Convolutional Neural Network Based on Dynamic Motion and Shape Variations for Elderly Fall Detection

Chadia Khraief, Faouzi Benzarti, and Hamid Amiri

***Abstract*—Fall detection became a major concern especially for elderly who lives alone at home. Unexpected situations might happen that influence their health, security and well-being. The development of an intelligent surveillance system is required to alleviate the negative effects of unforeseen circumstances and assisting the elderly in independent living. Currently, convolutional neural network has been successfully used for solving various computer vision tasks, such as object detection and recognition. In this paper, we propose a new vision system for elderly fall detection based on new two stream convolutional neural networks. First, human silhouette is extracted based on background subtraction and person recognition. Second, history of binary motion image HBMI is fed into the first stream characterizing the human shape variations. The second stream is based on amplitude and orientation of optical flow defining the velocity and the direction of the human motion. The system classifies fall events using score fusion schema. Transfer learning is performed to deal with the small amount of fall datasets. Our final network outperforms state-of-the-art results on standard fall datasets.**

***Index Terms*—Elderly people, fall detection, smart home, deep learning, CNN, transfer learning, video surveillance, motion and shape variations, optical flow.**

I. Introduction

With this fast growing population of seniors over the world, more and more elderly people want to live alone [1]. Unfortunately, unexpected situations might happen suddenly to elderly people. These circumstances influence their health, security and well-being.

Falls became one of the major causes of serious injuries and even mortality to elderly people [2]. Nearly one-third of adults aged over 65 reports a fall every year in United States and the annual medical cost of falls is about $31 billion [3]. The problem is aggravated by the fact that 12 million seniors in the US live alone [4]. They can't alert any one for help particularly if they were unconscious. It was reported that 50% of the elderly who lay on the floor for more than one hour after falls died within six months after the accident [2]. Moreover, fear of falling can limit the activities of old people and can result in social isolation, depression and helplessness [4]. In other hand, the long-term nursing care at home is very expensive [3]. Therefore, falls can be considered as one of the most important problems that prevent the elderly people from leading an independent life. Consequently, it is important to develop an intelligent system for monitoring and assisting elderly people in order to help them living safely and independently in their home. Automatic and reliable fall detection system is a primordial step towards smart home development and the healthcare of elderly people.

Many approaches are developed in order to automatically detect falls. They can be classified into wearable sensors based methods and vision-based methods. In the first group, we can set as example accelerometer [5], gyroscope, smart watch [6] or fusion of many wearable sensors. Unfortunately, elderly people often forget to wear or to recharge them. In the second group, different cameras are installed at home such as RGB, thermal, depth cameras. These sensors give rich information about the human environment without being worn or recharged. But, the development of these systems is very challenging due to light changing, illumination, occlusion, shadows, etc. Many vision-based fall detection methods are proposed [7]. They are based early on extracting the most discriminating hand-crafted features followed by the best classifier such as SVM or KNN. However, it is difficult to take into account all the complex fall models as well as the various daily life activities when developing such as system.

Recently, deep learning has been effectively used for solving various computer vision tasks, such as object detection and activity recognition. In this paper, we resort to deep learning architecture, which has demonstrated its performance over traditional methods based on hand-craft features. We have followed the same principle of the Two-stream Convolutional Networks method [8], a state-of-the-art action recognition method, by using two streams for fall detection. The architecture proposed by [8] for action recognition is based on two separate recognition information: spatial information and temporal information, which are then combined by class score fusion. The spatial information performs action recognition from entire still frames without background subtraction, whilst the temporal information is designed for motion extraction within video sequences represented by optical flow displacements.

In this paper, we propose new two stream convolutional neural networks for fall detection. The first stream is based on dense amplitude and orientation optical flow rather than inputting optical flow vertical and horizontal displacements. In the literature, there are frequently used to characterize the motion information in many handcrafted-based features [9]. But, they are not yet used with a CNN learning method. More exactly, our method captures the velocity of the movement based on the amplitude and the motion direction based on the orientation. The second stream doesn't use the still frames as

Manuscript received March 28, 2019; revised October 10, 2019.

C. Khraief, F. Benzarti and H. Amiri are with the Signal, Image and Technology Information Laboratory LR-SITI, National Engineering School of Tunis (ENIT) of University of Tunis El Manar, Tunis (e-mail: chadiaKhraief@gmail.com, benzartif@yahoo.fr, hamidlamiri@gmail.com).

doi: 10.18178/ijmlc.2019.9.6.878

input of spatial network as [8] but a slacked of human silhouettes, which are combined after background subtraction and person recognition. This stream captures the shape variations and gets the pertinent posture information.

Training our method requires a large number of datasets, otherwise it would be over-fitting. However, there are so few video datasets. In this paper, transfer learning is applied from multiple datasets to surmount the insufficiency of training fall detection datasets.

Our contributions are third-fold:

- First, this paper proposes the first fall detection method that uses shape and motion with CNN, to the best of our knowledge.
- Second, we design a new two-stream networks structure that analyzes both the motion and the shape variations of elderly. It is based on stacked human silhouette images for the shape stream. The second stream is based on orientation and magnitude optical flow images. Some action can be recognized from static image alone as phoning but some others actions, as fall, can't be identified using a single frame. That's why, we have considered stacked images for motion and shape variations analysis.
- Third, we demonstrate that transfer learning improve the fall detection performance. This not only speeds up the system but also contribute to a high recognition rate.

The rest of the paper is organized as follows. Section II reviews previous work on vision based fall detection methods. Section III details our proposed method. Section IV presents the experimental results compared to state-of-the-art methods. Section V concludes our paper and future work is suggested.

II. Related Work

Various vision-based approaches are proposed to detect falls. They can be divided into two categories based on the features used: hand crafted or deep learned:

Handcrafted features: Searching more efficient features is one of the vital interests in fall recognition. Human head position and center of mass velocity are extracted by [10] using depth images. False alarm is generated especially in sitting down posture due to fast head movement. Human fall detection is proposed by [11] based on subject position and velocity fetaures. Albawendi *et al.* [12] classified fall events based on tMNI, projection histogram and angle of fitting ellipse. Rougier *et al.* [13] recognize falls based on the motion history image (MHI) and human shape changes. The ratio and differences of width and height of the bounding box surrounding human are used by Liu *et al.* [14] as shape features to classify postures using k-nearest neighbor classifier. The silhouette area of the elderly person and block matching for motion estimation are exploited by Gnouma *et al.* [15] to classify fall events. Kamal *et al.* [16] extracted the person silhouette based on background subtraction technique and then a set of features were measured such as the vertical velocity of the head, area, height/width ratio, orientation. Khraief *et al.* [17] proposed vision-based fall detection method for elderly people using body parts movement and shape analysis.

Deep learned Features have been used to recognize fall actions based on huge quantity of training videos. Wang *et al.* [18] extract features from color images using PCAnet and then applied a SVM to classify activities. They proposed another method [19] for fall detection based on combination of Local Binary Pattern (LBP), Histograms of Oriented Gradients (HOG) and features learned from a Caffe neural network. Adrian *et al.* [20] applied a convolutional neural networks based on Optical flow only for elderly fall detection. Hwang *et a*l. [21] proposed a deep learning approach in order to maximize the accuracy of fall detection systems using depth cameras by applying the data augmentation technique and 3D-CNN model. Doulamis *et al.* [22] proposed an adaptative deep learning schema to distinguish humans from the background and to deal with dynamic variations of environment such as shadows and illumination. Lu *et al.* [23] proposed three-dimensional convolutional neural network (3D CNN) to extract motion feature from temporal sequence. To further locate the region of interest in each frame, a LSTM (Long Short-Term Memory) based spatial visual attention scheme is incorporated. In the same way, [24] propose an attention guided LSTM model fall events classification. We can conclude that the motion and shape information are less exploited with current state-of-the-art fall recognition methods based on deep learning despite its significant contribution in fall recognizing as a handcrafted feature.

Motivated by the success of deep learning methods and the importance of motion and shape information, we propose two-stream convolutional neural networks that combine complementary information. The first stream is based on stacked human silhouette after background subtraction and human recognition. The second stream is based on multi frame optical flow amplitude and orientation optical flow. In fact, many works in action recognition field demonstrate that abnormal event can be effectively detected based on the amplitude and the orientation of optical flow. Chaudhry *et al.* [25] proposed the Histogram of Oriented Optical Flow (HOOF) for event recognition. Benabbas *et al.* [26] extract the optical flow features globally from each video frame and then grouped into blocks of a certain size and each block is normalized using directional maps. Movement orientation is used to recognize events such as running, walking, assembling and spreading. Colque *et al.* [27] capture not only the orientation, but also the magnitude of the flow vectors, which provide information regarding the velocity of the moving objects and improves considerably the representation of the normal event. This novel feature descriptor is called Histograms of Optical Flow Orientation and Magnitude (HOFM). Optical flow amplitude and orientation are used to characterize motion information as handcrafted-based features not as deep features.

Inspired by the effectiveness of these features, we propose a novel temporal stream based on amplitude and orientation optical flow named AOOF Stream. Our proposed method will be more detailed in the next section.

III. Proposed Method

An abnormal activity is characterized by the presence of unusual movements, irregular shape or both of them. Consequently, we apply two stream to train convolutional

neural networks. The first stream is used to describe the shape variations after background subtraction in order to get only the pertinent information. The second stream is used to capture the motion information based on amplitude and orientation optical flow images. These inputs are fed into our deep learning architecture. Various architecture has been proposed in computer vision in the recent years such as AlexNet [28], VGG-16 [28], ResNet [29]. We adopted the VGG-16 net [28] for our task, motivated by its high accuracy obtained in other related domains.

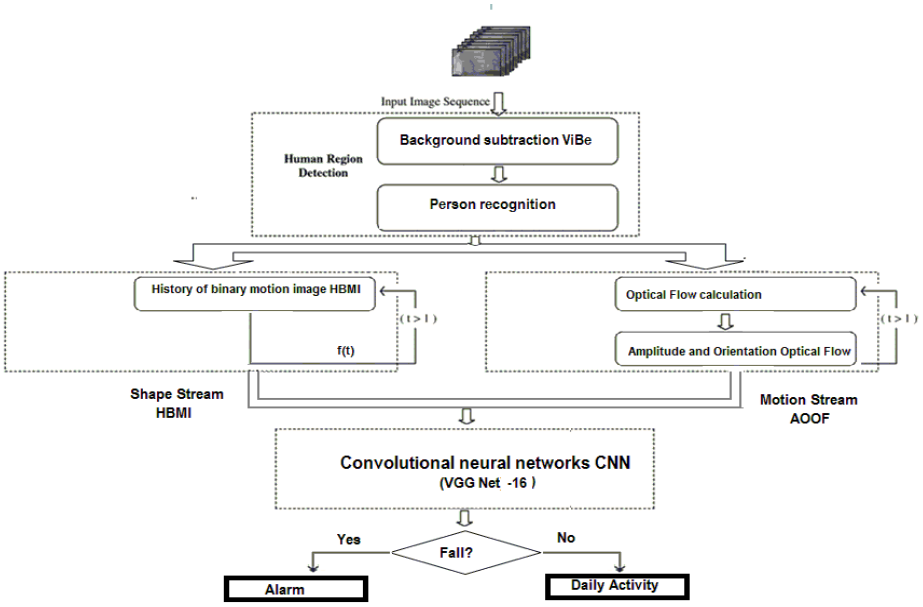

Fig. 1. Proposed method flow.

TABLE I: COMPARISON OF OUR METHOD WITH HAND-CRAFTED FEATURES BASED METHODS

Performance Measures	Proposed method	MHI based approach [13]	Yun *et al.* approach [37]	Chua *et al.*Approach [36]
Sensitivity (%)	99.2	85.7	98.55	90.5
Specificity (%)	99.5	80	95.84	93.3

TABLE II: COMPARISON OF PROPOSED METHOD WITH STATE OF ART APPROACHES

Performance Measures	Proposed method	Adrian *et al*. approach [20]	Wang *et al.* approach [19]	Wang *et al*. approach[18]
Sensitivity (%)	99.2	99	93.7	89.2
Specificity (%)	99.5	96	92	90.3

First, human silhouette is extracted based on our method [30] that combine background subtraction and person recognition method. Extracted silhouettes will be stacked for each posture and used to construct the first shape stream based CNN. In the same time, human motion will be generated based on staked optical flow images and it will construct the second steam. Finally, a score fusion will be used in order to classify fall events based on shape and motion variations. Each step will be explained in the sections bellow.

A. Human Silhouette Extraction

The first step of our proposed algorithm was foreground segmentation to distinguish moving regions from the background and to classify the foreground as a person or non-person.

There are innumerous background removal algorithms proposed in the literature and we have chosen ViBE [31] due to its good performance for surveillance videos and its capability of online adapting the background model. It is divided into three steps, model initialization, foreground detection, and model updating. The first background frame is initialized only with the first frame of the video. Background models are made of 20 background samples for each pixel. Then each pixel is labeled as foreground or background belongs to its history value. The last step is to select updatable neighbor pixels which are determined as the background pixel randomly. Background samples are selected randomly to update the model while other samples are discarded.

To ensure that the foreground region encloses only humans, we refine the foreground segmentation and eliminate the most invalid regions by using morphological operations including erosion and flood filling to fill the holes in this region. Then, connected components were computed and if its size is less than the minimum human size, it will be deleted from the foreground else it will be considered as human region. Therefore, a foreground region containing only a human object was obtained.

To further perform our method, we have combined it to person detection method of Dollar *et al* [32] in order to classify regions into human or not human. Dollar *et al*. [32] was chosen since it presents a good trade of between accuracy and speed. This method uses feature approximation for the features at nearby scales by computing the feature at one scale. Haar-like feature are computed over multiple channels including LUV color channels, grayscale and gradient magnitude.

Finally, detection is validated based on the percentage of foreground pixels inside the bounding box of each person

detected. If this value is below a threshold, the detection is rejected as a false positive.

B. Proposed Convolutional Neural Networks Model

Our model is based on two stream convolutional neural networks. First, we will describe the network architecture and then the two proposed streams.

1) Network architecture

Network architecture plays a very important role in the performance of a deep learning model. The original VGG-16 contains five convolutional block layers that use filters with size 3 × 3 each followed by max Pooling with filter size 2×2, a stride of 2 and Relu activation. Two fully-connected layers and Softmax activation produce the final classification output.

We have changed the classifier layer in order to get two output 'fall' or 'not fall'. With score fusion of the two streams, we have got the final prediction result.

In addition, we have replaced the input layer of VGG-16 net to treat effectively the shape variations and the dynamic motion.

TABLE III: COMPARISON OF OUR FUSION PROPOSED METHOD WITH OPTICAL FLOW STREAM AND RGB STREAM

Performance Measures	Proposed Fusion (RGB and Optical Flow stream)	Optical flow Stream	RGB Steam
Specificity (%)	99.5	96.7	97.3
Sensitivity (%)	99.2	98.1	98.5

TABLE IV: COMPARISON OF PROPOSED METHOD WITH STATE OF ART APPROACHES FOR URFD DATASET

Performance Measures	Proposed method	Adrian *et al.* [20]	Kwolek *et al.* [35]	Nizam *et al*. [38]
Sensitivity (%)	100	100	100	96.67
Specificity (%)	92.5	92	80	82.5

2) Shape stream HBMI

The first stream aims to extract the shape deformations after human silhouette extraction in order to get only the accurate posture variation. It is fed by history of binary motion image HBMI that is the combination of many consecutive binary silhouettes into one image in order to model human action. It is calculated as follows (1):

$$HBMI(X,Y) = \sum_{t=0}^{n} f(t) M_{xy}(t) \quad (1)$$

where $M(t)$ is the binary image containing the human silhouette only without background and $F(t)$ is the weight function that gives more higher weight to recent frames.

3) Motion stream AOOF

The second stream is used to describe the velocity and the direction of the human motion variations based on amplitude and orientation of optical flow. First, we have calculated the dense optical flow using Gunnar Farneback method [33]. This method uses quadratic polynomials that give us the local signal model expressed in a local coordinate system such that:

$$\mathcal{F}(x) \sim x^T Ax + b^T Ax \quad (2)$$

where A is a symmetric matrix, b a vector and c a scalar.

Optical flow generate vertical and horizontal components (dx, dy) for each image based on the optical flow constraint equation (OFCE) expressed as follow:

$$I_x MV_x + I_y MV_y + I_t = 0 \quad (3)$$

where MV_x, MV_y represent the optical flow vectors (δx, δy) and (Ix, Iy, It) represent the derivatives of image intensities at coordinate (x, y, t).

The vertical and horizontal displacements are not used as input to temporal stream as [8], [20]. But, we extract magnitude and orientation from the optical flow images to characterize more efficiently the fall model [8], [20].

Nevertheless, the dense computation of optical flow is computationally expensive [33]. So, to avoid calculating it for each pixel on the image, we use the background subtraction method if the resulting pixel difference is less than d, the pixel is discarded and then we calculate the magnitude OFmag and orientation OFphase information as follows:

$$OFmag = \sqrt{{V_x}^2 + {V_y}^2} \quad (4)$$

$$OFphase = tan^{-1}\left(\frac{V_y}{V_x}\right) \quad (5)$$

A HSV (Hue-Saturation-Value) model based color-coding converts the optical flow vector to an RGB image. At each pixel, the flow direction *OF* phase is coded as hue while the flow magnitude *OFmag* is coded as saturation. The input is composed by 10 stacked images.

C. Network Training and Transfer Learning

Deep learning learns complex features but it needs lots of data from scratch. Public fall detection datasets suffer from the low number of images. Consequently, we have applied transfer learning.

For the shape stream, we have trained VGG-16 net on ImageNet [28]. Learning millions of parameters during this pre-training will be done in order to get generic visual features. These features will be fed to our convolutional neural network. For the motion stream, we trained the network on the UCF101 dataset [34], a large activity recognition dataset with 13,320 video clips generated 101 classes.

IV. EXPERIMENTAL RESULTS

In this section, we present the datasets, evaluation metrics and our results.

We have implemented the architecture using the TensorFlow library, Keras, the Python programming language and MATLAB R2018b. All procedures, training and testing were performed on Dell system with Intel® Core™ i7-7500u - 2.9GHz, 8 GB of RAM, and Windows 10 Home Premium 64-bit operating system.

A. Datasets

We conducted experiments based on two publicly available datasets: the Multiple Cameras Fall (MCF) dataset [13] and the UR Fall Detection (URFD) dataset [35]:

The (MCF) dataset [13] is recorded from eight cameras mounted on the walls and contained 24 scenarios of simulated falls and normal daily activities such as sitting on a chair, walking, crouching, etc. Fig. 2 shows some frames from the MCF dataset.

Fig. 2. MCF Dataset . Top row: human falls. Bottom row: daily life activities.

The UR Fall Detection (URFD) [35] is captured by two Kinect sensors and contains frontal and overhead video sequences as illustrated by Fig. 3. The dataset contains 70 sequences resuming 30 falls and 40 activities of daily life in front view. Two types of falls were defined that are from standing position and from sitting on the chair.

Fig. 3. URFD Dataset . Top row: RGB human activities. Bottom row: Depth daily life activities.

B. Evaluation Metrics

We used two criteria widely used to evaluate fall detection systems. The sensitivity (Se) determines the capacity of the method to classify fall activities correctly and specificity (Sp) is its capability to distinguish daily life activities properly.

They are formulated as bellow:

$$\mathrm{Se} = \frac{TP}{TP+FN} \tag{6}$$

$$\mathrm{Sp} = \frac{TN}{TN+FP} \tag{7}$$

where

- TP (True Positives) is the number of falls correctly detected as a fall by the system,
- TN (True Negatives) is the number of the fall not occurred and correctly detected as a non-fall,
- FP (False Positives) is the number of incorrectly events detected as fall
- FN (False Negatives) is the number of fall events missed by the method and identified as non-fall

C. Results

First using the MCF dataset, we have compared our method to state-of the-art approaches that are based on hand-crafted low-level features such as Yun al. approach [37], MHI based approach [13] and Chua's approach based on three-point representation [36]. As shown in the table 1, the hand-crafted low-level features can usually work well in constraint environment. Yet, they are not universal for all conditions. Consequently, deep learning is treated as a better method to extract high-level features.

Second, we have compared our method to approaches based on deep learning such as Wang's approach based on PCANet [18], Wang's approach based on combination of features [19] and Adrian *et al.* approach [20] based on optical flow only. The results are illustrated in table 2, our approach outperforms their results by using only 10 frames for each stacks of optical flow instead of 30 frames for [20] and the use of the shape stream instead of motion stream only.

Third, we compare the results of each stream and the result of fusion of the two streams. The Table III proves that the fusion of two streams improves the classification result then using each stream independently. The performance of shape stream is higher compared to motion stream, which is reasonable because the motion stream only captures actions while the shape stream takes into consideration motion and appearance at the same time.

Finally, the evaluation of our method to other vision-based systems that have used the URFD is illustrated by Table IV. The results demonstrate the outperformance of the proposed method compared to the state-of-the-art fall detection methods.

In fact, the dataset contains 30 falls and 40 activities of daily life in front view. The proposed algorithm detected successfully all the 30 fall events. For the daily life activities included in the dataset, the proposed method identified 37 activities and failed to classify only 3 activities as illustrated in the Table V.

TABLE V: RESULTS OF THE PROPOSED METHOD FOR URFD DATASET

Performance Measures	Total	Detected	Missed
Fall events	30	30	0
Other activities	40	37	3

V. CONCLUSIONS

In this paper, we presented an efficient vision based method for elderly fall detection. Our method is based on deep learning architecture that analyzes both the motion and the shape variations. We have compared our method to various state-of-the-art methods and we have got encouraging results.

The key contributions of our work are as follows:

- Human motion detection before fall recognition exclude background thus let our method to better perform on various situations, in indoor or outdoor environment. The use of a stacked human silhouette as in input of shape stream compared with researches that use the entire RGB image without discarding the background. It speeds the method by extraction only features from the human body and diminishes the false predictions.
- A new motion stream is proposed based on amplitude

and orientation of optical flow. This information characterizes better the fall event then the displacements optical flow by calculating the velocity and the direction of the motion.

- Fusion of motion and shape steam makes the system more robust and efficient in many different scenarios.
- Transfer learning is used to overcome the low number of images in fall datasets and learn more generic features.

We can improve our methods by using depth cameras that are insensitive to illumination, invariant to color and texture changes. Second, we can use region-based CNNs (R-CNN) to improve our shape based stream by extracting features not only from the entire silhouette but also from different body shapes.

REFERENCES

[1] O. Kharrat, E. Mersni, O. Guebsi, F. Z. B.Salah, and C. Dziri, "Quality of life and elderly people in Tunisia," *NPG Neurologie - Psychiatrie - Gériatrie*, vol. 17, no. 97, pp. 5-11, 2017.

[2] M. Montero-Odasso, "Preventing falls and injuries and healthy ageing," *Coll P. Healthy Aging*, pp. 133-144, Springer, Cham, 2019.

[3] E. R. Burns, J. A. Stevens, and R. Lee, "The direct costs of fatal and non-fatal falls among older adults — United States," *J Safety Res* pp. 99–103, 2016,.

[4] G. Allali, E. I. Ayers, R. Holtzer, and J. Verghese, "The role of postural instability/gait difficulty and fear of falling in predicting falls in non-demented older adults," *Arch Gerontol Geriatr*, vol. 69, pp. 15–20, 2017.

[5] A. Sucerquia, J. López, and J. Vargas-Bonilla, "Real-life/real-time elderly fall detection with a triaxial accelerometer," *Sensors*, vol. 18, no. 4, 2018.

[6] L. Chen, R. Li, H. Zhang, L. Tian, and N. Chen, "Intelligent fall detection method based on accelerometer data from a wrist-worn smart watch," *Measurement*, vol. 140, 2019, pp. 215-226, 2019.

[7] T. Xu, Y. Zhou, and J. Zhu, "New advances and challenges of fall detection systems: A survey," *Applied Sciences*, vol. 8, no. 3, 2018.

[8] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, 2015.

[9] R. V. H. M. Colque, C. Caetano, M. T. L. D. Andrade, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude and entropy to detect anomalous events in videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 673–682, Mar. 2017

[10] F. Merrouche and N. Baha, "Fall detection using head tracking and centroid movement based on a depth camera," in *Proc. International Conference ICCES on Computing for Engineering and Sciences*, 2017, pp. 29-33.

[11] Y. Nizam, M. N. H. Mohd, and M. M. A. Jamil, "Human fall detection from depth images using position and velocity of subject," *Procedia Computer Science*, vol. 105, 2017.

[12] S. Albawendi, A. Lotfi, H. Powell, and K. Appiah, "Video based fall detection using features of motion, shape and histogram," in *Proc. the 11th PErvasive Technologies Related to Assistive Environments Conference*, pp. 529–536, 2018.

[13] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Fall detection from human shape and motion history using video surveillance," in *Proc. the 21st International Conference on Advanced Information Networking and Applications Workshops*, 2007, pp. 875–880.

[14] C. L. Liu, C.-H. Lee, and P.-M. Lin, "A fall detection system using k-nearest neighbor classifier," *Expert Systems with Applications*, vol. 37, no. 10, pp. 7174–7181, 2010.

[15] M. Gnouma, R. Ejbali, and M. Zaied, "Human fall detection based on block matching and silhouette area," in *Proc. the Ninth International Conference on Machine Vision*, 2016.

[16] K. Sehairi, F. Chouireb, and J. Meunier, " Elderly fall detection system based on multiple shape features and motion analysis," in *Proc. the 2018 International Conference on Intelligent Systems and Computer Vision*, 2018.

[17] C. Khraief, F. Benzarti, and H. Amiri, "Vision-based fall detection for elderly people using body parts movement and shape analysis," in *Proc. the 10th International Conference on Machine Vision*, 2018.

[18] S. Wang, L. Chen, Z. Zhou, X. Sun, and J. Dong, "Human fall detection in surveillance video based on PCANet," *Multimedia Tools and Applications*, vol. 75, no. 19, pp. 11603–11613, 2016.

[19] K. Wang, G. Cao, D. Meng, W. Chen, and W. Cao, "Automatic fall detection of human in video using combination of features," in *Proc. the 2016 IEEE International Conference on Bioinformatics and Biomedicine*, China, December 2016, pp. 1228–1233.

[20] A. Marcos, G. Azkune, and I. Arganda-Carreras, "Vision-based fall detection with convolutional neural networks," *Wireless Communications and Mobile Computing*, 2017.

[21] S. Hwang, D. Ahn, H. Park, and T. Park, "Maximizing accuracy of fall detection and alert systems based on 3D convolutional Neural network: Poster abstract," in *Proc. the Second International Conference on Internet-of-Things Design and Implementation*, 2017, pp. 343–344.

[22] A. Doulamis and N. Doulamis, "Adaptive deep learning for a vision-based fall detection," in *Proc. the 11th Pervasive Technologies Related to Assistive Environments Conference*, 2018.

[23] N. Lu, Y. Wu, L. Feng, and J. Song, "Deep learning for fall detection: 3D-CNN combined with LSTM on video kinematic data," in *Proc. the IEEE Journal of Biomedical and Health Informatics*.

[24] Q. Feng, C. Q. Gao, L. Wang, Y. Zhao, T. C. Song, and Q. Li, "Spatio-temporal fall event detection in complex scenes using attention guided LSTM," *Pattern Recognition Letters*, 2018.

[25] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proc. the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[26] Y. Benabbas, S. Amir, A. Lablack, and C. Djeraba, "Human action recognition using direction and magnitude models of motion," in *Proc. the International Conference on Computer Vision Theory and Applications (VISAPP)*, 2011.

[27] R. V. H. M. Colque, C. A. C. Júnior, and W. R. Schwartz, "Histograms of optical flow orientation and magnitude to detect anomalous events in videos," in *Proc. the 28th SIBGRAPI Conference on Graphics, Patterns and Images*, 2015, pp. 126-133.

[28] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. the 26th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, Nev, USA, December 2012, pp. 1097–1105.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[30] C. Khraief, F. Benzarti, and H. Amiri, "Multi person detection and tracking based on hierarchical Level-Set method," in *Proc. the 10th International Conference on Machine Vision (ICMV)*, 2017.

[31] O. Barnich and M. V. Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp.1709-1724, 2011.

[32] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. the British Machine Vision Conference*, 2010.

[33] G. Farneback, "Two-frame motion estimation based on polynomial expansion," *Image Analysis*, pp. 363-370, 2003.

[34] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human action classes from videos in the wild. CRCV-TR-12-01," November 2012.

[35] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, 2014.

[36] J. L. Chua, Y. C. Chang, W. K. Lim, "A simple vision-based fall detection technique for indoor video surveillance," *Signal, Image and Video Processing*, pp. 1–11, 2013.

[37] Y. Yun and I. Gu, "Human fall detection in videos via boosting and fusing statistical features of appearance, shape and motion dynamics on Riemannian manifolds with applications to assisted living," *Computer Vision and Image Understanding*, vol. 148, pp. 111–122, 2016.

[38] Y. Nizam, M. Mohd, and M. Jamil, "Development of a user-adaptable human fall detection based on fall risk levels using depth sensor," *Sensors*, vol. 18, no. 7, 2018.

Chadia Khraief received the engineer's degree and master degree in computer science from the National Engineering School of Tunis (ENIT) respectively in 2008 and 2011.

Currently, she is pursuing the Ph.D. degree in the National Engineering School of Tunis (ENIT) and she is a member of research group in the Image, Signal and Pattern Recognition SITI Laboratory. Her current research interests include human detection, activity recognition, tracking, moving object detection, active contour model, segmentation, machine learning, pattern recognition, deep learning, image processing and video analysis.

Faouzi Benzarti is a professor in the High School of Techniques and Sciences of Tunis. He received his engineer's degree in electrical engineering from the National Engineering School of Monastir, and his M.S degree in biomedical engineering from the Polytechnic School of Montreal Canada (Ecole Polytechnique de Montréal). He obtained his Ph.D degree in electrical engineering from the National Engineering School of Tunis (ENIT) in 2006.

He is presently a member of research group in the Image, Signal and Pattern Recognition SITI Laboratory. His current researches include: human activity recognition, image de convolution, image in painting, image retrieval, image segmentation, face recognition, video analysis and 3D image reconstruction.

Hamid Amiri received the diploma of electrotechnics, information technique in 1978 and the PhD degree in 1983 at the TU Braunschweig, Germany. He obtained the doctorates sciences in 1993.

He was a professor at the National School of Engineer of Tunis (ENIT), Tunisia, from 1987 to 2001. From 2001 to 2009 he was at the Riyadh College of Telecom and Information. Currently, he is again at ENIT. His research is focused on image processing, video analysis, speech processing and natural language processing.