," in - 2017 17th International Conference on Control, Automation and Systems (ICCAS), 2017. DOI: 10.23919/ICCAS.2017.8204202.
[173] S. Kasturi and K. Jo, "Classification of human fall in top viewed kinect depth images using binary support vector machine," in - 2017 10th International Conference on Human System Interactions (HSI), 2017. DOI: 10.1109/HSI.2017.8005016.
[174] S. Pattamaset et al, "Human fall detection by using the body vector," in - 2017 9th International Conference on Knowledge and Smart Technology (KST), 2017. DOI: 10.1109/KST.2017.7886075.
[175] S. Taghvaei, M. H. Jahanandish and K. Kosuge, "Autoregressive-moving-average hidden Markov model for vision-based fall prediction-An application for walker robot," Assist. Technol., vol. 29, (1), pp. 19-27, 2017. DOI: 10.1080/10400435.2016.1174178.
[176] Y. M. Galvão et al, "Anomaly detection in smart houses: Monitoring elderly daily behavior for fall detecting," in - 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2017. DOI: 10.1109/LA-CCI.2017.8285701.
[177] T. Tran et al, "Continuous detection of human fall using multimodal features from Kinect sensors in scalable environment," Comput. Methods Programs Biomed., vol. 146, pp. 151-165, 2017. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.cmpb.2017.05.007. Access 20/7/22.
[178] (). MULTIMODAL MULTIVIEW DATASET OF HUMAN ACTIVITIES. Available: http://mica.edu.vn/perso/Tran-Thi-Thanh-Hai/MFD.html. Access 22/8/22.

[179] X. Li et al, "Fall detection for elderly person care using convolutional neural networks," in - 2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2017. DOI: 10.1109/CISP-BMEI.2017.8302004.
[180] Y. Fan et al, "A deep neural network for real-time detection of falling humans in naturally occurring scenes," Neurocomputing, vol. 260, pp. 43-58, 2017. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.neucom.2017.02.082. Access 20/6/22.
[181] Greet Baldewijs et al, "Bridging the gap between real-life data and simulated data by providing a highly realistic fall dataset for evaluating camera-based fall detection algorithms," Healthcare Technology Letters, 2016.
[182] A. Abobakr, M. Hossny and S. Nahavandi, "A Skeleton-Free Fall Detection System From Depth Images Using Random Decision Forest," IEEE Systems Journal, vol. 12, (3), pp. 2994-3005, 2018. DOI: 10.1109/JSYST.2017.2780260.
[183] (). CMU Graphics Lab - motion capture library. Available: http://mocap.cs.cmu.edu/. Access 16/5/22.
[184] B. Dai et al, "A novel video-surveillance-based algorithm of fall detection," in - 2018 11th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), 2018. DOI: 10.1109/CISP-BMEI.2018.8633160.
[185] G. Mastorakis, T. Ellis and D. Makris, "Fall detection without people: A simulation approach tackling video data scarcity," Expert Syst. Appl., vol. 112, pp. 125-137, 2018. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.eswa.2018.06.019. Access 20/8/22.
[186] K. Sehairi, F. Chouireb and J. Meunier, "Elderly fall detection system based on multiple shape features and motion analysis," in - 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), 2018. DOI: 10.1109/ISACV.2018.8354084.
[187] K. Lu and E. T. -. Chu, "An Image-Based Fall Detection System for the Elderly," Applied Sciences, vol. 8, (10), 2018. . DOI: 10.3390/app8101995.
[188] L. Panahi and V. Ghods, "Human fall detection using machine vision techniques on RGB–D images," Biomedical Signal Processing and Control, vol. 44, pp. 146-153, 2018. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.bspc.2018.04.014. Access 20/2/22.
[189] M. Rahnemoonfar and H. Alkittawi, "Spatio-temporal convolutional neural network for elderly fall detection in depth video cameras," in - 2018 IEEE International Conference on Big Data (Big Data), 2018. DOI: 10.1109/BigData.2018.8622342.
[190] M. Ricciuti, S. Spinsante and E. Gambi, "Accurate Fall Detection in a Top View Privacy Preserving Configuration," Sensors, vol. 18, (6), 2018. DOI: 10.3390/s18061754.
[191] M. Ko et al, "A Novel Approach for Outdoor Fall Detection Using Multidimensional Features from A Single Camera," Applied Sciences, vol. 8, (6), 2018. DOI: 10.3390/app8060984.
[192] S. F. Ali et al, "Using Temporal Covariance of Motion and Geometric Features via Boosting for Human Fall Detection," Sensors, vol. 18, (6), 2018. DOI: 10.3390/s18061918.
[193] W. Min et al, "Support vector machine approach to fall recognition based on simplified expression of human skeleton action and fast detection of start key frame using torso angle," IET Computer Vision, vol. 12, (8), pp. 1133-1140, 2018. DOI: 10.1049/iet-cvi.2018.5324.
[194] W. Min et al, "Detection of Human Falls on Furniture Using Scene Analysis Based on Deep Learning and Activity Characteristics," IEEE Access, vol. 6, pp. 9324-9335, 2018. DOI: 10.1109/ACCESS.2018.2795239.
[195] X. ShanShan and C. Xi, "Fall detection method based on semi-contour distances," in - 2018 14th IEEE International Conference on Signal Processing (ICSP), 2018. DOI: 10.1109/ICSP.2018.8652332.
[196] (). CENTRE FOR DIGITAL HOME – MMU. Available: http://foe.mmu.edu.my/digitalhome/FallVideo.zip. Access 20/4/22.
[197] A. El Kaid, K. Baïna and J. Baïna, "Reduce False Positive Alerts for Elderly Person Fall Video-Detection Algorithm by convolutional neural network model," Procedia Computer Science, vol. 148, pp. 2-11, 2019. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.procs.2019.01.004. Access 12/8/22.
[198] C. Ma et al, "Fall detection using optical level anonymous image sensing system," Optics & Laser Technology, vol. 110, pp. 44-61, 2019. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.optlastec.2018.07.013. Access 12/6/22.
[199] D. Kumar et al, "Elderly health monitoring system with fall detection using multi-feature based person tracking," in - 2019 ITU Kaleidoscope: ICT for Health: Networks, Standards and Innovation (ITU K), 2019. DOI: 10.23919/ITUK48006.2019.8996141.
[200] MOT dataset. Available: https://motchallenge.net/. Access 5/6/22.
[201] T. Lin et al, "Microsoft coco: Common objects in context," in European Conference on Computer Vision, 2014, .
[202] F. Harrou et al, "An Integrated Vision-Based Approach for Efficient Human Fall Detection in a Home Environment," IEEE Access, vol. 7, pp. 114966-114974, 2019. DOI: 10.1109/ACCESS.2019.2936320.
[203] J. Brieva et al, "An intelligent human fall detection system using a vision-based strategy," in - 2019 IEEE 14th International Symposium on Autonomous Decentralized System (ISADS), 2019. DOI: 10.1109/ISADS45777.2019.9155767.
[204] M. Hua, Y. Nan and S. Lian, "Falls prediction based on body keypoints and Seq2Seq architecture," in - 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019. DOI: 10.1109/ICCVW.2019.00158.
[205] M. M. Hasan, M. S. Islam and S. Abdullah, "Robust pose-based human fall detection using recurrent neural network," in 2019 IEEE International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things (RAAICON), 2019, .
[206] P. K. Soni and A. Choudhary, "Automated fall detection from a camera using support vector machine," in - 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), 2019. DOI: 10.1109/ICACCP.2019.8882966.
[207] R. Espinosa et al, "A vision-based approach for fall detection using multiple cameras and convolutional neural networks: A case study using the UP-Fall detection dataset," Comput. Biol. Med., vol. 115, pp. 103520, 2019. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.compbiomed.2019.103520. Access 14/8/22.
[208] L. Martínez-Villaseñor et al, "UP-Fall Detection Dataset: A Multimodal Approach," Sensors (Basel, Switzerland), vol. 19, (9), pp. 1988, 2019. DOI: 10.3390/s19091988.
[209] S. Kalita, A. Karmakar and S. M. Hazarika, "Human fall detection during activities of daily living using extended CORE9," in - 2019 Second International Conference on Advanced Computational and Communication Paradigms (ICACCP), 2019. DOI: 10.1109/ICACCP.2019.8882928.
[210] S. Maldonado-Bascón et al, "Fallen People Detection Capabilities Using Assistive Robot," Electronics, vol. 8, (9), 2019. DOI: 10.3390/electronics8090915.

[211] X. Cai et al, "A novel method based on optical flow combining with wide residual network for fall detection," in - 2019 IEEE 19th International Conference on Communication Technology (ICCT), 2019. DOI: 10.1109/ICCT46805.2019.8947120.
[212] X. Kong et al, "Robust Self-Adaptation Fall-Detection System Based on Camera Height," Sensors (Basel), vol. 19, (17), 2019. DOI: 10.3390/s19173768.
[213] X. Kong et al, "A HOG-SVM Based Fall Detection IoT System for Elderly Persons Using Deep Sensor," Procedia Computer Science, vol. 147, pp. 276-282, 2019. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.procs.2019.01.264. Access 12/6/22.
[214] A. CARLIER et al, "Fall detector adapted to nursing home needs through an optical-flow based CNN," in - 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2020, . DOI: 10.1109/EMBC44109.2020.9175844.
[215] B. Wang et al, "Fall Detection Based on Dual-Channel Feature Integration," IEEE Access, vol. 8, pp. 103443-103453, 2020. DOI: 10.1109/ACCESS.2020.2999503.
[216] C. Menacho and J. Ordoñez, "Fall detection based on CNN models implemented on a mobile robot," in - 2020 17th International Conference on Ubiquitous Robots (UR), 2020. DOI: 10.1109/UR49135.2020.9144836.
[217] C. Zhong et al, "Multi-occupancy Fall Detection using Non-Invasive Thermal Vision Sensor," IEEE Sensors Journal, pp. 1, 2020. DOI: 10.1109/JSEN.2020.3032728.
[218] G. Sun and Z. Wang, "Fall detection algorithm for the elderly based on human posture estimation," in - 2020 Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2020. DOI: 10.1109/IPEC49694.2020.9114962.
[219] J. Liu et al, "Fall detection under privacy protection using multi-layer compressed sensing," in - 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2020. DOI: 10.1109/ICAIBD49809.2020.9137474.
[220] J. Thummala and S. Pumrin, "Fall detection using motion history image and shape deformation," in - 2020 8th International Electrical Engineering Congress (iEECON), 2020. DOI: 10.1109/iEECON48109.2020.229491.
[221] J. Zhang, C. Wu and Y. Wang, "Human Fall Detection Based on Body Posture Spatio-Temporal Evolution," Sensors, vol. 20, (3), 2020. DOI: 10.3390/s20030946.
[222] K. N. Kottari, K. K. Delibasis and I. G. Maglogiannis, "Real-Time Fall Detection Using Uncalibrated Fisheye Cameras," IEEE Transactions on Cognitive and Developmental Systems, vol. 12, (3), pp. 588-600, 2020. DOI: 10.1109/TCDS.2019.2948786.
[223] K. K. Delibasis, T. Goudas and I. Maglogiannis, "A novel robust approach for handling illumination changes in video segmentation," Eng Appl Artif Intell, vol. 49, pp. 43-60, 2016. DOI: https://doi.org/10.1016/j.engappai.2015.11.006. Access 15/3/22.
[224] PIROPO (People in Indoor ROoms with Perspective and Omnidirectional cameras). https://www.gti.ssr.upm.es/research/gti-data/databases. Access 12/6/22.
[225] Q. Feng et al, "Spatio-temporal fall event detection in complex scenes using attention guided LSTM," Pattern Recog. Lett., vol. 130, pp. 242-249, 2020. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.patrec.2018.08.031. Access 10/2/22.
[226] Q. Xu et al, "Fall prediction based on key points of human bones," Physica A: Statistical Mechanics and its Applications, vol. 540, pp. 123205, 2020. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.physa.2019.123205. Access 20/5/22.
[227] A. Shahroudy et al, "Ntu rgb d: A large scale dataset for 3d human activity analysis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[228] S. N. Htun, T. T. Zin and P. Tin, "Image Processing Technique and Hidden Markov Model for an Elderly Care Monitoring System," Journal of Imaging, vol. 6, (6), 2020. DOI: 10.3390/jimaging6060049.
[229] T. Kalinga et al, "A fall detection and emergency notification system for elderly," in - 2020 6th International Conference on Control, Automation and Robotics (ICCAR), 2020. DOI: 10.1109/ICCAR49639.2020.9108003.
[230] W. Chen et al, "Fall Detection Based on Key Points of Human-Skeleton Using OpenPose," Symmetry, vol. 12, (5), 2020. DOI: 10.3390/sym12050744.
[231] X. Cai et al, "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.
[232] Y. Chen et al, "Vision-Based Fall Event Detection in Complex Background Using Attention Guided Bi-Directional LSTM," IEEE Access, vol. 8, pp. 161337-161348, 2020. DOI: 10.1109/ACCESS.2020.3021795.
[233] Y. Chen et al, "An Edge Computing Based Fall Detection System for Elderly Persons," Procedia Computer Science, vol. 174, pp. 9-14, 2020. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.procs.2020.06.049. Access 12/5/22.
[234] X. Wang and K. Jia, "Human fall detection algorithm based on YOLOv3," in - 2020 IEEE 5th International Conference on Image, Vision and Computing (ICIVC), 2020. DOI: 10.1109/ICIVC50857.2020.9177447.
[235] K. Chaccour et al, "From Fall Detection to Fall Prevention: A Generic Classification of Fall-Related Systems," IEEE Sensors Journal, vol. 17, (3), pp. 812-822, 2017. DOI: 10.1109/JSEN.2016.2628099.
[236] L. Ren and Y. Peng, "Research of Fall Detection and Fall Prevention Technologies: A Systematic Review," IEEE Access, vol. 7, pp. 77702-77722, 2019. . DOI: 10.1109/ACCESS.2019.2922708.
[237] Nor Surayahani Suriani, Fadilla 'Atyka Nor Rashid and Nur Yuzailin Yunos, "Optimal Accelerometer Placement for Fall Detection of Rehabilitation Patients," vol. 10, 2017.
[238] J. Jacob et al, "A fall detection study on the sensors placement location and a rule-based multi-thresholds algorithm using both accelerometer and gyroscopes," in 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2017), 2017, pp. 666-671.
[239] C. Krupitzer et al, "Hips do lie! A position-aware mobile fall detection system," in 2018 IEEE International Conference on Pervasive Computing and Communications (PerCom), 2018. DOI: 10.1109/PERCOM.2018.8444583.
[240] Casilari E, Santoyo-Ramón JA,Cano-García JM, "Analysis of a Smartphone-Based Architecture with Multiple Mobility Sensors for Fall Detection," PLoS ONE 11, pp. 12, 2016. Available: https://doi.org/10.1371/journal.pone.0168069. Access 8/3/22.
[241] J. He et al, "A wearable method for autonomous fall detection based on kalman filter and k-NN algorithm," in 2016 IEEE Biomedical Circuits and Systems Conference (BioCAS), 2016. DOI: 10.1109/BioCAS.2016.7833821.
[242] J. He et al, "Application of kalman filter and k-NN classifier in wearable fall detection device," in 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing,

Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), 2017. DOI: 10.1109/UIC-ATC.2017.8397482.

[243] A. Sucerquia, J. Lopez and J. Vargas-Bonilla, "SisFall: A Fall and Movement Dataset," Sensors, vol. 17, pp. 198, 2017. . DOI: 10.3390/s17010198.

[244] S. D. Bersch et al, "Activity detection using frequency analysis and off-the-shelf devices: Fall detection from accelerometer data," in 2011 5th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth) and Workshops, 2011. DOI: 10.4108/icst.pervasivehealth.2011.246119.

[245] E. Casilari-Pérez and F. García-Lagos, "A comprehensive study on the use of artificial neural networks in wearable fall detection systems," Expert Syst. Appl., vol. 138, pp. 112811, 2019. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.eswa.2019.07.028. Access 11/3/22.

[246] S. Fudickar et al, "Fall-detection simulator for accelerometers with in-hardware preprocessing," in Jun 6, 2012, Available: http://dl.acm.org/citation.cfm?id=2413149 http://dl.acm.org/citation.cfm?id=2413149. DOI: 10.1145/2413097.2413149.

[247] S. Fudickar, A. Lindemann and B. Schnor, Threshold-Based Fall Detection on Smart Phones. 2015.

[248] D. Razum et al, "Optimal threshold selection for threshold-based fall detection algorithms with multiple features," in 2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2018 . DOI: 10.23919/MIPRO.2018.8400272.

[249] N. P. Pham et al, "Classification different types of fall for reducing false alarm using single accelerometer," in 2018 IEEE Seventh International Conference on Communications and Electronics (ICCE), 2018. DOI: 10.1109/CCE.2018.8465736.

[250] A. K. Thella et al, "Smart unit care for pre fall detection and prevention," in 2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS), 2016. DOI: 10.1109/NAECON.2016.7856802.

[251] M. Hemmatpour et al, "Nonlinear Predictive Threshold Model for Real-Time Abnormal Gait Detection," Journal of Healthcare Engineering, vol. 2018, pp. 1-9, 2018. DOI: 10.1155/2018/4750104.

[252] Y. wu et al, "A Multi-sensor Fall Detection System Based on Multivariate Statistical Process Analysis," Journal of Medical and Biological Engineering, vol. 39, 2018. DOI: 10.1007/s40846-018-0404-z.

[253] L. Ren and W. Shi, "Chameleon: Personalised and adaptive fall detection of elderly people in home-based environments," International Journal of Sensor Networks, vol. x x, 2015. DOI: 10.1504/IJSNET.2016.075365.

[254] O. Aziz et al, "A comparison of accuracy of fall detection algorithms (threshold-based vs. machine learning) using waist-mounted tri-axial accelerometer signals from a comprehensive set of falls and non-fall trials," Med. Biol. Eng. Comput., vol. 55, 2016. DOI: 10.1007/s11517-016-1504-y.

[255] T. Xu, Y. Zhou and J. Zhu, "New Advances and Challenges of Fall Detection Systems: A Survey," Applied Sciences, vol. 8, pp. 418, 2018. DOI: 10.3390/app8030418.

[256] X. Hu and X. Qu, "An individual-specific fall detection model based on the statistical process control chart," Saf. Sci., vol. 64, pp. 13-21, 2015. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.ssci.2013.11.010. Access 2/7/22.

[257] S. Ray, "A quick review of machine learning algorithms," in 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019. DOI: 10.1109/COMITCon.2019.8862451.

[258] D. Berrar, "Bayes' theorem and naive bayes classifier," in Encyclopedia of Bioinformatics and Computational Biology, S. Ranganathan et al, Ed. 2019. DOI: https://doi-org.ezproxy.uned.es/10.1016/B978-0-12-809633-8.20473-1. Access 12/5/22.

[259] L. Ren and Y. Peng, "Research of Fall Detection and Fall Prevention Technologies: A Systematic Review," IEEE Access, vol. 7, pp. 77702-77722, 2019. . DOI: 10.1109/ACCESS.2019.2922708.

[260] M. Kangas et al, "Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects," Gait Posture, vol. 35, (3), pp. 500-505, 2012.

[261] J. Klenk et al, "Comparison of acceleration signals of simulated and real-world backward falls," Med. Eng. Phys., vol. 33, (3), pp. 368-373, 2011. DOI: https://doi.org/10.1016/j.medengphy.2010.11.003. Access 10/6/22.

[262] Timo Jämsä et al, "Fall detection in the older people: from laboratory to real-life," Proceedings of the Stonian Academy of Sciences, pp. 341–345, 2014. DOI: 10.3176/proc.2014.3.08.

[263] E. Casilari and M. A. Oviedo-Jiménez, " Automatic Fall Detection System Based on the Combined Use of a Smartphone and a Smartwatch." PLoS ONE 10(11), 2015. Available: https://doi.org/10.1371/journal.pone.0140929. Access 2/9/22.

[264] Qijia Cheng et al, "Energy harvesting from human motion for wearable devices," in 10th IEEE International Conference on Nano/Micro Engineered and Molecular Systems, 2015. DOI: 10.1109/NEMS.2015.7147455.

[265] T. Wu et al, "Flexible wearable sensor nodes with solar energy harvesting," in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2017. DOI: 10.1109/EMBC.2017.8037555.

[266] K. Li et al, "Wearable energy harvesters generating electricity from low-frequency human limb movement." Microsystems & Nanoengineering, 4, 2018. Available: https://www.nature.com/articles/s41378-018-0024-3. Access 12/6/22.

[267] Y. Cai et al, "The distributed system of smart wearable energy harvesting based on human body," in 2018 37th Chinese Control Conference (CCC), 2018. DOI: 10.23919/ChiCC.2018.8483790.

[268] S. M. Noghabaei, R. L. Radin and M. Sawan, "Efficient dual-band ultra-low-power RF energy harvesting front-end for wearable devices," in 2018 IEEE 61st International Midwest Symposium on Circuits and Systems (MWSCAS), 2018. DOI: 10.1109/MWSCAS.2018.8623832.

[269] S. Roundy et al, "Inertial energy harvesting for wearables," in 2018 Ieee Sensors, 2018. DOI: 10.1109/ICSENS.2018.8589603.

[270] M. Mohammadifar et al, "A skin-mountable bacteria-powered battery system for self-powered medical devices," in 2020 IEEE 33rd International Conference on Micro Electro Mechanical Systems (MEMS), 2020. DOI: 10.1109/MEMS46641.2020.9056174.

[271] J. Silva, I. Sousa and J. Cardoso, "Transfer learning approach for fall detection with the FARSEEING real-world dataset and simulated falls," in - 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018. DOI: 10.1109/EMBC.2018.8513001.

[272] B. Mungamuru and P. Aarabi, "Enhanced sound localization," IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 34, (3), pp. 1526-1540, 2004. DOI: 10.1109/TSMCB.2004.826398.
[273] D. L. Donoho, "Compressed sensing," IEEE Transactions on Information Theory, vol. 52, (4), pp. 1289-1306, 2006. DOI: 10.1109/TIT.2006.871582.
[274] T. Horprasert, D. Harwood and L. S. Davis, "A statistical approach for real-time robust background subtraction and shadow detection," in Ieee Iccv, 1999.
[275] A. Mittal and N. Paragios, "Motion-based background subtraction using adaptive kernel density estimation," in - Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. 2004. DOI: 10.1109/CVPR.2004.1315179.
[276] M. A. Mousse, C. Motamed and E. C. Ezin, "Fast moving object detection from overlapping cameras," in - 2015 12th International Conference on Informatics in Control, Automation and Robotics (ICINCO), 2015.
[277] M. I. Chacon M., G. D. Sergio and V. P. Javier, "Simplified SOM-neural model for video segmentation of moving objects," in - 2009 International Joint Conference on Neural Networks, 2009. DOI: 10.1109/IJCNN.2009.5178632.
[278] V. Nguyen et al, "A new hand representation based on kernels for hand posture recognition," in - 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015. DOI: 10.1109/FG.2015.7163110.
[279] J. Kennedy and R. Eberhart, "Particle swarm optimization," in - Proceedings of ICNN'95 - International Conference on Neural Networks, 1995. DOI: 10.1109/ICNN.1995.488968.
[280] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, (3), pp. 257-267, 2001. DOI: 10.1109/34.910878.
[281] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," 1981.
[282] Jianbo Shi and Tomasi, "Good features to track," in - 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1994. DOI: 10.1109/CVPR.1994.323794.
[283] E. J. Candès et al, "Robust principal component analysis?" Journal of the ACM (JACM), vol. 58, (3), pp. 1-37, 2011.
[284] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in - 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005. DOI: 10.1109/CVPR.2005.177.
[285] Y. Nizam, M. N. Haji Mohd and M. M. Abdul Jamil, "A Study on Human Fall Detection Systems: Daily Activity Classification and Sensing Techniques," International Journal of Integrated Engineering, vol. 8, 2016.
[286] S. Kalita, A. Karmakar and S. Hazarika, "Efficient extraction of spatial relations for extended objects vis-à-vis human activity recognition in video," Appl. Intell., vol. 48, 2017. DOI: 10.1007/s10489-017-0970-8.
[287] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," Psychol. Rev., vol. 65, (6), pp. 386-408, 1958. DOI: 10.1037/h0042519.
[288] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," Proceedings of the National Academy of Sciences, vol. 79, (8), pp. 2554-2558, 1982.
[289] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, (8), pp. 1735-1780, 1997.
[290] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in Neural Information Processing Systems, 2014.
[291] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," J. Physiol. (Lond.), vol. 160, (1), pp. 106, 1962.
[292] K. Fukushima, "Biological cybernetics neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position," Biol. Cybern., vol. 36, pp. 193-202, 1980.
[293] Y. Lecun et al, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, (11), pp. 2278-2324, 1998. DOI: 10.1109/5.726791.
[294] M. Tygert et al, "A mathematical motivation for complex-valued convolutional networks," Neural Comput., vol. 28, (5), pp. 815-825, 2016.
[295] N. Wojke, A. Bewley and D. Paulus, "Simple online and realtime tracking with a deep association metric," in 2017 IEEE International Conference on Image Processing (ICIP), 2017.
[296] I. N. Junejo and H. Foroosh, "Euclidean path modeling for video surveillance," Image Vision Comput., vol. 26, (4), pp. 512-528, 2008.
[297] C. Maldonado et al, "Feature selection to detect fallen pose using depth images," in - 2016 International Conference on Electronics, Communications and Computers (CONIELECOMP), 2016. DOI: 10.1109/CONIELECOMP.2016.7438558.
[298] R. Labayrade, D. Aubert and J. -. Tarel, "Real time obstacle detection in stereovision on non flat road geometry through "v-disparity" representation," in - Intelligent Vehicle Symposium, 2002. IEEE, 2002. DOI: 10.1109/IVS.2002.1188024.
[299] T. Schmiedel, E. Einhorn and H. Gross, "IRON: A fast interest point descriptor for robust NDT-map matching and its application to robot localization," in - 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015. DOI: 10.1109/IROS.2015.7353812.
[300] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Commun ACM, vol. 24, (6), pp. 381-395, 1981.
[301] M. D. Solbach and J. K. Tsotsos, "Vision-based fallen person detection for the elderly," in - 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017. DOI: 10.1109/ICCVW.2017.170.
[302] Keqi Zhang et al, "A progressive morphological filter for removing nonground measurements from airborne LIDAR data," IEEE Transactions on Geoscience and Remote Sensing, vol. 41, (4), pp. 872-882, 2003. DOI: 10.1109/TGRS.2003.810682.
[303] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learning, vol. 20, (3), pp. 273-297, 1995.
[304] D. M. J. Tax and R. P. W. Duin, "Support Vector Data Description," Mach. Learning, vol. 54, (1), pp. 45-66, 2004. DOI: 10.1023/B:MACH.0000008084.60811.49.
[305] L. Breiman, "Random forests," Mach. Learning, vol. 45, (1), pp. 5-32, 2001.

[306] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," Annals of Statistics, pp. 1189-1232, 2001.
[307] L. Rabiner and B. Juang, "An introduction to hidden Markov models," Ieee Assp Magazine, vol. 3, (1), pp. 4-16, 1986.
[308] P. Natarajan and R. Nevatia, "Online, real-time tracking and recognition of human actions," in - 2008 IEEE Workshop on Motion and Video Computing, 2008. DOI: 10.1109/WMVC.2008.4544064.
[309] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.
[310] N. Gordon, "Bayesian methods for tracking," 1993.
[311] M. Kangas et al, "Comparison of real-life accidental falls in older people with experimental falls in middle-aged test subjects," Gait Posture, vol. 35, (3), pp. 500-505, 2012. DOI: https://doi-org.ezproxy.uned.es/10.1016/j.gaitpost.2011.11.016. Access 16/5/22.
[312] Anonymous "ISO/IEC/IEEE International Standard - Systems and software engineering -- Life cycle processes -- Requirements engineering," Iso/Iec/Ieee 29148:2018(E), pp. 1-104, 2018. DOI: 10.1109/IEEESTD.2018.8559686.
[313] Anonymous "IEEE Recommended Practice for Software Requirements Specifications," IEEE Std 830-1998, pp. 1-40, 1998. DOI: 10.1109/IEEESTD.1998.88286.
[314] S. Elo and H. Kyngäs, "The qualitative content analysis process," J. Adv. Nurs., vol. 62, (1), pp. 107-115, 2008. DOI: https://doi.org/10.1111/j.1365-2648.2007.04569.x. Access 11/3/22.
[315] U. Flick, "The SAGE Handbook of Qualitative Data Analysis," 2022. DOI: 10.4135/9781446282243.
[316] Y. Zang et al, "Pose estimation at night in infrared images using a lightweight multi-stage attention network," Signal, Image and Video Processing, vol. 15, (8), pp. 1757-1765, 2021.
[317] B. Sapp and B. Taskar, "Modec: Multimodal decomposable models for human pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[318] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation." in Bmvc, 2010.
[319] M. Andriluka et al, "2d human pose estimation: New benchmark and state-of-the-artanalysis," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[320] T. Lin et al, "Microsoft coco: Common objects in context," in Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, 2014.
[321] C. Ionescu et al, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, (7), pp. 1325-1339, 2013.
[322] D. Mehta et al, "Monocular 3d human pose estimation in the wild using improved cnn supervision," in 2017 International Conference on 3D Vision (3DV), 2017.
[323] M. Habermann et al, "Deepcap: Monocular human performance capture using weak supervision," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[324] W. Gong et al, "Human pose estimation from monocular images: A comprehensive survey," Sensors, vol. 16, (12), pp. 1966, 2016.
[325] A. Elgammal and C. Lee, "Inferring 3D body pose from silhouettes using activity manifold learning," in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. 2004.
[326] A. Agarwal and B. Triggs, "Recovering 3D human pose from monocular images," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, (1), pp. 44-58, 2005.
[327] D. M. Gavrila, "A bayesian, exemplar-based approach to hierarchical shape matching," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, (8), pp. 1408-1421, 2007.
[328] P. Viola, M. J. Jones and D. Snow, "Detecting pedestrians using patterns of motion and appearance," International Journal of Computer Vision, vol. 63, pp. 153-161, 2005.
[329] B. Sapp, A. Toshev and B. Taskar, "Cascaded models for articulated pose estimation," in Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part II 11, 2010, .
[330] M. Dimitrijevic, V. Lepetit and P. Fua, "Human body pose detection using bayesian spatio-temporal templates," Comput. Vision Image Understanding, vol. 104, (2-3), pp. 127-139, 2006.
[331] C. Weinrich, M. Volkhardt and H. Gross, "Appearance-based 3D upper-body pose estimation and person re-identification on mobile robots," in 2013 IEEE International Conference on Systems, Man, and Cybernetics, 2013.
[332] C. T. Zahn and R. Z. Roskies, "Fourier descriptors for plane closed curves," IEEE Trans. Comput., vol. 100, (3), pp. 269-281, 1972.
[333] L. Gorelick et al, "Shape representation and classification using the poisson equation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 28, (12), pp. 1991-2005, 2006.
[334] H. Jiang, "Human pose estimation using consistent max covering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, (9), pp. 1911-1918, 2011.
[335] C. H. Ek, P. H. Torr and N. D. Lawrence, "Gaussian process latent variable models for human pose estimation," in Machine Learning for Multimodal Interaction: 4th International Workshop, MLMI 2007, Brno, Czech Republic, June 28-30, 2007, Revised Selected Papers 4, 2008.
[336] G. Mori, S. Belongie and J. Malik, "Efficient shape matching using shape contexts," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, (11), pp. 1832-1837, 2005.
[337] S. Belongie, J. Malik and J. Puzicha, "Shape matching and object recognition using shape contexts," IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, (4), pp. 509-522, 2002.
[338] Q. Zhu et al, "Fast human detection using a cascade of histograms of oriented gradients," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006.
[339] N. DARAL, "Histograms of oriented gradients for human detection," Proc.of CVPR, 2005, pp. 886-893, 2005.
[340] D. G. Lowe, "Object recognition from local scale-invariant features," in Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999.
[341] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, pp. 91-110, 2004.

[342] P. Sabzmeydani and G. Mori, "Detecting pedestrians by learning shapelet features," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[343] Y. Wu, J. Lin and T. S. Huang, "Analyzing and capturing articulated hand motion in image sequences," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, (12), pp. 1910-1922, 2005.
[344] H. Jiang, "Human pose estimation using consistent max covering," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, (9), pp. 1911-1918, 2011.
[345] A. Agarwal and B. Triggs, "A local basis representation for estimating human pose from cluttered images," in Computer Vision–ACCV 2006: 7th Asian Conference on Computer Vision, Hyderabad, India, January 13-16, 2006. Proceedings, Part I 7, 2006.
[346] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005.
[347] A. Kanaujia, C. Sminchisescu and D. Metaxas, "Semi-supervised hierarchical models for 3d human pose reconstruction," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[348] S. Lazebnik, C. Schmid and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006.
[349] A. Agarwal and B. Triggs, "Hyperfeatures–multilevel local coding for visual recognition," in Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, may 7-13, 2006. Proceedings, Part I 9, 2006.
[350] T. Serre, L. Wolf and T. Poggio, "Object recognition with features inspired by visual cortex," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005.
[351] G. Gkioxari et al, "Using k-poselets for detecting people and localizing their keypoints," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[352] L. Bourdev et al, "Detecting people using mutually consistent poselet activations," in Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part VI 11, 2010.
[353] T. J. Roberts, S. J. McKenna and I. W. Ricketts, "Human pose estimation using learnt probabilistic region similarities and partial configurations," in Eccv (4), 2004.
[354] G. Oleinikov et al, "Task-based control of articulated human pose detection for openvl," in IEEE Winter Conference on Applications of Computer Vision, 2014.
[355] C. Sminchisescu and B. Triggs, "Covariance scaled sampling for monocular 3D body tracking," in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001.
[356] Y. Lu and H. Jiang, "Human movement summarization and depiction from videos," in 2013 IEEE International Conference on Multimedia and Expo (ICME), 2013.
[357] S. Zuffi, O. Freifeld and M. J. Black, "From pictorial structures to deformable structures," in 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012.
[358] M. Yamamoto and K. Yagishita, "Scene constraints-aided tracking of human body," in Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. no. PR00662), 2000.
[359] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. no. 98CB36231), 1998.
[360] J. Wang et al, "Locality-constrained linear coding for image classification," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.
[361] A. Bissacco, M. Yang and S. Soatto, "Fast human pose estimation using appearance and motion via multi-dimensional boosting regression," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[362] M. Andriluka, S. Roth and B. Schiele, "Pictorial structures revisited: People detection and articulated pose estimation," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.
[363] M. Andriluka, S. Roth and B. Schiele, "Monocular 3d pose estimation and tracking by detection," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.
[364] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation." in Bmvc, 2010.
[365] L. Pishchulin et al, "Poselet conditioned pictorial structures," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[366] S. Johnson and M. Everingham, "Learning effective human pose estimation from inaccurate annotation," in Cvpr 2011, 2011.
[367] N. Komodakis, N. Paragios and G. Tziritas, "MRF energy minimization and beyond via dual decomposition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 33, (3), pp. 531-552, 2010.
[368] Y. Wang and G. Mori, "Multiple tree models for occlusion and spatial constraints in human pose estimation." in Eccv (3), 2008.
[369] K. Tashiro et al, "Refinement of ontology-constrained human pose classification," in 2014 IEEE International Conference on Semantic Computing, 2014.
[370] A. M. Lehrmann, P. V. Gehler and S. Nowozin, "A non-parametric bayesian network prior of human pose," in Proceedings of the IEEE International Conference on Computer Vision, 2013.
[371] A. Baumberg and D. Hogg, "Learning flexible models from image sequences," in Computer Vision—ECCV'94: Third European Conference on Computer Vision Stockholm, Sweden, may 2–6, 1994 Proceedings, Volume I 3, 1994.
[372] O. Freifeld et al, "Contour people: A parameterized model of 2D articulated human shape," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.
[373] T. F. Cootes et al, "Active shape models-their training and application," Comput. Vision Image Understanding, vol. 61, (1), pp. 38-59, 1995.
[374] H. Jiang, "Finding human poses in videos using concurrent matching and segmentation," in Computer Vision–ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part I 10, 2011.
[375] J. Gall et al, "Motion capture using joint skeleton tracking and surface estimation," in 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009.

[376] S. Ge and G. Fan, "Non-rigid articulated point set registration for human pose estimation," in 2015 IEEE Winter Conference on Applications of Computer Vision, 2015.
[377] S. Ge and G. Fan, "Articulated non-rigid point set registration for human pose estimation from 3D sensors," Sensors, vol. 15, (7), pp. 15218-15245, 2015.
[378] A. O. Balan et al, "Detailed human shape and pose from images," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[379] P. Guan et al, "Estimating human shape and pose from a single image," in 2009 IEEE 12th International Conference on Computer Vision, 2009.
[380] J. F. Blinn, "Models of light reflection for computer synthesized pictures," in Proceedings of the 4th Annual Conference on Computer Graphics and Interactive Techniques, 1977.
[381] A. O. Balan et al, "Shining a light on human pose: On shadows, shading and the estimation of pose and shape," in 2007 IEEE 11th International Conference on Computer Vision, 2007.
[382] T. B. Moeslund and E. Granum, "A survey of computer vision-based human motion capture," Comput. Vision Image Understanding, vol. 81, (3), pp. 231-268, 2001.
[383] R. Urtasun et al, "Priors for people tracking from small training sets," in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005.
[384] D. Metaxas and D. Terzopoulos, "Shape and nonrigid motion estimation through physics-based synthesis," IEEE Trans. Pattern Anal. Mach. Intell., vol. 15, (6), pp. 580-591, 1993.
[385] M. A. Brubaker, D. J. Fleet and A. Hertzmann, "Physics-based person tracking using simplified lower-body dynamics," in 2007 IEEE Conference on Computer Vision and Pattern Recognition, 2007.
[386] W. Zhang, L. Shang and A. B. Chan, "A robust likelihood function for 3D human pose tracking," IEEE Trans. Image Process., vol. 23, (12), pp. 5374-5389, 2014.
[387] U. Güdükbay, I. Demir and Y. Dedeoğlu, "Motion capture and human pose reconstruction from a single-view video sequence," Digital Signal Processing, vol. 23, (5), pp. 1441-1450, 2013.
[388] G. Pons-Moll and B. Rosenhahn, "Model-based pose estimation," Visual Analysis of Humans: Looking at People, pp. 139-170, 2011.
[389] V. Parameswaran and R. Chellappa, "View independent human body pose estimation from a single perspective image," in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004. 2004.
[390] G. Pons-Moll et al, "Efficient and robust shape matching for model based human motion capture," in Pattern Recognition: 33rd DAGM Symposium, Frankfurt/Main, Germany, August 31–September 2, 2011. Proceedings 33, 2011.
[391] V. Ganapathi et al, "Real-time human pose tracking from range data," in Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12, 2012.
[392] M. A. Brubaker, D. J. Fleet and A. Hertzmann, "Physics-based person tracking using the anthropomorphic walker," International Journal of Computer Vision, vol. 87, (1-2), pp. 140, 2010.
[393] J. Gall et al, "Optimization and filtering for human motion capture: A multi-layer framework," International Journal of Computer Vision, vol. 87, pp. 75-92, 2010.
[394] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," International Journal of Computer Vision, vol. 61, pp. 185-205, 2005.
[395] R. Okada and S. Soatto, "Relevant feature selection for human pose estimation and localization in cluttered images." in Eccv (2), 2008.
[396] R. Ronfard, C. Schmid and B. Triggs, "Learning to parse pictures of people," in Computer Vision—ECCV 2002: 7th European Conference on Computer Vision Copenhagen, Denmark, may 28–31, 2002 Proceedings, Part IV 7, 2002.
[397] W. Zhang et al, "A latent clothing attribute approach for human pose estimation," in Computer Vision–ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part I 12, 2015.
[398] S. Sedai, M. Bennamoun and D. Q. Huynh, "Evaluating shape and appearance descriptors for 3D human pose estimation," in 2011 6th IEEE Conference on Industrial Electronics and Applications, 2011.
[399] S. Li and A. B. Chan, "3d human pose estimation from monocular images with deep convolutional neural network," in Computer Vision--ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part II 12, 2015.
[400] H. Ning et al, "Discriminative learning of visual words for 3D human pose estimation," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[401] M. I. Jordan and R. A. Jacobs, "Hierarchical mixtures of experts and the EM algorithm," Neural Comput., vol. 6, (2), pp. 181-214, 1994.
[402] O. Freifeld and M. J. Black, "Lie Bodies: A Manifold Representation of 3D Human Shape." Eccv (1), vol. 7572, pp. 1-14, 2012.
[403] A. Baak et al, "A data-driven approach for real-time full body pose reconstruction from a depth camera," Consumer Depth Cameras for Computer Vision: Research Topics and Applications, pp. 71-98, 2013.
[404] C. M. Christoudias and T. Darrell, "On modelling nonlinear shape-and-texture appearance manifolds," in 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005.
[405] J. Gall, A. Yao and L. Van Gool, "2D action recognition serves 3D human pose estimation." in Eccv (3), 2010.
[406] V. I. Morariu and O. I. Camps, "Modeling correspondences for multi-camera tracking using nonlinear manifold learning and target dynamics," in 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), 2006.
[407] G. Mori et al, "Pose embeddings: A deep architecture for learning to match human poses," arXiv Preprint arXiv:1507.00302, 2015.
[408] A. Gupta et al, "Context and observation driven latent variable model for human pose estimation," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[409] J. Wang et al, "Locality-constrained linear coding for image classification," in 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010.

[410] L. Sun et al, "Motionlet LLC coding for discriminative human pose estimation," Multimedia Tools Appl, vol. 73, pp. 327-344, 2014.
[411] H. Ning et al, "Discriminative learning of visual words for 3D human pose estimation," in 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008.
[412] V. Belagiannis et al, "Holistic human pose estimation with regression forests," in Articulated Motion and Deformable Objects: 8th International Conference, AMDO 2014, Palma De Mallorca, Spain, July 16-18, 2014. Proceedings 8, 2014.
[413] L. Breiman, "Random forests," Mach. Learning, vol. 45, pp. 5-32, 2001.
[414] 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, (6), pp. 949-959, 2013.
[415] M. Dantone et al, "Human pose estimation using body parts dependent joint regressors," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[416] R. Girshick et al, "Efficient regression of general-activity human poses from depth images," in 2011 International Conference on Computer Vision, 2011.
[417] S. Wei et al, "Convolutional pose machines," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[418] A. Newell, K. Yang and J. Deng, "Stacked hourglass networks for human pose estimation," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, 2016.
[419] Z. Su et al, "Cascade feature aggregation for human pose estimation," arXiv Preprint arXiv:1902.07837, 2019.
[420] Z. Cao et al, "Realtime multi-person 2d pose estimation using part affinity fields," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[421] E. Insafutdinov et al, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VI 14, 2016.
[422] Y. Chen et al, "Cascaded pyramid network for multi-person pose estimation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[423] A. Toshev and C. Szegedy, "Deeppose: Human pose estimation via deep neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[424] A. Dosovitskiy et al, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv Preprint arXiv:2010.11929, 2020.
[425] S. Khan et al, "Transformers in vision: A survey," ACM Computing Surveys (CSUR), vol. 54, (10s), pp. 1-41, 2022.
[426] A. Krizhevsky, I. Sutskever and G. E. Hinton, "2012 AlexNet," Adv.Neural Inf.Process.Syst, pp. 1-9, 2012.
[427] J. Tompson et al, "Efficient object localization using convolutional networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[428] S. Wei et al, "Convolutional pose machines," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[429] A. Newell, K. Yang and J. Deng, "Stacked hourglass networks for human pose estimation," in Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part VIII 14, 2016.
[430] J. Carreira et al, "Human pose estimation with iterative error feedback," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[431] W. Mao et al, "Tfpose: Direct human pose estimation with transformers," arXiv Preprint arXiv:2103.15320, 2021.
[432] X. Zhu et al, "Deformable detr: Deformable transformers for end-to-end object detection," arXiv Preprint arXiv:2010.04159, 2020.
[433] Y. Xu et al, "Vitpose: Simple vision transformer baselines for human pose estimation," arXiv Preprint arXiv:2204.12484, 2022.
[434] G. Ning, Z. Zhang and Z. He, "Knowledge-guided deep fractal neural networks for human pose estimation," IEEE Transactions on Multimedia, vol. 20, (5), pp. 1246-1259, 2017.
[435] X. Liu et al, "LightPose: A lightweight and efficient model with transformer for human pose estimation," in ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.
[436] J. Wang et al, "Deep high-resolution representation learning for visual recognition," IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, (10), pp. 3349-3364, 2020.
[437] H. Tao et al, "Few shot cross equipment fault diagnosis method based on parameter optimization and feature mertic," Measurement Science and Technology, vol. 33, (11), pp. 115005, 2022.
[438] X. Song et al, "Event-driven NN adaptive fixed-time control for nonlinear systems with guaranteed performance," Journal of the Franklin Institute, vol. 359, (9), pp. 4138-4159, 2022.
[439] Z. Zhuang et al, "Iterative learning control for repetitive tasks with randomly varying trial lengths using successive projection," Int J Adapt Control Signal Process, vol. 36, (5), pp. 1196-1215, 2022.
[440] D. A. Winter, "Human balance and posture control during standing and walking," Gait Posture, vol. 3, (4), pp. 193-214, 1995.
[441] S. N. Robinovitch et al, "Video capture of the circumstances of falls in elderly people residing in long-term care: an observational study," The Lancet, vol. 381, (9860), pp. 47-54, 2013.
[442] L. I. Wolfson et al, "Stressing the postural response: a quantitative method for testing balance," J. Am. Geriatr. Soc., vol. 34, (12), pp. 845-850, 1986.
[443] M. A. Holbein and M. S. Redfern, "Functional stability limits while holding loads in various positions," Int. J. Ind. Ergonomics, vol. 19, (5), pp. 387-395, 1997.
[444] M. R. Popović et al, "Stability criterion for controlling standing in able-bodied subjects," J. Biomech., vol. 33, (11), pp. 1359-1368, 2000.
[445] B. E. Maki, P. J. Holliday and G. R. Fernie, "Aging and postural control: a comparison of spontaneous-and induced-sway balance tests," J. Am. Geriatr. Soc., vol. 38, (1), pp. 1-9, 1990.
[446] R. Hamideh, "A comparison between static and dynamic stability in postural sway and fall risk," Journal of Ergonomics, vol. 7, pp. 1-7, 2017.

[447] S. M. Bruijn et al, "Assessing the stability of human locomotion: a review of current measures," Journal of the Royal Society Interface, vol. 10, (83), pp. 20120999, 2013.
[448] A. L. Hof, M. Gazendam and W. E. Sinke, "The condition for dynamic stability," J. Biomech., vol. 38, (1), pp. 1-8, 2005.
[449] D. L. Wight, E. G. Kubica and D. W. Wang, "Introduction of the foot placement estimator: A dynamic measure of balance for bipedal robotics," Journal of Computational and Nonlinear Dynamics, vol. 3, (1), 2008.
[450] T. Caderby et al, "Influence of gait speed on the control of mediolateral dynamic stability during gait initiation," J. Biomech., vol. 47, (2), pp. 417-423, 2014.
[451] A. L. Hof et al, "Control of lateral balance in walking: experimental findings in normal subjects and above-knee amputees," Gait Posture, vol. 25, (2), pp. 250-258, 2007.
[452] M. Millard et al, "Human foot placement and balance in the sagittal plane," 2009.
[453] M. Millard, J. McPhee and E. Kubica, "Foot placement and balance in 3D," 2012.
[454] F. Huo, "Limits of stability and postural sway in young and older people." 2000.
[455] S. M. Bruijn et al, "Gait stability in children with Cerebral Palsy," Res. Dev. Disabil., vol. 34, (5), pp. 1689-1699, 2013.
[456] S. M. Bruijn and J. H. Van Dieën, "Control of human gait stability through foot placement," Journal of the Royal Society Interface, vol. 15, (143), pp. 20170816, 2018.
[457] G. Casiez, N. Roussel and D. Vogel, "1€ filter: A simple speed-based low-pass filter for noisy input in interactive systems," in Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 2012.
[458] D. Pavllo et al, "3d human pose estimation in video with temporal convolutions and semi-supervised training," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[459] Y. Zou et al, "Reducing footskate in human motion reconstruction with ground contact constraints," in - 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), 2020. DOI: 10.1109/WACV45572.2020.9093329.
[460] C. Lea et al, "Temporal convolutional networks for action segmentation and detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.
[461] M. Holschneider et al, "A real-time algorithm for signal analysis with the help of the wavelet transform," in WaveletsAnonymous 1990.
[462] Y. Wei et al, "Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[463] X. Zhang, Y. Zou and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in 2017 22nd International Conference on Digital Signal Processing (DSP), 2017.
[464] L. Zhou, C. Zhang and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018.
[465] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in Icml, 2010.
[466] K. He et al, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[467] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in International Conference on Machine Learning, 2015.
[468] N. Srivastava et al, "Dropout: a simple way to prevent neural networks from overfitting," The Journal of Machine Learning Research, vol. 15, (1), pp. 1929-1958, 2014.
[469] G. Pavei et al, "On the estimation accuracy of the 3D body center of mass trajectory during human locomotion: inverse vs. forward dynamics," Frontiers in Physiology, vol. 8, pp. 129, 2017.
[470] P. De Leva, "Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters," J. Biomech., vol. 29, (9), pp. 1223-1230, 1996.
[471] V. Z. V. Seluyanov, "The mass and inertia characteristics of the main segments of the human body," Biomechanics VIII-B, pp. 1152-1159, 1983.
[472] C. Napier et al, "The use of a single sacral marker method to approximate the centre of mass trajectory during treadmill running," J. Biomech., vol. 108, pp. 109886, 2020.
[473] M. A. Thirunarayan et al, "Comparison of three methods for estimating vertical displacement of center of mass during level walking in patients," Gait Posture, vol. 4, (4), pp. 306-314, 1996. DOI: https://doi.org/10.1016/0966-6362(95)01058-0. Access 15/5/22.
[474] S. A. Gard, S. C. Miff and A. D. Kuo, "Comparison of kinematic and kinetic methods for computing the vertical motion of the body center of mass during walking," Human Movement Science, vol. 22, (6), pp. 597-610, 2004. DOI: https://doi.org/10.1016/j.humov.2003.11.002. Access 18/5/22.
[475] M. Roser, C. Appel and H. Ritchie, "Human height," Our World in Data, 2013.
[476] P. Davidovits, Physics in Biology and Medicine. 2018.
[477] M. Batista and J. Peternelj, "The Falling Time of an Inverted Plane Pendulum," arXiv Preprint Physics/0607080, 2006.
[478] I. Habibie et al, "In the wild human pose estimation using explicit 2d features and intermediate 3d representations," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
[479] A. Kanazawa et al, "End-to-end recovery of human shape and pose," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.
[480] C. Ionescu et al, "Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments," IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, (7), pp. 1325-1339, 2013.
[481] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," International Journal of Data Mining & Knowledge Management Process, vol. 5, (2), pp. 1, 2015.
[482] V. T. Inman and H. D. Eberhart, "The major determinants in normal and pathological gait," Jbjs, vol. 35, (3), pp. 543-558, 1953.
[483] F. Yang and Y. Pai, "Can sacral marker approximate center of mass during gait and slip-fall recovery among community-dwelling older adults?" J. Biomech., vol. 47, (16), pp. 3807-3812, 2014.

[484] N. Gill et al, "Are the arms and head required to accurately estimate centre of mass motion during running?" Gait Posture, vol. 51, pp. 281-283, 2017. DOI: https://doi.org/10.1016/j.gaitpost.2016.11.001. Access 4/1/23.
[485] J. Liu et al, "NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, (10), pp. 2684-2701, 2020. DOI: 10.1109/TPAMI.2019.2916873.
[486] Q. Xu et al, "Fall prediction based on key points of human bones," Physica A: Statistical Mechanics and its Applications, vol. 540, pp. 123205, 2020.
[487] A. Shojaei-Hashemi et al, "Video-based human fall detection in smart homes using deep learning," in 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018.
[488] T. Tsai and C. Hsu, "Implementation of fall detection system based on 3D skeleton for deep learning technique," IEEE Access, vol. 7, pp. 153049-153059, 2019.
[489] V. Bazarevsky et al, "Blazepose: On-device real-time body pose tracking," arXiv Preprint arXiv:2006.10204, 2020.
[490] E. Auvinet et al, "Multiple cameras fall dataset," DIRO-Université De Montréal, Tech.Rep, vol. 1350, pp. 24, 2010.
[491] M. Kepski and B. Kwolek, "Embedded system for fall detection using body-worn accelerometer and depth sensor," in 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 2015.
[492] J. Gutiérrez, V. Rodríguez and S. Martin, "Comprehensive Review of Vision-Based Fall Detection Systems," Sensors, vol. 21, (3), 2021. DOI: 10.3390/s21030947.
[493] C. Menacho and J. Ordoñez, "Fall detection based on CNN models implemented on a mobile robot," in - 2020 17th International Conference on Ubiquitous Robots (UR), 2020. DOI: 10.1109/UR49135.2020.9144836.
[494] J. Gutiérrez, V. Rodriguez and S. Martin, "Fall detection system based on far infrared images," in 2022 Congreso De Tecnología, Aprendizaje Y Enseñanza De La Electrónica (XV Technologies Applied to Electronics Teaching Conference), 2022.
[495] J. Gutiérrez, S. Martin and V. Rodriguez, "Human stability assessment and fall detection based on dynamic descriptors," IET Image Processing.

Annex A – Electronic folder description

This annex describes the contain and structure of the electronic folder which includes both this PhD thesis manuscript itself and the most relevant software needed to develop it:

- CoGNet folder contains both the trained network itself and the datasets used to train it.
    - CoGNet_dataset_generation_SW contains the software needed to develop and label the datasets used to train and validate the CoGNet neural network.
    - CoGNet_Datasets contains the datasets used to train and validate the datasets used to train and validate CoGNet.
    - CoGNet_implementation_trx contains the CoGNet network and the training and validation codes needed to develop it as well as the weights obtained in the training process.
- Fall_detection_network_implementation_trx contains the considered pose estimation neural networks, the training and validation codes needed to develop it and the weights obtained in the training process.
    - 1_DeepPose contains the DeepPose network.
    - 2_ConvNet contains the ConvNet Pose network.
    - 3_CPM contains the Convolutional Pose Machines network.
    - 4_Hourglass contains the HourGlass network.
    - 5_HPE contains the Human Pose Estimation with iterative error feedback network.
    - 6_CFA contains the Cascade feature aggregation for human pose estimation network.
    - 7_TFP contains the TFPose network.
    - 8_ViTP contains the ViTPose network.
- FIR-Human contains the FIR-Human dataset.
    - BLOCK_1_TRX contains the module developed with training purposes.
    - BLOCK_2_VAL contains the module developed with validation purposes.
    - BLOCK_3_FALL contains the module of falls.

Annex B – FIR-Human

This dataset contains video-clips of five volunteers developing daily life activities. Each video-clip is recorded with a FIR camera and includes an associated file which contains the three-dimensional and two-dimensional coordinates of the main body joints in each frame of the clip. This way, it is possible to train human pose estimation networks using FIR imagery.

Diversity and Size

- Over 250.000 2D and 3D human poses and their corresponding FIR images.
- 5 volunteers (4 males, 1 female).
- 27 action classes including falls of different kinds.

Accurate Capture and Synchronization

- FIR video clips recorded at 23.98 frames per second with a resolution of 480 x 640 pixels.
- Accurate 3D positions of the 19 main body joints from high-speed motion capture system.
- Precise 2D projections of the body joints onto the image plane.

Subjects

The motions were performed by five volunteers: four males and one female, whose body mass indexes (BMI) ranged from 16 to 24. This guarantees body type and movement variability. Additionally, the volunteers wore a range of diverse clothes, and the thermal conditions of the laboratory where the recording was made have changed to provide a dataset as rich and diverse as possible.

Action classes and dataset structure

The dataset contains 27 action classes in total. 26 of them are daily life activities while the other one includes different types of falls. The different actions are repeated by the volunteers in four different positions so frontal, rear and side views of the same actions are 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. These actions include:

1. Giving directions.
2. Discussing.
3. Eating.
4. Taking photos.
5. Exercising on the ground.
6. Running in place.
7. Walking.
8. Sitting and standing up.
9. Coughing.
10. Exercising.
11. Playing basketball.

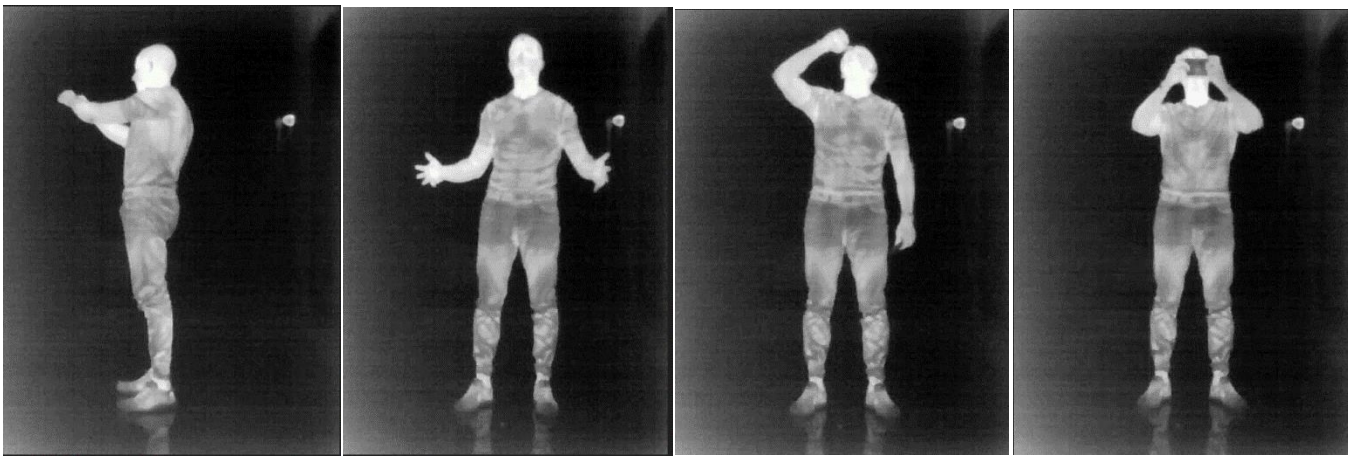
12. Picking up objects.
13. Limping.

The second block includes a single person who executes a different set of actions with validation purposes. These activities include:

1. Brushing teeth.
2. Encouraging your team.
3. Toasting.
4. Taking a selfie.
5. Crouching for meditation.
6. Walking a dog.
7. Throwing a stone.
8. Talking on the phone.
9. Stretching yourself.
10. Hopping.
11. Kicking a ball.
12. Tying shoelaces.
13. Rotating your trunk.

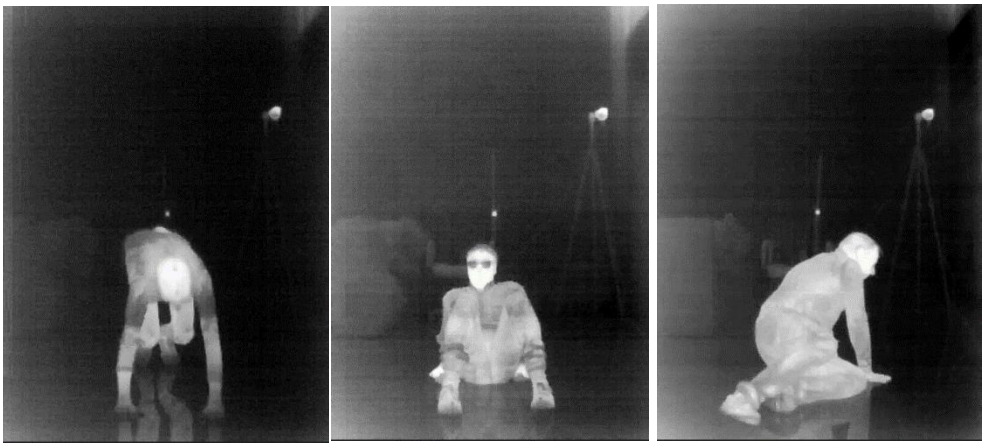
Finally, the third block includes four volunteers who are recorded from different perspectives falling forward, falling backwards and side falling. The falls start from static or dynamic situations and a number of them are slow falls, a common type of fall in the elderly community.

First block



Second block

Third block



Labeling

All video-clips include a joint file in .mat format (label.mat) which contains two data structures. The first one, called data_2D, includes the two-dimensional position of the 19 main body joints in each frame of the video-clip it is associated to. The second data

structure, data_3D, includes the same information but, this time, the information is three-dimensional.