FALL DETECTION SYSTEM USING YOLO VERSION 3

S.P. Saranya

pg scholar, Sir Issac Newton College of Engineering and Technology

M.kavitha

Assistant Professor, Sir Issac Newton College of Engineering and Technology

P.Arivazhagan

Assistant Professor, Sir Issac Newton College of Engineering and Technology

***Abstract*— In this paper, With the increase of the elderly population, the phenomenon of the elderly falling at home or out is more and more common. Therefore, fall detection is of great significance for the health protection of the elderly. Throughout the research of fall detection at home and abroad, most of the fall detection based on video monitoring is complex and redundant, which affects the real-time and accuracy of detection. Given the above problems, this paper proposes a fall detection method based on a video in a complex environment, aiming to detect fall behavior more accurately and quickly. The main work of this paper is as follows: firstly, the YOLOv3 network model is proposed for the detection algorithm. Secondly, the human fall detection data set is constructed by referring to the Pascal VOC data set format. Then, the algorithm model is optimized and trained in GPU (graphic processing unit) deep learning server. Finally, a comparison of test results with our YOLOv3 network model and other detection algorithms shows that our detection algorithm has a good recognition effect.**

***Index Terms*—Convolution Neural Networks (CNN), Human activity recognition (HAR), Web Server Gateway Interface (WSGI), YOLO.**

INTRODUCTION

HUMANS exist commonly in our daily visual data, e.g. private photographs, public surveillance videos, etc. Therefore, the fundamental of interpreting these visual contents lies in a comprehensive analysis of humans via obtaining their poses, body parts, and identities, etc. However, we notice that the performance of these human-centric under- standing techniques drops dramatically when the resolution of the input image becomes extremely low. To overcome this problem, a simple strategy is to upsample the low resolution(LR) images using some interpolation methods, which inevitably introduce blurriness. Alternatively, single image super resolution (SR) techniques are introduced to reconstruct High-Resolution(HR) images from LR counterparts with higher visual quality. As a pre-processing for understanding humans in visual data, human body image SR can facilitate other human centric tasks such as human pose estimation [1], human parsing [2], pedestrian attribute recognition [3], and person re-identification [4], etc.

Recent advances in generic image SR take advantage of the powerful representation ability of convolutional neural networks (CNNs). From the first CNN based super resolution network SRCNN [5] (3 convolutional layers) to RCAN [6] (more than 400 layers), the overall performance has improved dramatically. Although deep networks are expected to achieve higher performance, they are too expensive w.r.t. memory and computational time. For many human image analysis related applications, especially those on mobile devices, a lightweight model is required. To this end, another line of works focusing on efficiency is of great interest. An effective strategy is to aggregate features of multiple receptive fields so as to enhance the compactness. Some works [7], [8] use different filter sizes, e.g., 5 5 or 7 7, to build multi-scale blocks. The disadvantage is that the kernel sizes are large, resulting in rather high number of parameters. Aiming to reduce the parameters, we alternatively use dilated convolution and extract features of multiple receptive fields using different dilation rates.

In this paper we propose Fall is one of the main life-threaten factors for humans, especially the elderly who live alone. It is caused by the inability of their muscle to support and balance their body due to the aging process. Fall events may cause serious injuries especially in the elderly community and some may be fatal. Several fall prevention solutions had been deployed by different manufactures and industries, but they are still some falls that are unpreventable. Following a fall event, immediate help and treatment are extremely critical. Therefore, fall should be noticed immediately to prevent life-threatening risk. The outcomes of fall events are far beyond physical injuries as they may also lead to psychological, medical, and social consequences. This work aims in developing an automated image based fall detection system utilizing the YOLOv3 algorithm that can help monitor elderly activity. The fall events will be detected and notified upon detection. Our project proposed to integrate the YOLOv3 object detection algorithm with the IFADS fall detection algorithm to achieve low cost, high accuracy, and real-time computing requirements.

RELATED WORK

The one and only existing work on human body image super-resolution [16] was proposed for surveillance images and used an exemplar-based method. In this paper we address the human body image SR in diverse scenarios via a deep learning based approach, which is considered more powerful than exemplar-based methods. Specifically, we propose a CNN based method using

prior knowledge to solve human body image SR. In this section, we contrast our method with state- of-the-art image SR methods in Table I, and review related works from the following two perspectives.

A. CNN-Based Generic Image SR

In general, existing CNN-based methods are composed of two crucial components: the feature extraction module and up-sampling module. For the up-sampling module, previous works [17], [18] use bicubic interpolation to pre-upsample the LR images before feeding them into networks, resulting in significant increase of the computational costs. Alternatively, many methods perform the up-sampling operation after feature extraction. Some of them use sub-pixel convolutional layer, e.g., ESPCN [19] and LP-KPN [20]; while others use deconvolutional layer, e.g., FSRCNN [21], DBPN [22], and IDN [23].For the feature extraction module, recent works tend to adopt deep CNN architectures to solve SISR. To make the training easier, residual connections are adopted to build fea- ture extraction blocks, e.g., RCAN [6] and CARN [24]. Some other works borrow dense connections for better performance, e.g., SRDenseNet [25], RDN [26], and SRFBN [27]. Recently, many works [7], [8], [28] adopt multi-scale block to build their feature extraction module and exhibit improved performance. For many human analysis applications, efficiency is an important concern. Thus, in this paper, we propose an efficient feature extraction block to greatly reduce the model parameters while obtaining reasonable SR performance.

B. Prior-Based Specific Image SR

Compared to generic image SR, there are many spe cific image SR methods that use prior knowledge to better super-resolve LR images. However, some early methods [29],[30] are hard to train and apply in practice, as they solve the image SR and prior estimation separately. Alternatively, some methods integrate the image recon- struction and prior estimation into a multi-task framework so as to allow for end-to-end training and inference. For example, Stimpel *et al.* [31] combine the locally linear guided filter with a learned guidance map for medical image SR. Wang *et al.* [32] use the loss of text recognition to guide the training of the SR network, and thus it pays more attention to the text content. Some other methods try to explore facial prior information or attributes to facilitate face SR. In these methods, different kinds of priors including face attributes [11], [33], parsing maps [12], [34], and facial component heatmaps [35], [36] show to be beneficial to reconstruct HR face images with higher visual quality. Similar to face images, human body images also exhibit distinct structures and texture details for different body com- ponents. To this end, this paper adopts the human parsing and NSST to represent the human body prior knowledge, which can be leverage to better super-resolve human body images.

PROPOSED METHOD

In this section, we will first give an overview of our method. Then, we will describe the proposed feature of Darknet 53 YOLO V3, system architecture, Neural Networks respectively.

Overview

As illustrated in Figure 1, we propose a novel CNN-based human body image SR method by jointly solving image reconstruction and prior estimation in a coherent framework. Specifically, our CNN consists of two branches: (1) the image reconstruction branch takes an original LR image as input and reconstructs an HR image through a deep SR network; (2) the prior estimation branch estimates the human body prior and injects it back into image reconstruction branch as useful cues. The loss function L of our model is defined as:

$$L(\hat{\theta}) = \alpha L_{image}(\theta_1) + (1-\alpha) L_{prior}(\theta_2), \quad (1)$$

where θ_1 and θ_2 are network parameters of image recon- struction branch and prior estimation branch, respectively. $\hat{\theta}$ θ_1, θ_2 denotes the parameters to be estimated, L_{Image} and L_{Prior} are the loss functions of image reconstruction branch and prior estimation branch, respectively, α is the hyper-parameter to control the contribution between two losses.

1) Image Reconstruction Branch: We propose an efficient lightweight multi-scale block (LMSB) to serve as a basic block for feature extraction in the image reconstruction branch. Let x denote the input LR image and y denote the reconstructed HR image. In the image reconstruction branch, we first stack two convolutional layers to extract shallow feature-maps F_0 from x . Then, the F_0 is fed into the network consisting of multiple stacked LMSBs. Suppose we have N LMSBs in the feature extraction module, the output F_n of the n-th block can be represented as:

$$F_n = H_n(F_{n-1}) = H_n(H_{n-1}(\cdots(H_1(F_0))\cdots)), \quad (2)$$

where H_n is the operation of the n-th block. Inspired by [6], we use global residual learning to improve the information flow. Finally, a deconvolutional layer is utilized to reconstruct $\hat{y}$as follows,

$$\hat{y} = C_1(C_0(F_0 + F_N)), \quad (3)$$

where C_0 and C_1 denote the operation of 3 × 3 convolution and 12 × 12 deconvolution, respectively. And we use the $L1$ loss for image reconstruction.

Specifically, our CNN consists of two branches: (1) the image reconstruction branch takes an original LR image as input and reconstructs an HR image through a deep SR network; (2) the prior estimation branch estimates the human body prior and injects

it back into image reconstruction branch as useful cues. The loss function L of our model is defined as:

$$L(\hat{\theta}) = \alpha L_{image}(\theta_1) + (1-\alpha) L_{prior}(\theta_2), \quad (1)$$

where θ_1 and θ_2 are network parameters of image recon- struction branch and prior estimation branch, respectively. $\hat{\theta}$ θ_1, θ_2 denotes the parameters to be estimated, $_{Image}$
and $_{Prior}$ are the loss functions of image reconstruction branch and prior estimation branch, respectively, α is the hyper-parameter to control the contribution between two losses.

IMAGE RECONSTRUCTION BRANCH:
We propose an efficient lightweight multi-scale block (LMSB) to serve as a basic block for feature extraction in the image reconstruction branch. Let In existing system the main objective of human fall detection systems, they need to be trained so that they can discriminate between a human fall and other activities of daily life (ADL) (walking, standing, sitting, lying) ,This data can be collected from different types of sensors installed in the environment such as pressure sensors, floor vibration sensors, infrared sensors, microphones, and cameras. The data collected in the form of acceleration signals, pressure signals, audio, or videos are then processed and passed to the classifier which then classifies whether the collected data represents a fall or an ADL.

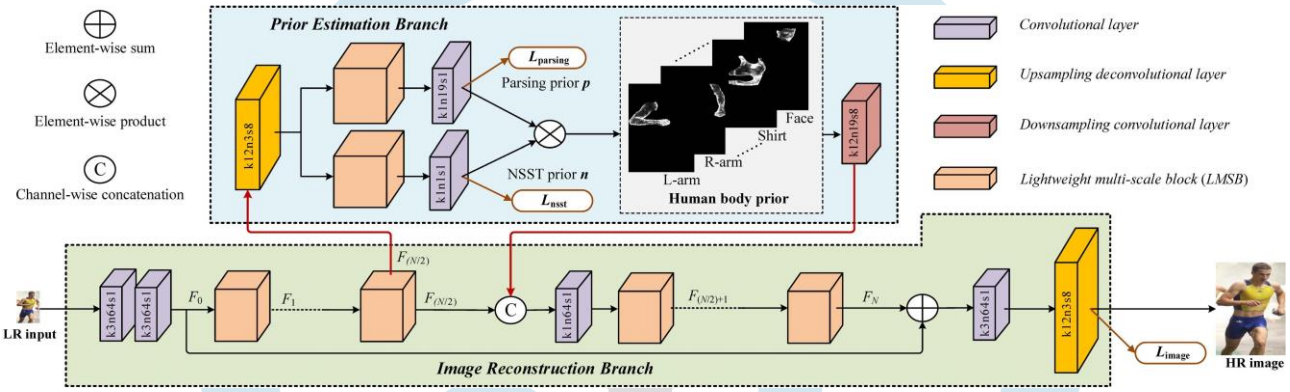


Fig. 1. Overview of coherent framework.

CONTEXT-AWARE FALL DETECTION SYSTEMS:
These systems include sensors which are deployed in the environment around the humans to detect a fall. Systems designed based on this class of technology include ambience such as pressure sensors, floor sensors, infrared sensors, and microphones. These systems consist of a set of the above mentioned sensors to collect data. Along with them, a dedicated PC is attached. The collected data is passed to the PC for further processing and analysis The algorithm running in PC, based on certain threshold values and conditions decides whether a fall has taken place or not. Machine learning classification algorithms are also being used for the classification of human activity in two classes i.e. fall and non-fall. Systems developed under this category are further divided into two sub categories depending on the type of data collected by these sensors: (a) ambient-based fall detection systems for numerical sensor data and (b) acoustic-based fall detection systems for audio data.

AMBIENT-BASED FALL DETECTION SYSTEMS:
Floor and pressure sensors have to be installed on the ground so that they can read the frequency of vibrations generated during the fall. Along with them, an infrared sensor is also used to detect motion in the surrounding environment.
the fifth and sixth layers, seventh and eighth layers. The number of feature maps in the eight convolutional layers is respectively 64, 128, 256, 256, 512, 512, 512, and 512. The dimension of each layer has been indicated in Fig.2. There are five pooling layers where maxpooling has been employed, where the max filter is employed on non-overlapping regions to down-sample the input representation. To further retain the motion information embedded between frames, the convolution kernel size of the first pooling layer is set as 1×2×2, and for the rest pooling layers, the kernel size is respectively set as 2×2×2, 2×2×2, 2×2×2, and 1×1×2.

With the goal of fall detection where the kinetic feature is salient, the dataset of sports videos Sports-1M has been selected to train the 3D CNN, which is also the largest video classification benchmark currently. Similar to the operations in [34], each frame has been resized into 128×171 and then the input image cube is randomly clipped from the video with a length of 16 frames and size of 112×112 for each frame toThese systems are environment dependent since every house has a different flooring. Hence, specific configurations are required for their setup which makes it difficult to install them.

FALL DETECTION MODELS USING MACHINE LEARNING:
For classification between a fall and an ADL, Boyle and Karunanithi, and Chen used threshold-based method (fall is detected when values of certain features cross their specified threshold), while others used predefined machine learning classifiers. Few surveys on threshold- based fall detection techniques have already been done by researchers (Yu, 2008; Sposaro and In this paper, we have focused on a survey of fall detection models, developed using machine learning algorithms, as described below. The architecture of the 3D CNN is as described below in Fig. 2. The input to the CNN is an image cube composed of 16 frames segmented from the video sequence. there are eight convolutional layers, where the third and fourth convolutional layers are connected directly, and so

are the fifth and sixth layers, seventh and eighth layers. The number of feature maps in the eight convolutional layers is respectively 64, 128, 256, 256, 512, 512, 512, and 512. The dimension of each layer has been indicated in Fig.2. There are five pooling layers where maxpooling has been employed, where the max filter is employed on non-overlapping regions to down-sample the input representation. To further retain the incorporate temporal and spatial jittering. Thus, the input to the CNN is 112×112×16 as shown in Fig. 2. With one input propagated through the network, a feature cube of 7×7×512 will be obtained at the fifth pooling layer, which will then be fed to train the following LSTM.

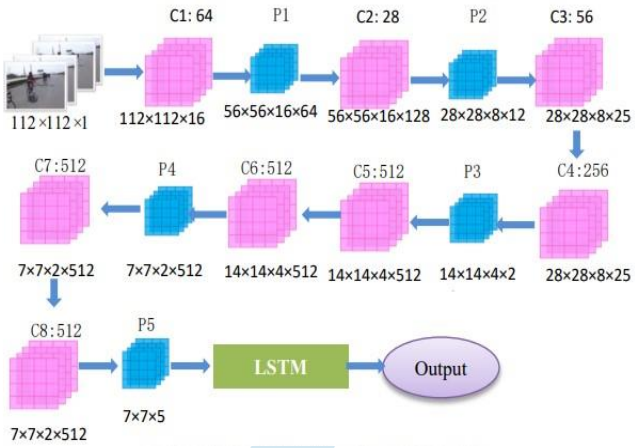


Fig.2. STATE OF THE ART

Background noise can emerge randomly because the optical flow technique analyzes the whole image and computes the flow for every pixel.The features of the current system. The useful features are highlighted inside the broken blue line. The limitations in the system are highlighted inside the broken red border. The system was proposed by Núñez-Marcos et al. They used deep CNN to decide if a video contains a person falling or not. This approach uses optical flow images as an input to the deep network. However, the optical flow images ignore any appearance-related features such as color, contrast, and brightness. The proposed approach minimizes the hand-crafted image processing steps by using CNN. CNN can learn a set of features and improved the performance when enough examples are provided during the training phase. However, the proposed system has been made more generic. Núñez-Marcos et al. presented a vision-based fall detection system using a CNN, which applies transfer learning from the action recognition domain to fall detection. Three different public datasets were used to evaluate the proposed approach. This model consists of two main stages: Preprocessing stage, and feature extraction, and classification stage.

Fig.3. State of the Art System

PRE-PROCESSING STAGE:

The video consists of contiguous frames stacked together. For a fall detection system, just a few adjacent frames are needed to detect a fall in the video. Besides that, some techniques process each frame individually and do not consider the correlation between the stacked frames. To solve this problem, Núñez-Marcos et al. utilized an optical flow algorithm that describes the displacement vector between two frames. The optical flow generator receives consecutive images and uses the TVL-1 optical flow algorithm that considers just motion in the video and removes any static background. Although static background can be removed from the motion of video, the perfect foreground cannot be obtained by utilizing the optical flow technique alone due to

various background noise, which may arise from the change in brightness.
Due to the presence of background noise, the accuracy of flow estimation decreased in the low light situation. Additionally, computing flow estimation for every pixel and generating a stack of optical flow images requires very high processing power, which is eventually increased the estimating of processing time. New solutions try to overcome general optical flow problem problems such as constant brightness assumption. This assumption is based on three successive images of a sequence. The approach assumes that motion is translational on a region large enough to regularize the aperture problem. In the way of avoiding outliers, a robust multi-resolution method was used. It is composed of a low-pass pyramid and an M-estimator technique. This method offered some good results on artificial and natural image sequences.The optical flow algorithm represents the patterns of the motion of objects as displacement vector fields between two consecutive images. Núñez-Marcos et al. selected the TVL-1 optical flow algorithm due to its performance with changing lighting conditions compared to another algorithm. However, the TVL1 optical flow algorithm can capture just small events. In a real-world situation, the human fall event involves many complex human actions. TVL-1 algorithm does not require modeling the dynamics of video content, which is crucial to increase the accuracy of any human activities such as fall. Eq. (1) describes the TVL-1 optical flow equation.

TVL-1 has not been considered a temporal evolution of human action. The first part of the TVL-1 of Eq. (1) is the data term that includes a brightness consistency assumption, which measures the accuracy with velocity field that describes the observable image motion. The second part is the regularization term, which penalizes high variations in the optical flow field to obtain smooth displacement fields. The regularization term in TVL-1 optical flow poses several issues. Some of them are not feasible in the learning of 'flow patterns' if different image structures are given, and the motion of the camera may give rise to a multitude of motion patterns with little resemblance between motion fields from different videos. So, it is necessary to use an appropriate regularization term. Due to these issues with the regularization term, it is difficult to obtain accurate optical flow estimation.

FEATURE EXTRACTION AND CLASSIFICATION STAGE :

In this stage, CNN was used for feature extraction and classification of images. In particular, VGG-16 CNN was used. The architecture of VGG-16 was chosen because of its high accuracy in the feature extraction process. VGG-16 CNN architecture consists of 3 layers: convolution layers, max-pooling layers, and pooling layers. In convolution layers, VGG-16 restricts the use of 3x3 convolution kernels size. The choice of small kernel size leads to reduce the number of parameters that can result in practical training and testing. Most importantly, with a small kernel size, it can stack more layers in deep networks; as a result, it would increase the performance of fall detection. By stacking series of 3x3 size kernel filters, the effective receptive can be increased to larger values, for example, 5x5 with two layers, 7x7 size with three layers. In VGG architecture, each convolution layer was followed by the rectified linear unit (ReLU) layer. Pooling layers operate on the blocks of the input feature map and combines the featured activations. This combination task is defined by a max function, which selects the maximum activation from the chosen group of blocks. The Max-pooling layer minimizes the spatial size of the feature, and it helps in reducing the number of parameters and the amount of computation in the network. A fully connected layer takes the input from the ReLU layer and generates the class score that is used for the classification of fall detection. VGG16 architecture uses dropout layers in the first two fully connected layers to avoid over fitting. The input layer of VGG-16 accepts a stack of optical flow images. The ImageNet dataset was utilized to address a small number of fall samples in the dataset. Besides, it used to learn the generic features as it has 14 million images. Based on the CNN-trained ImageNet dataset, the input to the CNN is modified to accept input images of size 224x224 and stack size of 20. The ImageNet dataset was utilized to address a small number of fall samples in the dataset. Besides, it used to learn the generic features as it has 14 million images. Based on the CNN-trained.

DISADVANTAGE:

- Number of frame processing is low in this existing system.
- Low accuracy.
- Need to improve the overall performance of the system.

I. PROPOSED FALL DETECTION

Fall detection devices automatically employ the technology to detect and get fast assistance for a senior that is prone to falls. A fall detection medical alert system allows the user to summon help without having to press the call button. These systems automatically activate the sensor if the user suffers a fall. The built-in technology can be worn around the neck or, depending on the device, on your wrist or your waist.

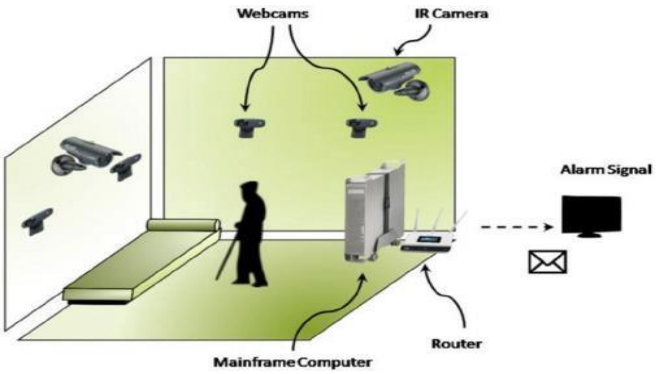


Fig.4. Built in Technology of Detection falls

Worldwide, falls are a leading cause of unintentional injuries in adults older than 65 years old, with 37.3 million falls requiring medical attention and 646,000 resulting in deaths annually. Seniors living alone are at high risk. Many common neurological problems result in falls2: Peripheral neuropathy manifesting with numbness and imbalance, spinal stenosis4 resulting in pain and in coordination, acute strokes5 leading to sudden weakness, and Parkinson's disease characterized by postural instability, etc. In addition, cardiovascular, musculoskeletal, and medication-induced problems often coexist. Orthostatic hypotension, knee arthritis, and iatrogenic dizziness are only a few examples. Even healthy senior activities such as climbing ladders, taking showers, going downstairs, and walking in snow could be dangerous.

Falls are not exclusively problems for the elderly, but may also be a concern for the young. Postural Orthostatic Tachycardia Syndrome, seizures, anemia, pregnancy, and sports all can lead to unexpected falls. Without timely detection and treatment, complications such as bone fractures, intracranial hemorrhage, or nerve avulsion can result. Permanent disabilities and death are not unusual1. In 2015, the medical cost for falls exceeded $50 billion. As the world population ages, the number of serious falls and subsequent financial burdens rises accordingly. It is imperative to detect falls timely to initiate appropriate medical responses to reduce the significant physical, social, and financial damages.

Currently, fall detection methods are broadly classified as wearable devices, environmental sensors, and image detectors, but significant limitations exist. Wearable sensors utilize tri-axial accelerometers to measure body inclination and are mounted to the wrist or another body part, or attached to the shoe insoles or garment fabrics. Gyroscopes estimate the rotational acceleration. Unfortunately, these sensors need frequent manual calibration due to fluctuations in temperature and humidity and could send false signals. Also, people could forget or feel uncomfortable wearing these devices or fail to replace batteries.

Furthermore, healthy seniors can also fall accidentally but they usually do not have these devices, which require a costly 24-h monitoring team and monthly subscription fee.

The second detection category utilizes various environmental elements. For example, acoustic sensors measure the sound of falls; pressure sensors measure the weight changes on the floor. Infrared sensors map out a person's heat signature and ultrasound detectors process the return signal. Near-field imaging with matrices of electrodes under floors track fall patterns. A common challenge is differentiating humans from animals or objects. Usually, the detection accuracy by acoustic arrays decreases if the person is five or more meters (m) away. Besides, some technologies are too expensive and impractical to be installed in every room. Even for the less costly wireless physical layer using channel state information, detection fails with multiple people in the room or if the furniture is pushed away during the fall, interfering with the mathematics designed only for single-entity monitoring.

The third category is image-based and is sub-classified into multiple cameras, single cameras, or images with three-dimensional (3D) depth data. The multiple-camera network reconstructs a 3D image, analyzes the volume distribution of the individual along the vertical axis, and triggers an alarm when most of the volume is near the floor for a predefined period38. This system requires a complicated setup with time-consuming calibration and fails to detect falls when there is more than one person in the room or when one is partially occluded by furniture. Fall is one of the main life-threaten factors for humans, especially the elderly who live alone. It is caused by the inability of their muscle to support and balance their body due to the aging process. Fall events may cause serious injuries especially in the elderly community and some may be fatal. Several fall prevention solutions had been deployed by different manufactures and industries, but they are still some falls that are unpreventable. Following a fall event, immediate help and treatment are extremely critical. Therefore, fall should be noticed immediately to prevent life-threatening risk. The outcomes of fall events are far beyond physical injuries as they may also lead to psychological, medical, and social consequences. This project aims in developing an automated image based fall detection system utilizing the YOLOv3 algorithm that can help monitor elderly activity. The fall events will be detected and notified upon detection. Our project proposed to integrate the YOLOv3 object detection algorithm with the IFADS fall detection algorithm to achieve low cost, high accuracy, and real-time computing requirements.

FALL ANALYSIS

We first characterized the biomechanics of the most common types of falls. A fall occurs when the center of the gravity (CG) of a person's trunk becomes misaligned with the base of support provided by the feet against the foor48. the CG is an imaginary point at the level of the sternum anterior to the spine, at which all the weight of the torso is evenly distributed. Stumbling (Fig. 1a) results from accidentally stepping onto an unperceived object while inertia keeps the CG moving, resulting in an imbalance in the torso. The trunk can fix anteriorly and so the edge of support approximates a vertical line passing through the tars metatarsal joints of

the foot. When the CG is shifted beyond this edge, due to resistance encountered in the moving feet, the person stumbles and falls. Stumbling commonly occurs in poorly lit rooms with misplaced items on the floor.

Fig.5. Automated images-based fall detection system

Selected fall types. (a) Stumbling. (b) Slipping. (c) Fainting. (d) Getting up from a sitting position (i.e., a chair) and falling as in orthostasis. (e) Falling from a high structure (i.e., stairs, ladders, etc.). (f) Jumping down from a high structure and falling.

Individuals with neurological or musculoskeletal disorders are more at risk for stumbling. Slipping (Fig.1b) occurs when the frictional force opposing the direction of foot movement is less than the horizontal shear force of the foot immediately after the heel contacts the foor49. The legs slide out of place and the person can no longer stay upright. Seniors have a reduced density of sensorimotor nerve fibers in the feet and often slip in bathrooms and kitchens or when walking downstairs50. Improper footwear and environmental obstacles further increase the likelihood of slipping. Individuals with preexisting gait difficulty such as from back pain, Parkinson disease, multiple sclerosis, or stroke particularly slip easily. Fainting (Fig. 1c) is due to impaired cerebral perfusion and transient brain hypoxia51, leading to a loss of postural tone. It is characterized by direct descent of the head and torso, while the CG remains in line with the feet, Followed by bending of the torso and the knee, and then the whole body stumbles and collapses. Any pathology impeding adequately oxygenated blood flow to the brain can result in fainting, and this could range from chronic anemia, vasovagal syncope, paroxysmal arrhythmia, to dysautonomia, to name just a few. Other common types of falls present as variations of stumbling or slipping. For example, falls develop while getting up from a chair (Fig. 1d) or sitting into it are commonly observed when the elderly use lightweight chairs or stools with wheels. Other examples include falling or jumping down from a high structure such as ladders and desks (Fig. 1e,f), and tumbling down or slipping while walking down stairs, etc.

YOLO-Version-3

You only look once, or YOLO is one of the faster object detection algorithms out there. Though it is no longer the most accurate object detection algorithm, it is a very good choice when you need real-time detection, without loss of too much accuracy.

A few weeks back, the third version of YOLO came out, and this post aims at explaining the changes introduced in YOLO v3. This is not going to be a post explaining what YOLO is from the ground up. I assume you know how YOLO v2 works. If that is not the case, I recommend you to check out the following papers by Joseph Redmon et all, to get a hang of how YOLO works.

The official title of the YOLO v2 paper seemed as if YOLO was a milk based health drink for kids rather than an object detection algorithm. It was named “YOLO9000: Better, Faster, Stronger”. For its time YOLO 9000 was the fastest, and also one of the most accurate algorithms. However, a couple of years down the line and it’s no longer the most accurate with algorithms like RetinaNet, and SSD outperforming it in terms of accuracy. It still, however, was one of the fastest. But that speed has been traded off for boosts inaccuracy in YOLO v3. While the earlier variant ran on 45 FPS on a Titan X, the current version clocks about 30 FPS. This has to do with the increase in complexity of underlying architecture called Darknet.

DARKNET-53

YOLO v2 used a custom deep architecture darknet-19, an originally 19-layer network supplemented with 11 more layers for object detection. With a 30-layer architecture, YOLO v2 often struggled with small object detections. This was attributed to the loss of fine-grained features as the layers down-sampled the input. To remedy this, YOLO v2 used identity mapping, concatenating feature maps from a previous layer to capture low-level features.

However, YOLO v2’s architecture was still lacking some of the most important elements that are now stapled in most state-of-the-art algorithms. No residual blocks, no skip connections, and no up sampling. YOLO v3 incorporates all of these. First, YOLO v3 uses a variant of Darknet, which originally has a 53 layer network trained on Image net. For the task of detection, 53 more layers are stacked onto it, giving us a 106 layer fully convolutional underlying architecture for YOLO v3. This is the reason behind the slowness of YOLO v3 compared to YOLO v2. Here is what the architecture of YOLO now looks like.

ImageNet dataset, the input to the CNN is modified to accept input images of size 224x224 and stack size of 20. The network was retrained on the optical flow stack of the UCF101 dataset. In the final step of transfer learning, the weight of the convolution layer was frozen to make the weight unaltered during the training stage.

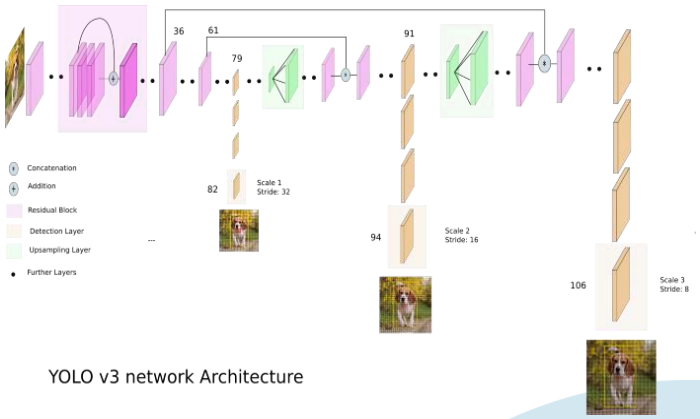


Fig.6. YOLO V3 Network architecture

DEEP LEARNING:

Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised. Deep-learning architectures such as deep neural networks, deep belief networks, deep reinforcement learning, recurrent neural networks and convolutional neural networks have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance.Artificial neural networks (ANNs) were inspired by information processing and distributed communication nodes in biological systems. ANNs have various differences from biological brains. Specifically, artificial neural networks tend to be static and symbolic, while the biological brain of most living organisms is dynamic (plastic) and analogue. The adjective "deep" in deep learning refers to the use of multiple layers in the network. Early work showed that a linear perceptron cannot be a universal classifier, but that a network with a no polynomial activation function with one hidden layer of unbounded width can. Deep learning is a modern variation which is concerned with an unbounded number of layers of bounded size, which permits practical application and optimized implementation, while retaining theoretical universality under mild conditions. In deep learning the layers are also permitted to be heterogeneous and to deviate widely from biologically informed connectionist models, for the sake of efficiency, trainability and understand ability, whence the "structured" part.

In an artificial neural network, there are three kinds of layers: the input layer, hidden layer and output layer. In the input layer, input vectors x=(x1,x2,…,xp) are provided to a system in order to test that system. In the output layer, final outputs are provided. The hidden layer is located between the input layer and output layer. When the hidden layers are increased, it becomes Deep. Deep Learning is a Machine Learning paradigm that use this Deep artificial neural network as a learning model. Also, Deep Learning is extremely useful because it is an unsupervised machine learning approach which means that it does not need labeled data. Because of this technology, we have a wealth of data that can be used for deep learning. Let's look at some common examples of data.

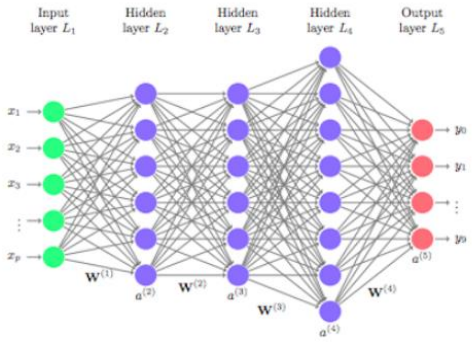


Fig.7. Artificial Neural Networks

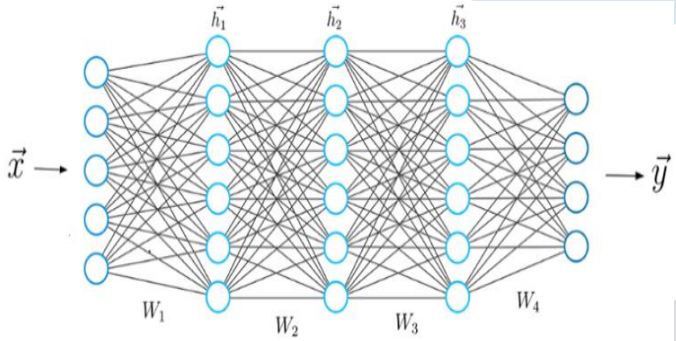
Deep Learning is a subset of Machine Learning, which on the other hand is a subset of Artificial Intelligence. Artificial Intelligence is a general term that refers to techniques that enable computers to mimic human behavior. Machine Learning represents a set of algorithms trained on data that make all of this possible.

Deep Learning, on the other hand, is just a type of Machine Learning, inspired by the structure of a human brain. Deep learning algorithms attempt to draw similar conclusions as humans would by continually analyzing data with a given logical structure. To achieve this, deep learning uses a multi-layered structure of algorithms called neural networks.

The design of the neural network is based on the structure of the human brain. Just as we use our brains to identify patterns and classify different types of information, neural networks can be taught to perform the same tasks on data. The individual layers of neural networks can also be thought of as a sort of filter that works from gross to subtle, increasing the likelihood of detecting and

outputting a correct result. The human brain works similarly. Whenever we receive new information, the brain tries to compare it with known objects. The same concept is also used by deep neural networks. Neural networks enable us to perform many tasks, such as clustering, classification or regression. With neural networks, we can group or sort unlabeled data according to similarities among the samples in this data. Or in the case of classification, we can train the network on a labeled dataset in order to classify the samples in this dataset into different categories.

In general, neural networks can perform the same tasks as classical algorithms of machine learning. However, it is not the other way around. Artificial neural networks have unique capabilities that enable deep learning models to solve tasks that machine learning models can never solve. All recent advances in artificial intelligence in recent years are due to deep learning. Without deep learning, we would not have self-driving cars, chatbots or personal assistants like Alexa and Siri. The Google Translate app would continue to be as primitive as 10 years ago (before Google switched to neural networks for this App), and Netflix or Youtube would have no idea which movies or TV series we like or dislike. Behind all these technologies are neural networks. We can even go so far as to say that today a new industrial revolution is taking place, driven by artificial neural networks and deep learning. At the end of the day, deep learning is the best and most obvious approach to real machine intelligence we've had so far.



FALL DETECTION

The dataset for out project is collected from the Third party website called Kaggle. Each two classes contains 150 images. The labelIMg is a image labelling tool which is used in the proposed method. The darknet framework help to train the model with help of darknet50 network. Trained model has been applied in the deployment process. Web cam capture is directly applied to the deployed model. The deployed model is process the video frame by frame.

At present, target detection tasks based on deep learning mainly include two types of detection methods: one is a two stage detection model based on region recommendation to extract target candidate regions, such as Fast-RCNN; the other is a one-stage detection model based on regression thought, such as YOLO, SSD, etc. Although the accuracy of one-stage detection model is slightly inferior, it can achieve real-time detection better in detection speed. Human fall is a sudden situation. It is necessary to detect the fall target accurately at the fastest speed. Therefore, this paper uses the YOLOv3 algorithm which takes both speed and detection accuracy into account. This network is mainly composed of a series of 1*1 and 3*3 convolution layers, because the network has 53 convolutions, it's called Darknet-53.

Darknet-53.

First, the image is scaled to a unified form of 416 in length and width, which is the input of the whole network. Secondly, feature extraction is carried out through the Darknet-53 network, and convolution operations are carried out alternately using convolution kernels of sizes of 3*3 and 1*1. The outputs of 13*13*512 dimension, 26*26*768 dimension and 52*52*384 dimension obtained at the 77th, 84th and 94th layers are taken as three features respectively, which are sent into the system after dimension reduction. In the YOLO layer, the final weight model is obtained through the training of three scales. Finally, the marked image of human fall test is output. The fall detection algorithm is introduced by Lu et al. The Image-Based Fall Detection System (IFADS) algorithm is designed to detect fall events based on the frames captured by a camera. It focuses on tracking the posture state of the person in every frame and fall events will be declared when any suspicious posture changes were detected. It is designed for the detection of fall in real-time video and can be integrated into any of the surveillance cameras. It involved a combination of object detection and a fall detection algorithm. IFADS compares the human's posture states frame by frame and get track of the posture states in ever y frame. The IFADS algorithm included a process of person detection and fall detection and carried out by a different algorithm. The YOLO algorithm used in this system is to detect a moving person and to draw a bounding box surrounding the detected person to get track of the person's movement. The fall detection algorithm will then be calculated based on the tracked person's bounding box. The YOLOv3 which is the enhanced version of the YOLO algorithm was chosen for the proposed system due to the higher accuracy and higher fps to detect an object in real-time.

The efficiency testing is done to test the frame rate per second and the confidence value of the YOLOv3 algorithm in detecting person. The input image is resized into different resolution before sending into the YOLOv3 algorithm. The resolution may affect the confidence and performance of the real-time tracking. Therefore, this test was carried out to identify the resolution with highest efficiency. The evaluation result for the confidence of the YOLOv3 object detection algorithm in detecting person in different resolution. The results revealed the image that is resized to 224 x 224 resolution has the highest confidence in classifying object. This means that the YOLOv3 object can detect object in this resolution more accurately. The efficiency with highest rate is more efficiency and will be chosen to be implement into the fall detection system. Equation to calculate the efficiency rate: Efficiency rate = confidence * fps.

STRUCTURE OF DARKNET-53

At present, target detection tasks based on deep learning mainly include two types of detection methods: one is a two stage detection model based on region recommendation to extract target candidate regions, such as Fast-RCNN; the other is a one-stage detection model based on regression thought, such as YOLO, SSD, etc. Although the accuracy of one-stage detection model is slightly inferior, it can achieve real-time detection better in detection speed. Human fall is a sudden situation. It is necessary to detect the fall target accurately at the fastest speed. Therefore, this paper uses the YOLOv3 algorithm which takes both speed and detection accuracy into account. This network is mainly composed of a series of 1*1 and 3*3 convolution layers, because the network has 53 convolutions, it's called Darknet-53.

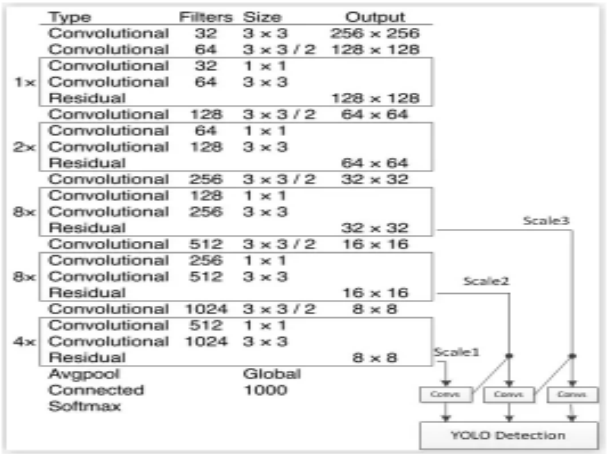


Fig.8. Darknet-53 Network Structure.

DARKNET RESIDUAL COMPONENT

The residuals refer to the residuals structure of resent network and set up fast links between some layers, the degradation of deep network is solved, and the network structure can be deeper. The network of Darknet-53 uses 256*256*3 as the input, and the number of 1, 2, 8 in the left column represents how many duplicate residual components. Each residual component has two convolution layers and a fast link.

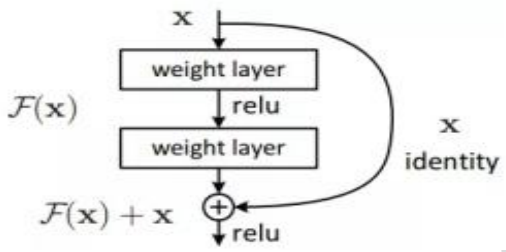


Fig.9. Darknet-53 Network Flowchart

Loss Function

In the training process of the model, the parameters in the network are adjusted continuously, the loss function is optimized to the minimum value, and the training of the model is completed. The loss function of YOLOv3 is mainly composed of four parts: the prediction error of the center point (x, y), the prediction error of the width and height (w, h), the confidence error and the classification prediction error. The loss function of YOLOv3 is mainly divided into three parts.

$$L(O,o,C,c,l,g) = \lambda_1 L_{conf}(o,c) + \lambda_2 L_{cla}(O,C) + \lambda_3 L_{loc}(l,g)$$

Fig.10. Loss Function

The target confidence can be understood as the probability of the target existing in the rectangular box of the target. The target confidence loss and the target category loss adopt the binary cross entropy loss. The target positioning loss adopts the adjustment sum of the difference between the real deviation value and the predicted deviation value.

Feature Extraction

First, the image is scaled to a unified form of 416 in length and width, which is the input of the whole network. Secondly, feature extraction is carried out through the Darknet-53 network, and convolution operations are carried out alternately using convolution kernels of sizes of 3*3 and 1*1. The outputs of 13*13*512 dimension, 26*26*768 dimension and 52*52*384 dimension obtained at the 77th, 84th and 94th layers are taken as three features respectively, which are sent into the system after dimension reduction. In the YOLO layer, the final weight model is obtained through the training of three scales. Finally, the marked image of human fall test is output.

Network Prediction Process:

The human fall detection method is based on the network structure of YOLOv3 algorithm, and the specific detection process is as follows:

1) The fall data in the training set is processed by image preprocessing, and the unified image after processing is used as the input of the whole training network.
2) The processed image is sent to the Darknet-53 network for fall feature extraction.
3) The 77th layer output is extracted as the first feature, and the feature is convoluted and sampled once.
4) The output of the 83rd layer and the 61st layer are spliced to get the second feature, which is convoluted and sampled once.
5) The output of the 93rd layer and the 36th layer are combined to get the third feature.
6) Three features are sent to the YOLO layer for training, and after the training times, the iteration is stopped to generate the final weight model.
7) Input the image of the test set into the same network, call the training weight model to detect the image of the test set, and output the detection results.

ALGORITHM FOR YOLO MODEL

YOLOv3 (You Only Look Once, Version 3) is a real-time object detection algorithm that identifies specific objects in videos, live feeds, or images. YOLO uses features learned by a deep convolutional neural network to detect an object.

YOLO is a Convolutional Neural Network (CNN) for performing object detection in real-time. CNNs are classifier-based systems that can process input images as structured arrays of data and identify patterns between them (view image below). YOLO has the advantage of being much faster than other networks and still maintains accuracy.

FALL DETECTION ALGORITHM

The fall detection algorithm is introduced by Lu et al. The Image-Based Fall Detection System (IFADS) algorithm is designed to detect fall events based on the frames captured by a camera. It focuses on tracking the posture state of the person in every frame and fall events will be declared when any suspicious posture changes were detected. It is designed for the detection of fall in real-time video and can be integrated into any of the surveillance cameras. It involved a combination of object detection and a fall detection algorithm. IFADS compares the human's posture states frame by frame and get
track of the posture states in ever y frame. The IFADS algorithm included a process of person detection and fall detection and carried out by a different algorithm.

EVALUATING INDICATOR

The evaluation indicator selected in this paper is the mainstream mAP (mean average precision) indicator. The mAP indicator is the average value of all target category AP (average precision), and the AP of each target category are different recall values (including 0 and 1). Next, select the maximum precision when it is greater than or equal to these recall values, take recall value as the independent variable, the maximum precision under this recall as the dependent variable, draw the curve, and the area under the curve as the AP of this target. Recall rate is used to describe the missed detection rate of identifying a target, and precision rate is used to describe the accuracy rate of identifying a target. The calculation method is as (2) (3) follows. True positive is the number of correctly recognized target classes, false positive is the number of incorrectly recognized target classes, false negative is the number of incorrectly recognized or recognized target classes.

$$recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

Fig.11. Recall & Precision

MODEL TRAINING

In this paper, Ubuntu 18.04, python programming, tensorflow open source framework 1.14, keras open source framework 2.2.4, cuda10.0 cudnn7.4, video card 2070s are selected as the model training environment. Model training is the process of parameter fitting in the model. In this paper, supervised learning is used to calculate the error between the network output score and the expected score by loss function. The obtained error is used to modify the internal parameters of the network to reduce the error. In the whole training process, batch random gradient descent method is used to optimize the loss function, and 200000 batches are trained in total. The initial learning rate is set to 0.001, the weight attenuation value is set to 0.0005, the batch size is set to 16 and the average loss tends to 0, indicating that the training has reached convergence.

TEST RESULT

The performance comparison of existing popular target detection algorithms. The test result on COCO dataset shows that YOLOv3 algorithm has higher map and faster inference time. The effect comparison between Darknet-53 and other networks on the ImageNet dataset is shown in Table I. It can be seen that the effect of Darknet-53 is similar to that of ResNet-152, but it is twice as fast.

SYSTEM ARCHITECTURE

System architecture is the conceptual model that defines the structure, behavior, and more views of a system. An architecture description is a formal description and representation of a system, organized in a way that supports reasoning about the structures and behaviors of the system.

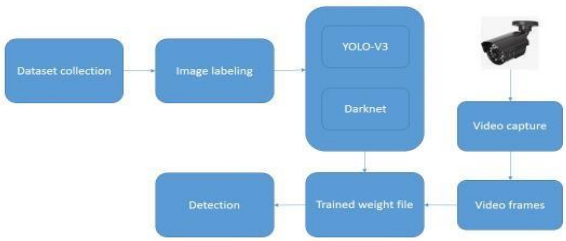


Fig.12. Block diagram of SYSTEM ARCHITECTURE

RESULTS

Fig.13. Samples of collected DATASET

Fig.14. shows the Man before Fall

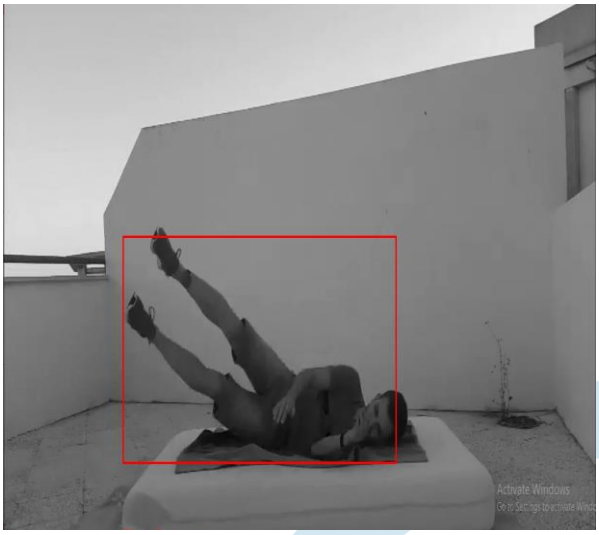


Fig.15. Shows the Man after Fall Detection

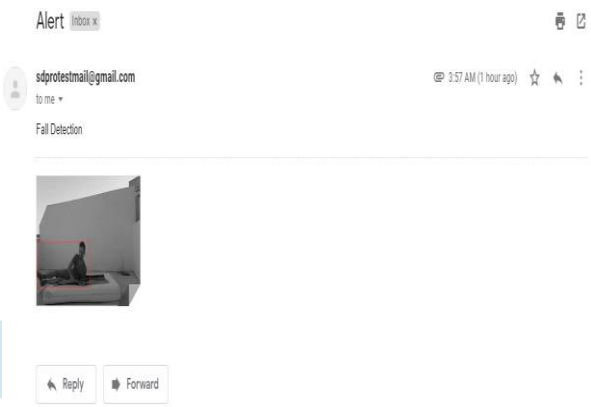


Fig.16. Shows the Mail output

CONCLUSION

In this paper, the fall detection system with mail alert has been implemented. Transfer learning method based training process reduces the training time of the proposed method. The loss function of the proposed method is 0.245. Growing demand on services oriented to elderly makes justified the development of improved system to help elderly live longer in their home increasing their quality of life. The product presented here represents an important step beyond the actual state of the art in services to elderly. Indeed, the service offers complete activity monitoring, automatic fall detection and user localization on a small autonomous mobile module both for indoor and outdoor use. The system, composed by a mobile module worn by the user and a call centre to analyze and save the information, has been developed as easy to use and reliable, and final user requirements have been taken into account on every stage of the development. The first tests and validations, both realized in laboratory conditions as well as with final users (elderly in gerontology Centre) show that the system meets well the requirements, is reliable (above 90 % of fall detection) and well accepted by final users.

REFERENCES

[1] W. Li, X. Hu, R. Gravina, and G. Fortino, "A neuro-fuzzy fatigue-tracking and classification system for wheelchair users," *IEEE Access*, vol. 5, pp. 19420–19431, 2017.

[2] Y. Liu, L. Zhao, S. Zhang, and J. Yang, "Hybrid resolution network using edge guided region mutual information loss for human parsing," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1670–1678.

[3] J. P. F. E. Ocampo, J. A. T. Dizon, C. V. I. Reyes, J. J. C. Capitulo, J. K. G. Tapang, and S. V. Prado, "Evaluation of muscle fatigue degree using surface electromyography and accelerometer signals in fall detection systems," in Proceedings of the 2017 *IEEE* International Conference on Signal and Image Processing Applications (ICSIPA), pp. 21–26, *IEEE*, Kuching, Malaysia, September 2017.

[4] S. Mao, S. Zhang, and M. Yang, "Resolution-invaraint person re- identification," in *Proc. Int. Joint Conf. Artificial Intell.*, 2019, pp. 883–889.

[5] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 184–199.

[6] A. Leone, G. Rescio, A. Caroppo, and P. Siciliano, "An EMGbased system for pre-impact fall detection," in Proceedings of the 2015 *IEEE Sensors*, pp. 1–4, *IEEE*, Busan, South Korea, November 2015.

[7] A. Leone, G. Rescio, A. Caroppo, and P. Siciliano, "An EMGbased system for pre-impact fall detection," in Proceedings of the 2015 *IEEE Sensors*, pp. 1–4, *IEEE*, Busan, South Korea, November 2015.

[8] Y. Zhang, P. Li, X. Zhu et al., "Extracting time-frequency feature of singlechannel vastus medialis EMG signals for knee exercise pattern recognition," *PLoS One*, vol. 12, no. 7, Article ID e0180526, 2017.

[9] W. Wu, J. Fong, V. Crocher et al., "Modulation of shoulder muscle and joint function using a powered upper-limb exoskeleton," *Journal of Biomechanics*, vol. 72, pp. 7–16, 2018.

[10] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[11] M.-A. Dragan and I. Mocanu, "Human activity recognition in smart environments," in Proceedings of the 2013 19th *International Conference on Control Systems and Computer Science*, pp. 495–502, Bucharest, Romania, May 2013.

[12] F. M. Hasanuzzaman, X. Yang, Y. Tian, Q. Liu, and E. Capezuti, "Monitoring activity of taking medicine by incorporating RFID and video analysis," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 2, no. 2, pp. 61– 70,

2013.

[13] L. Mo, F. Li, Y. Zhu, and A. Huang, "Human physical activity recognition based on computer vision with deep learning model," in Proceedings of the 2016 *IEEE International Instrumentation and Measurement Technology Conference Proceedings*, pp. 1–6, Taipei, Taiwan, May 2016.

[14] M. Al Ameen and K. S. Kwak, "Social Issues in wireless sensor networks with healthcare perspective," *International Arab Journal of Information Technology*, vol. 8, no. 1, pp. 52–58, 2011.

[15] A. Fleury, N. Noury, and M. Vacher, "Supervised classification of activities of daily living in health smart homes using SVM," in Proceedings of the 2009 Annual International Conference of the *IEEE Engineering in Medicine and Biology Society*, pp. 6099–6102, Minneapolis, MN, USA, September 2009.

[16] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: a review," *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–43, 2011.

[17] K. Aminian and B. Najafi, "Capturing human motion using body-fixed sensors: outdoor measurement and clinical applications," *Computer Animation and Virtual Worlds*, vol. 15, no. 2, pp. 79–94, 2004.

[18] S. H. Roy, M. S. Cheng, S.-S. Chang et al., "A combined sEMG and accelerometer system for monitoring functional activity in stroke," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 6, pp. 585–594, 2009.

[19] M. N. Nyan, F. E. H. Tay, A. W. Y. Tan, and K. H. W. Seah, "Distinguishing fall activities from normal activities by angular rate characteristics and high-speed camera characterization," *Medical Engineering & Physics*, vol. 28, no. 8, pp. 842– 849, 2006.

[20] A. A. Adewuyi, L. J. Hargrove, and T. A. Kuiken, "An analysis of intrinsic and extrinsic hand muscle EMG for improved pattern recognition control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 4, pp. 485–494, 2016.