



Visual Surveillance for Human Fall Detection in Healthcare IoT

Yinlong Zhang , State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China, also Key Laboratory of Networked Control Systems, Chinese Academy of Sciences, Shenyang, 110016, China, and also Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China

Xiaoyan Zheng, School of Information and Control Engineering, Shenyang Jianzhu University, Shenyang, 110168, China

Wei Liang , Sichao Zhang, and Xudong Yuan, State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang, 110016, China, also Key Laboratory of Networked Control Systems, Chinese Academy of Sciences, Shenyang, 110016, China, and also Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China

This article designs a visual surveillance framework for human fall detection. In order to solve the conventional issues in fall detection, such as unsatisfactory feature generalization, low recall rates, and large computational time, we design a model that incorporates the deep convolutional neural network and the aggregated heuristic visual features in detecting the occurrence of falls. First, the convolutional neural network (Openpose model) is utilized to extract human skeleton in the image. Second, the hand-crafted spatial features, such as the angle of human shank inclination, are aggregated to determine the fall presence. It should be noticed that our fall detection method has been integrated to healthcare Internet of Things (IoT) video surveillance architecture, which has multiple graphic processing unit groups to perform real-time monitoring and alarming for the elderly in need. The experimental results prove that our method is able to accurately distinguish fall and nonfall activities with a competitive false-alarm rate.

The health of the elderly has become a major concern in the modern society. The statistics provided by the United Nations shows that the aging population is gradually increasing.^{1,2} It also shows that elderly falls have become a serious safety hazard.³ More than 30% of the elderly over 65 years fall at least once a year, which could be attributed to trip hazards, lighting changes, physical, or neurological decline of the subject, etc. Falls can affect the health of the elderly considerably, or cause skin and soft tissue scratches, abrasions, and even long-term disability. Furthermore, long-term staying on the floor or the delayed medical assistance, somehow, increase the risk of both physical and psychological complications. Therefore, a reliable fall detection should be

designed to timely and reliably detect the elderly falls in the healthcare environments. The aim of this article is to design a visual surveillance framework for human fall detection. The surveillance could cover a large area with multiple cameras fixed at the specific locations.

The state-of-the-art fall detection methods could be categorized into the following two types: nonvision based and vision based. In terms of nonvision-based fall detection, wearable sensors [e.g., inertial measurement units (IMUs)] are mounted to the human chest and knees, and the measurements are transmitted wirelessly by WiFi, UWB, etc., installed in the surrounding environment to capture the movements of the senior people.⁴ In Seneviratne *et al.*'s work,⁵ the miniaturized IMU is embedded into the eyeglass frame. The daily activities and fall events could be monitored by analyzing the acceleration and angular changes. Unfortunately, these wearable sensors are susceptible to noises and the detection results are not very satisfactory.

1070-986X © 2022 IEEE

Digital Object Identifier 10.1109/MMUL.2022.3155768

Date of publication 3 March 2022; date of current version 4 May 2022.

Comparably, the vision-based fall detection is able to collect the abundant human poses, and extract the human skeletons for recognizing abnormal behaviors. In Shu and Shu work,⁶ multiple cameras are utilized to capture human movements. The fall detector will be triggered when the acceleration of head and the center of body (computed from the image sequences) exceeds the predefined threshold. It is interesting that convolutional neural network (CNN) exerts the powerful abilities in extracting the high-level features especially in complex surveillance context. For instance, Cao *et al.*⁷ presented a multiperson pose estimation using the part affinity fields. The human poses could be robustly and quickly detected in the bottom-up architecture. Despite the merits of vision-based methods, there are still several challenges that need to be solved, which are listed as follows.

1) The extracted features are not discriminative. In fall feature extraction, there exist dark or strong sunlight disturbances, as well as tracking view changes, which will inevitably result in inaccurate fall features in the noisy environments.

2) Fall detection lacks generalization. Although the traditional methods are able to define the fall detection, there still lack meaningful semantic representations of fall features, which will result in model inadequate generalization.

3) The computational resources are limited. The fall detection, sometimes, are unable to be processed in real time due to the limited computational resources in the healthcare scenarios. It will result in delayed or even missing fall alarms.

In order to solve the aforementioned issues, this article presents a novel fall detection method that combines the CNN model and multidiscriminant hand-crafted features. Our method is able to extract and track the human skeletons in the captured image sequence. The human joints could be robustly detected even in the presence of body part occlusions. Unlike the traditional methods that extract the single feature in detecting the occurrence of human falls, our work uses multiple discriminant human fall features, i.e., the angle of human shank inclination, the angle of spine inclination, and the body bounding box height-width ratio. In addition, our method has been integrated in the healthcare Internet of Things (IoT) framework. The captured multichannel video sequences are fed into the cloud processors that have groups of powerful graphic processing units (GPUs). Through this way, the falls in multiple scenes could be detected, which saves favorable computational resources.

The rest of this article is organized as follows. In the "Related work" section, the related works are summarized. In the "IoT Visual Surveillance Framework"

section, our healthcare IoT visual surveillance framework is briefly introduced. In the "Methodology" section, our proposed method is described in detail. In the "Experimental Results and Analysis" section, the developed platform is introduced; the fall detections are extensively analyzed and compared with state-of-the-art. The "Conclusion" section concludes this article. In the "Discussion and Future Work" section, the discussion and future work are introduced.

RELATED WORK

A brief review of the related literatures on human fall detection is presented in this section. The relevant visual surveillance techniques are introduced and compared. By and large, the human fall detection could be categorized into signal-view and multiview computer vision techniques.

In signal-view human fall detections, Alvarez *et al.*² proposed to use multimodal sensing strategy to analyze the behaviors of patients suffering from Parkinson and Alzheimer long-term diseases. It is impressive that a variety of physical signals from diverse sources in health monitoring environments are collected and used to infer the user behavior and context, and trigger the proper actions for improving the patients quality of life. Ren *et al.*⁸ designed a directed graph CNNs to recognize the human action. The residual split structure could help us to avoid the gradient disappearing. Kong *et al.*⁹ designed a three-stream CNNs to classify the fall events. Afterward, the voting is performed on the high-quality spatiotemporal representations of appearance and motion information. In Ozcan *et al.*'s work,¹⁰ the modified histogram of oriented gradients, together with the local gradient binary patterns, are used to train the fall thresholds, which will be compared with the relative entropy changes for fall detection. Although the features are descriptive and discriminative, there are still some false alarms in presence of body bending the waist or picking up the things on the floor. In Leite *et al.*'s work,¹¹ a multichannel fall detection is designed that uses the openpose and object recognition model. It classifies the fall into dynamic and static processes. The multilayer perception and random forest are used to classify the selected features. In Dusmanu *et al.*'s work,¹² the CNN is designed to extract the human joints and derive the human poses. Afterward, the local and global information are jointly used in the coding model for fall detection.

In multiview human fall detections, to deal with the issue of limited surveillance areas using the single camera capturing, a variety of researchers have proposed the distributed surveillance cameras and IoT-sensing strategy for large-scale scenarios (e.g., elderly

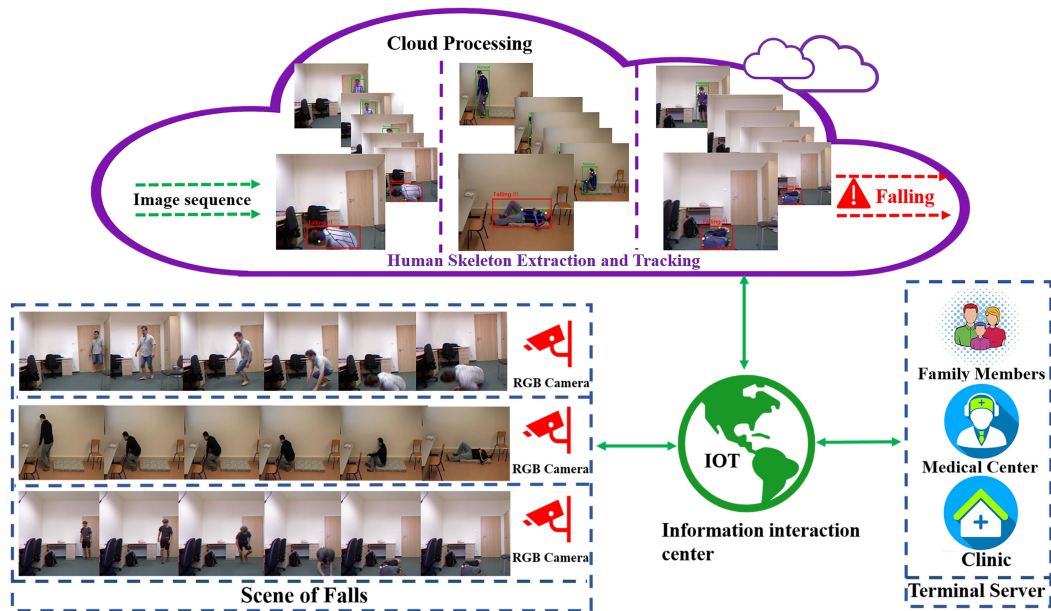


FIGURE 1. Illustration on our proposed fall detection in IoT healthcare visual surveillance framework. The scenes of elderly daily activities are monitored by IP cameras. The captured image sequences are fed into the cloud processing units with multiple GPUs. The human skeleton in the image sequences are extracted and warning alarms will be triggered in presence of elderly falls.

nursing centers). Rougier *et al.*¹³ designed the multicamera framework and employed the context matching techniques in detecting the human falls across several cameras. Wang *et al.*¹⁴ adopted the PCANet to extract the multiview human silhouettes in human posture description and fall detection. However, the multicamera system for large-area human surveillance requires more computational resources, which is impractical for home users and medical clinics to install the expensive equipment with powerful computations. In order to solve this issue, Wang *et al.*¹ proposed the healthcare IoT, which mitigates the pressures of hospitals and medical resources by transmitting the large volume of signals collected on the edge devices to the cloud computing end with powerful computational resources. Yet, the transmission and computation latency are still intractable and there is still a large margin between the practical use and the theoretical surveillance analysis.

IOT VISUAL SURVEILLANCE FRAMEWORK

Inspired by the remarkable progress of IoT healthcare,¹⁵ the fall detection has been used in the IoT visual surveillance framework in this work. As illustrated in Figure 1, the scenes are captured by IP cameras installed at the

corners of surveillance areas. The corresponding image sequences are uploaded and processed in the cloud server and the corresponding fall detection results will be transferred to information center, where the terminals will notify the elderly family members, medical centers, and clinics the occurrence of falls.

Our IoT visual surveillance system consists of the following three parts: scenes captured by IP cameras, image processing, and fall detection in cloud servers, detection results transferred to specified users.

- 1) The captured scenes for fall detection: In order to achieve the multiview fall detection, multiple RGB cameras are installed in various rooms in elderly daily activity ranges to ensure that the elderly have been monitored all the time. The captured image sequences will be uploaded to the cloud processing server.
- 2) The cloud server will process the image sequences. Due to the fact that there are large volumes of video sequences, which are intractable for the local computers to process the results in real time. Comparably, the cloud server has powerful computational resources, which satisfy the real-time human skeleton extraction and fall detection.
- 3) The terminal will upload the human fall detection results to medical centers, which could help the elderly timely.

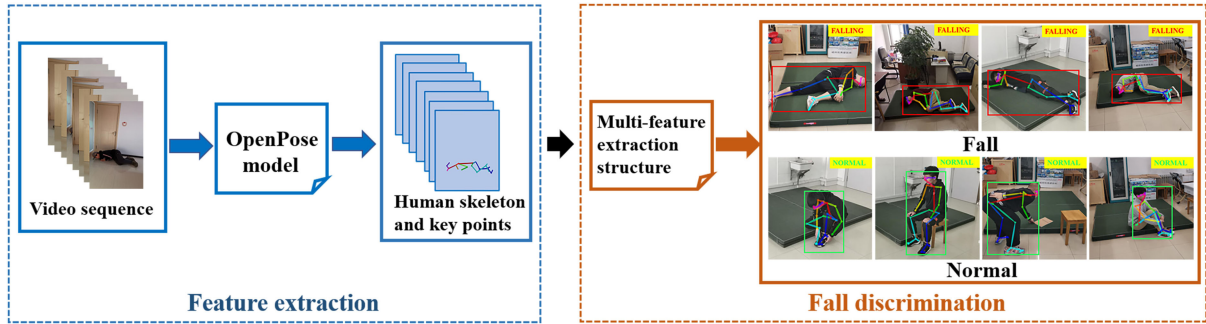


FIGURE 2. Pipeline of our fall detection method. It consists of two parts. In part one, the fall features will be extracted using the openpose model from the video sequences. The human skeleton and key points will be obtained. In part two, the discriminant features (i.e., the angle of human shank inclination, the angle of spine inclination, the height–width ratio of the human bounding box) will be calculated to predict the presence of falls.

METHODOLOGY

The schematic diagram of our fall detection method consists of two parts, i.e., feature extraction and fall discrimination, as shown in Figure 2. In the feature extraction stage, the openpose model⁷ will be used to extract the human features in the captured video sequence. The skeleton and key points of humans in the images will be computed and fed into the fall discrimination stage. The discriminant features (i.e., the angle of human shank inclination, the angle of spine inclination, the height–width ratio of the human) will be calculated given the human skeleton and key point positions in the images. The captured human fall state “Fall” and “Normal” will be given by using the falling criteria, which will be given in the following sections.

Feature Extraction

Before the image feature extraction, we have performed the image preprocessing to tackle the issues of dark and strong sunlight. Besides, the Gaussian smoothing filter and median filter are performed to remove the salt and pepper noise. Afterward, the human feature extraction is implemented based on the multistage CNN model-openpose.⁷ The human key points are extracted in real time by the bottom-up strategy. It uses the first ten layers of VGG-19 network structure to yield the feature map and applies the greedy analysis to encode the whole human skeleton.

The feature extraction model proceeds in two stages. The 2-D confidence maps will be generated to predict the feature positions in the first stage. Meanwhile, the part affinity fields will be used to associate the key points in the second stage. The feature confidence map for human key point $C_{n,m}^*(Q)$ is given in the following:

$$C_{n,m}^*(Q) = \exp\left(-\frac{\|Q - X_{n,m}\|_2^2}{\sigma_{n,m}^2}\right) \quad (1)$$

where m and n represent the human index and key point index, respectively; Q symbolizes each point in the confidence map $C_{n,m}^*$; $X_{n,m}$ is the true position of the n th key point of the m th human. $\sigma_{n,m}$ is used to describe the key point probability distribution. Ideally, the key point corresponds to the maximum in the feature map. Therefore, the n th human key point position $C_n^*(Q)$ could be derived by

$$C_n^*(Q) = \max_m C_{n,m}^*(Q). \quad (2)$$

Furthermore, the part affinity values between the adjoint joints (e.g., arm and wrist, ankle and knee) could be computed by the calculus upon H , which is given in the following equation:

$$H = \int_{t=0}^{t=1} k_b(Q(t)) \cdot \frac{n_2 - n_1}{\|n_2 - n_1\|_2} dt Q(t) \\ = (1 - t)n_2 + tn_1, t \in (0, 1) \quad (3)$$

where n_1 and n_2 symbolize the successive joints; $\|n_2 - n_1\|_2$ means the body length between the joints n_1 and n_2 . $Q(t) \in [n_1, n_2]$ means the body part between n_1 and n_2 . When the point Q lies on the body, $k_b(Q(t)) = (m_2 - m_1)/\|m_2 - m_1\|_2$, else $k_b(Q(t)) = 0$. Each calculus H between the key point and its successive joint will be computed and the correct link between the successive joints will be obtained by selecting the link with the maximum H . Through this way, all the joints will be connected and the human skeleton will be grouped.

In this work, we define the human skeleton with 25 joints and label these joints with indexes from 0 to 24, as shown in Figure 3. 0 stands for nose; 1 stands for

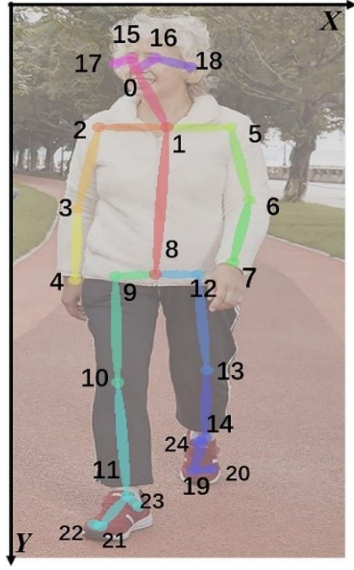


FIGURE 3. Human skeleton and key points. 0 stands for nose; 1 stands for neck; 5 and 2 stand for left and right shoulder; 6 and 3 stand for left and right elbow; 7 and 4 stand for left and right wrist; 8 stands for the waist; 12 and 9 stand for left and right span; 13 and 10 stand for left and right knee; 14 and 11 stand for left and right ankle; 16 and 15 stand for left and right eye; 24 and 23 stand for left and right knee.

neck; 5 and 2 stand for left and right shoulder; 6 and 3 stand for left and right elbow; 7 and 4 stand for left and right wrist; 8 stands for the waist; 12 and 9 stand for left and right span; 13 and 10 stand for left and right knee; 14 and 11 stand for left and right ankle; 16 and 15 stand for left and right eye; 24 and 23 stand for left and right heel.

Fall Detection

In this work, we consider the angle of human spine inclination θ_1 , the angle of right knee inclination θ_2 , the angle of left knee inclination θ_3 , and the height-width ratio of human bounding box R as the discriminant features to judge the presence of falls, as shown in Figure 4. The corresponding joints include the neck (x_1, y_1) , waist (x_8, y_8) , right knee (x_{10}, y_{10}) , left knee (x_{11}, y_{11}) , right angle (x_{13}, y_{13}) , and left ankle (x_{14}, y_{14}) . The vectors between the related joints, i.e., spline vector \vec{S}_1 , right calf vector \vec{S}_2 , and left calf vector \vec{S}_3 are given in the following equation:

$$\begin{aligned}\vec{S}_1 &= (x_8 - x_1, y_8 - y_1) \\ \vec{S}_2 &= (x_{14} - x_{13}, y_{14} - y_{13}) \\ \vec{S}_3 &= (x_{11} - x_{10}, y_{11} - y_{10}).\end{aligned}\quad (4)$$

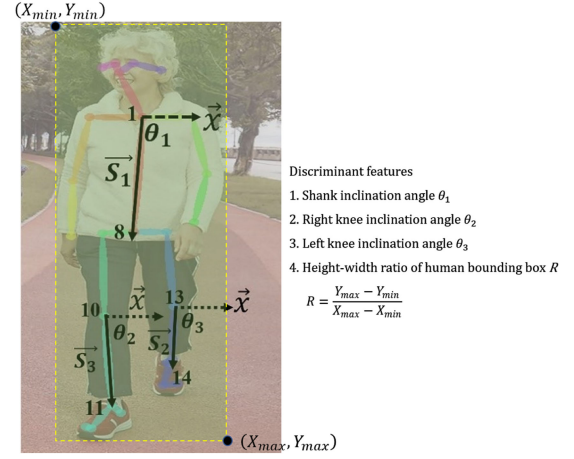


FIGURE 4. Discriminant human fall features, which include the angle of human shank inclination θ_1 , the angle of right knee inclination θ_2 , the angle of left knee inclination θ_3 , the height-width ratio of human bounding box R .

The corresponding angles could be derived by

$$\theta_i = \arccos\left(\frac{\|\vec{S}_i \cdot \vec{x}\|}{\|\vec{S}_i\| \cdot \|\vec{x}\|}\right), i = 1, 2, 3 \quad (5)$$

where \vec{x} is the vector parallel to the ground. The height-width ratio of human bounding box is

$$R = \frac{Y_{\max} - Y_{\min}}{X_{\max} - X_{\min}} \quad (6)$$

where (X_{\min}, Y_{\min}) and (X_{\max}, Y_{\max}) are the bounding box upperleft and lowerright coordinates. $X_{\max} - X_{\min}$ symbolize the bounding width; $Y_{\max} - Y_{\min}$ symbolize the bounding box height.

During the human fall process, the features θ_i and R will be varying. We define the angle thresholds α and β , ratio threshold J to predict the fall occurrence.

When $0^\circ < \theta_i < \alpha$, ($i = 1, 2, 3$), $R < J$ and it lasts for a certain period of time (for example, 10 seconds), our method would detect the presence of falls and warning alarms will be triggered.

When $\alpha < \theta_i < \beta$, ($i = 1, 2, 3$), the human is standing or walking.

When $0^\circ < \theta_1 < \alpha$ and $\alpha < \theta_2, \theta_3 < \beta$ or $\beta < \theta_1 < 180^\circ$ and $\alpha < \theta_2, \theta_3 < \beta$, the human is bending the waist.

It should be noticed that the human fall could occur along the longitudinal direction in the image and the inclination angles may not be estimated due to self-occlusions. Under this circumstance, if the human bounding box R is below J and not varying for a relatively

TABLE 1. Selected candidates for fall tests.

| Candidate | Gender | Age(year) | Height(cm) |
|-----------|--------|-----------|------------|
| 1 | Female | 45 | 162 |
| 2 | Female | 23 | 158 |
| 3 | Female | 22 | 160 |
| 4 | Male | 50 | 175 |
| 5 | Male | 24 | 180 |
| 6 | Male | 22 | 183 |

longer period of time (for example, one minute), our method would also trigger the warning alarms. Although it may increase the odds of false alarms, the recall rate could be guaranteed.

EXPERIMENTAL RESULTS AND ANALYSIS

Experimental Platform

In order to evaluate our method competitiveness, we have compared ours with state-of-the-art datasets, e.g., SDU Fall dataset, Multicam Fall dataset, SIMPLE Fall dataset, CHARFI2012 dataset, LE2I dataset, UPFALL dataset, URFD dataset, to name a few.¹⁶ This work focuses on the URFD dataset for the performance evaluation. In addition, we have developed our own fall datasets. All the tests and analysis have been performed on the cloud processing group, which has Intel Core i9-10900T 4.60-GHz processing unit, four groups of K80 GPUs, and 32-gigabit RAM.

Design and Test

In our fall detection tests, six candidates are chosen. In order to achieve the satisfactory generalization of our method, we have selected the participants with diversities on ages, height, gender, coordination, and balance. The detailed information is listed on Table 1. The participants have been required to perform left and right inclinations for 6 times. The angle between the body and ground has been observed.

Table 2 tabulates the body angles of subjects at the moment of fall occurrence. As can be seen, the angle of subject falls at the left side are no less than 70°. In a similar manner, the angle of subject falls at the right side are less than 110°. Thus, the human fall threshold angles on the left and right sides are 70° and 110°, respectively, i.e., $\alpha = 70^\circ$, $\beta = 110^\circ$, which could be applicable in the elderly.

During the tests, we have also recorded the height-width ratios of each subject at the state of standing, setting, seating, bending the knee, squatting,

TABLE 2. Angles of the tester body at the moment of fall (lean to the left).

| Tester | First | Second | Third | Fourth | Fifth | Sixth |
|---------|-------|--------|-------|--------|-------|-------|
| Tester1 | 75.6° | 76.1° | 75.5° | 77.0° | 74.0° | 75.0° |
| Tester2 | 74.5° | 74.5° | 73.0° | 75.0° | 74.0° | 73.5° |
| Tester3 | 76.5° | 77.1° | 76.0° | 77.5° | 77.0° | 78.5° |
| Tester4 | 73.2° | 72.5° | 72.0° | 72.5° | 71.5° | 72.0° |
| Tester5 | 72.1° | 71.5° | 72.5° | 71.3° | 72.0° | 71.0° |
| Tester6 | 73.9° | 72.3° | 73.1° | 72.4° | 71.5° | 72.0° |

and falling. We have computed the mean values of each subject movement. The height-width ratio will be less than 1.1 when the subject is in the nonfall state. Comparably, the ratio will be less than 0.8 when the subject is falling. Thus, we choose the height-width ratio threshold $J=0.9$. It should be noticed that the threshold values on α , β , and J are the same for both the public datasets and our collected datasets in the following evaluations and analysis.

Fall Detection Tests on the URFD Datasets

In this work, we have chosen the URFD datasets to evaluate the performance of ours against other approaches. The dataset consists of 70 videos (that include 30 falling pieces and 40 pieces of human daily activities) collected by two cameras. The dataset is suitable for our application scenarios since it includes various human indoor activities, which are representative in terms of subject different habits. To fully evaluate the performance of our method, we have used four metrics, i.e., precision, sensitivity, specificity, and accuracy. The comparisons are shown in Figure 5. Ours achieves the competitive results, especially on the accuracy (which reaches up to 99%), which could partly prove our method advantages in discriminating the fall and no-fall activities. It can also be observed that both the precision and sensitivity of our method are above 98%, which are greater than the state of the art by a margin of approximately 2% and 1%, respectively, which shows the effectiveness of fall detection. In terms of specificity, ours is 99%, while the others are lower than 95%. The main reason is that ours considers not only the multiple spatial features, but also the geometrical relationship of human body.

We have also analyzed the human fall detection in the presence of occlusions. As shown in Figure 6, there exist complex occlusions, such as tables and chairs. Impressively, ours is able to detect the human

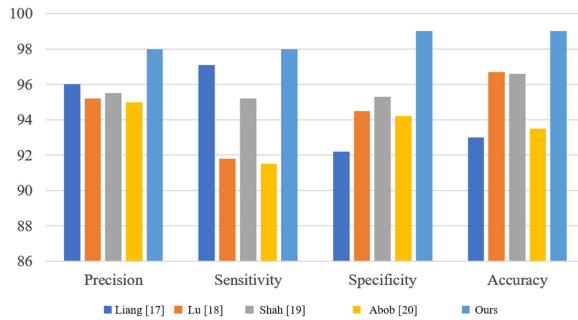


FIGURE 5. Comparisons of ours with state-of-the-art in terms of fall detection precision, sensitivity, specificity, and accuracy on the URFD datasets.

fall and nonfall states accurately; while the others failed to some extent. In Figure 6(b), (d), and (e), the traditional methods failed due to the fact that they only extract parts of the human skeleton as the feature to determine the fall. In Figure 6(c), the fall detection fails because of inadequate segmentation of human and background, which incurs the failures in labeling the human bounding box.

For all the tests, we have run the fall detection on the developed cloud server. The processed videos lasts about 24 seconds and the image resolution is 1920×1080 . As shown on Table 3, the running speed and computational time using ours are 38.2 fps and 24.1 s, respectively. While, the others use relatively longer time. The fastest method is able to run at 18.3 fps, the videos could not be processed smoothly and quickly. It could be attributed to two aspects. On one hand, most of the state-of-the-art used the top-down

TABLE 3. Running speed and computational time between ours and state-of-the-art.

| | Liang ¹⁷ | Lu ¹⁸ | Shah ¹⁹ | Abob ²⁰ | Ours |
|-----------|---------------------|------------------|--------------------|--------------------|------|
| Speed/fps | 13.4 | 15.6 | 14.5 | 18.3 | 38.2 |
| Time/s | 42.3 | 38.5 | 40.8 | 30.5 | 24.1 |

human pose estimation framework (for instance, alpha-pose model), which relatively requires more computational time. On the other hand, the selected features are neither compact in size nor representative in discriminating the fall presence. Through the test, we can see that ours achieve competitive results in terms of computational time and running speed.

Fall Detection Tests on the Collected Datasets

In the tests, we have also collected the fall datasets in the indoor surveillance areas. The subjects are chosen to perform the similar fall and nonfall movements of elderly people. They were required to imitate the motions of elderly people. In the tests, the motions are recorded by five HikVision surveillance cameras, which are deployed at the corner of rooms. The examples of captured scenes are shown in Figure 7(a). The dataset includes the videos of various human motions, i.e., 50 segments of normal walking, 50 segments of human falling, 50 segments of bending the waist, 50 segments of squatting, 50 segments of seating. Five subjects are chosen and each one randomly walks into the separate rooms to complete the designed

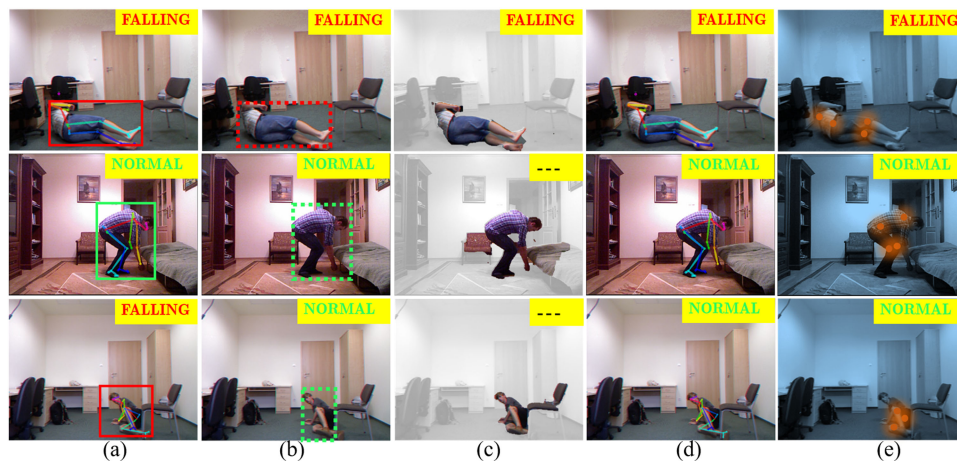


FIGURE 6. Comparisons of ours with state-of-the-art in presence of human occlusions on the URFD datasets. (a) Ours. (b) Liang.¹⁷ (c) Lu.¹⁸ (d) Shah.¹⁹ (e) Abob.²⁰

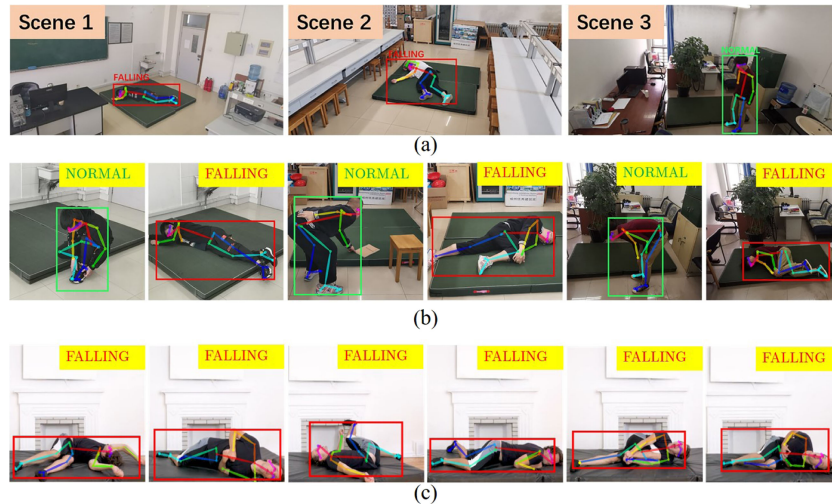


FIGURE 7. Examples of people fall detection results. (a) Captured scenes by three surveillance cameras, which are fixed at the corner of rooms. (b) Examples of fall detection results in Dataset1. (c) Examples of fall detection results in Dataset2.

motions. The subject is asked to perform the movements with different speed, amplitude, and postures. To ensure the safety of falling motions, we have put two soft mats on the ground with the size $2 \times 1 \times 0.1$ m. Each video clip lasted about 15–20 s. The collected datasets have been divided into training, validation, and testing subsets. The percentages of training, validation, and testing are 50%, 25%, and 25%, respectively. The test results are illustrated in Figure 7(b). In the test, there exists the challenging scenario that the person falls temporarily and quickly gets up. Our method is able to identify this false fall since it not satisfies the falling lasting period condition.

In order to analyze the human inclination angle and aspect ratio during the falling process, we have plotted the figures on angle and ratio evolutions, as illustrated in Figure 8. It could be observed that when the shank angles (θ_1 , θ_2) and spine angle θ_3 fall under the corresponding thresholds, and the aspect ratio is lower than the threshold 0.9 for 10

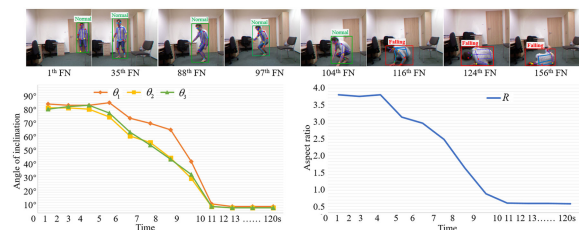


FIGURE 8. Illustration on inclination angle and aspect ratio during the human falling process.

seconds, our method will give that the human falling warning.

The detection results are analyzed, as shown in Table 4. The precision reaches up to 98.5%. In addition, we have also collected the samples of 50 ways of falling, e.g., drunk, stomachache, tightrope, tripping, ankle injury, and shot, as shown in Figure 7(c). Impressively, all of the falls could be detected correctly using our method.

It is noticeable that we have fixed ten cameras in five rooms to capture the presence of human falls. Compared with the single HikVision camera DS-2CD2183G0 (with vertical and horizontal view 53° and 102° , respectively) that could only capture the areas of approximately 15 m^2 our configured multicamera surveillance system could cover up to 150 m^2 areas.

In this work, we have compared the running speed on the local laptop and cloud server. The local laptop is equipped with Intel Core i7-10900 3.60-GHz processor, 16G RAM. The comparisons on the running speed is shown in Figure 9. The local laptop is able to run the fall detection at 20 fps when only one channel from camera captured scenes is connected. Comparably, the cloud server running speed reaches up to 38 fps. They are able to perform the real-time fall detection. However, the running speed degrades over the increase of cameras. When three cameras are connected to the laptop, it is unable to satisfy the real-time fall detection and there exist obvious delayed results. By contrast, the cloud server is able to perform the real-time fall detection at approximately 37 fps, thanks to its powerful GPU groups. It proves that the

TABLE 4. Detection results of our proposed method on the collected fall-detection dataset.

| | Precision | Sensitivity | Specificity | Accuracy |
|----------|-----------|-------------|-------------|----------|
| Dataset1 | 98.5% | 98.2% | 98.4% | 98.5% |
| Dataset2 | 96.1% | 98.7% | 96.5% | 98.3% |

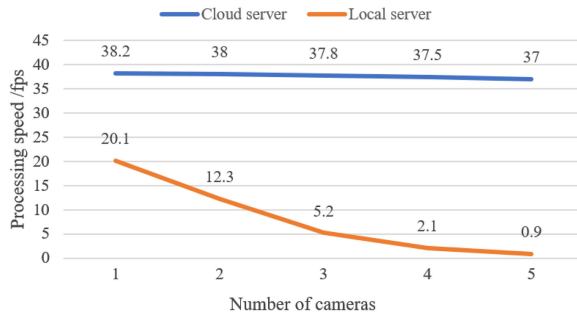


FIGURE 9. Video running speed with the various number of surveillance cameras.

cloud server is more suitable for the large-scale health-care visual surveillance. Besides, the video transferring time from the local camera to the cloud is estimated. The cameras and the cloud server are connected by the network cables. The video (image resolution 1920×1080) is transferred to the cloud with 21 milliseconds latency approximately in average. The computational latency in fall detection is 115 milliseconds in cloud servers, which could satisfy the real-time human fall detection.

In presence of elderly falls detected by the IP cameras, the captured scenes will be transmitted to the family members, medical centers, and clinics. As shown in Figure 10, the family members are able to track the state of the elderly. When the elderly falls, the warning will be issued and appeared on the terminals (i.e., cellphone screen and monitors in the medical centers). The host could make the phone call to the surveillance staff. The monitored video segments could be replayed. The captured scenes could be zoomed in locally. The medical staff could be timely notified and make the emergency calls to the surveillance manager and the related family members. The captured scenes could be locally zoomed for better views. Also, the replay function is integrated in the surveillance software.

CONCLUSION

This article presents a fall detection method using the deep CNN model and hand-crafted discriminant features. It first uses the VGG neural network to extract the



FIGURE 10. Captured scenes are remotely transmitted to the terminals, i.e., cellphone app and medical centers. The medical staff and family members could replay the falling video segments; zoom in and out the scenes locally; give the emergency call to the surveillance staff and the elderly family members.

human skeleton. Afterward, the hand-crafted features, i.e., the angle of human shank inclination, the angle of spine inclination, the height-width ratio of human bounding box are aggregated to detect the occurrence of human falls. Our method has been deployed on the cloud server to relieve the computational burdens. The performance of our detection method has been evaluated on both URFD datasets and the collected datasets. The effectiveness and accuracy have been evaluated by comparing the precision, accuracy, specificity, and accuracy, which are all above 96%.

DISCUSSION AND FUTURE WORK

Our method has been evaluated on both the public datasets and our own collected datasets. Notice that multiple person falls could be detected using our method. However, there exist false alarms or missing alarms due to similar actions, such as human kneeling down, crawling, or severe body occlusions. We will further investigate these scenarios and improve the fall detection and prediction model for that.

In the future, we plan to evaluate our method on other public datasets (such as SDU fall dataset, multi-cam fall dataset¹⁶) and apply to the elderly fall risk assessment for the potential clinical interventions or rehabilitations. Also, the clinical research on the nature and impact of fall evaluation as well as the human walking ability decline and fall risk will be performed for elderly healthcare, on the base of which the research works on other similar activities will be performed using the video-based skeleton extraction methods.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grants 61903357 and 62022088, in part by the Liaoning Provincial Natural Science Foundation of China under

Grants 2021JH6/10500114 and 2020-MS-032, in part by the International Partnership Program of Chinese Academy of Sciences under Grant 173321KYSB20200002, in part by LiaoNing Revitalization Talents Program under Grant XLYC1902110, in part by Young and Middle-aged Science and Technology Innovation Talent Plan of Shenyang City under Grant RC210482, and in part by Guangzhou Science and Technology Planning Project under Grant 202102021300.

REFERENCES

1. K. Wang, Y. Shao, L. Xie, J. Wu, and S. Guo, "Adaptive and fault-tolerant data processing in healthcare IoT based on fog computing," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 1, pp. 263–273, Jan.–Mar. 2020.
2. F. Alvarez *et al.*, "Behavior analysis through multimodal sensing for care of Parkinson's and Alzheimer's patients," *IEEE MultiMedia*, vol. 25, no. 1, pp. 14–25, Jan.–Mar. 2018.
3. Y. Yun and I. Y. Gu, "Human fall detection in videos by fusing statistical features of shape and motion dynamics on Riemannian manifolds," *Neurocomputing*, vol. 207, pp. 726–734, 2016.
4. Y. He, Y. Chen, Y. Hu, and B. Zeng, "WiFi vision: Sensing, recognition, and detection with commodity MIMO-OFDM WiFi," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8296–8317, Sep. 2020.
5. S. Seneviratne *et al.*, "A survey of wearable devices and challenges," *IEEE Commun. Surv. Tut.*, vol. 19, no. 4, pp. 2573–2620, Oct.–Dec. 2017.
6. F. Shu and J. Shu, "An eight-camera fall detection system using human fall pattern recognition via machine learning by a low-cost android box," *Sci. Rep.*, vol. 11, no. 1, pp. 1–17, 2021.
7. Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
8. B. Fu, S. Fu, L. Wang, Y. Dong, and Y. Ren, "Deep residual split directed graph convolutional neural networks for action recognition," *IEEE MultiMedia*, vol. 27, no. 4, pp. 9–17, Oct.–Dec. 2020.
9. Y. Kong, J. Huang, S. Huang, Z. Wei, and S. Wang, "Learning spatiotemporal representations for human fall detection in surveillance video," *J. Vis. Commun. Image Representation*, vol. 59, pp. 215–230, 2019.
10. K. Ozcan, S. Velipasalar, and P. K. Varshney, "Autonomous fall detection with wearable cameras by using relative entropy distance measure," *IEEE Trans. Hum.-Mach. Syst.*, vol. 47, no. 1, pp. 31–39, Feb. 2017.
11. G. V. Leite, G. P. da Silva, and H. Pedrini, "Three-stream convolutional neural network for human fall detection," in *Deep Learning Applications*, vol. 2. Singapore: Springer, 2021, pp. 49–80.
12. M. Dusmanu *et al.*, "D2-Net: A trainable CNN for joint description and detection of local features," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8084–8093.
13. C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Robust video surveillance for fall detection based on human shape deformation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 5, pp. 611–622, May 2011.
14. S. Wang, L. Chen, Z. Zhou, X. Sun, and J. Dong, "Human fall detection in surveillance video based on PCANet," *Multimedia Tools Appl.*, vol. 75, no. 19, pp. 11603–11613, 2016.
15. S. Boll, J. Meyer, and N. E. O'Connor, "Health media: From multimedia signals to personal health insights," *IEEE MultiMedia*, vol. 25, no. 1, pp. 51–60, Jan.–Mar. 2018.
16. F. Harrou, N. Zerrouki, Y. Sun, and A. Houacine, "Vision-based fall detection system for improving safety of elderly people," *IEEE Instrum. Meas. Mag.*, vol. 20, no. 6, pp. 49–55, Dec. 2017.
17. J. Gutiérrez, V. Rodríguez, and S. Martin, "Comprehensive review of vision-based fall detection systems," *Sensors*, vol. 21, no. 3, pp. 947–996, 2021.
18. N. Lu, Y. Wu, L. Feng, and J. Song, "Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 1, pp. 314–323, Jan. 2019.
19. A. Shahzad and K. Kim, "FallDroid: An automated smart-phone-based fall detection system using multiple kernel learning," *IEEE Trans. Ind. Inform.*, vol. 15, no. 1, pp. 35–44, Jan. 2019.
20. A. Abobakr, M. Hossny, and S. Nahavandi, "A skeleton-free fall detection system from depth images using random decision forest," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2994–3005, Sep. 2018.

YINLONG ZHANG is an associate professor with the State Key Laboratory of Robotics, Shenyang Institute of Automation, the Key laboratory of Networked Control Systems, and the Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China. His research interests include multisource data fusion, visual surveillance and pervasive computing, machine learning, and computer vision. Zhang received his Ph.D. degree in control theory and control engineering from the University of Chinese Academy of Sciences. He is a member of IEEE. Contact him at zhangyinlong@sia.cn.

XIAOYAN ZHENG is currently working toward the M.Eng. degree in control theory and control engineering with Shenyang Jianzhu University, Shenyang, 110168, China. His research interests include computer vision, visual surveillance, and Internet of Things. Zheng received the B.Eng. degree in electrical and electronic engineering from Shenyang Jianzhu University. Contact him at zhengxy3636@163.com.

WEI LIANG is currently a professor at the State Key Laboratory of Robotics, Shenyang Institute of Automation, the Key laboratory of Networked Control Systems, and the Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China. Her research interests include industrial wireless sensor networks and image processing. Liang received her Ph.D. degree in mechatronic engineering from the Shenyang Institute of Automation, Chinese Academy of Sciences. She is the senior member of IEEE. She is the corresponding author of this article. Contact her at weiliang@sia.cn.

SICHAO ZHANG is currently an associate professor with the State Key Laboratory of Robotics, Shenyang Institute of Automation, the Key laboratory of Networked Control Systems, and the Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China. His research interests include industrial control network security and wireless network security. Zhang received his M.Eng. degree in instrument and meter engineering from Chongqing University, Chongqing, China. Contact him at zhangsichao@sia.cn.

XUDONG YUAN is an associate professor with the State Key Laboratory of Robotics, Shenyang Institute of Automation, the Key laboratory of Networked Control Systems, and the Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang, 110169, China. His research interests include time sensitive network, information security of ICA, and industrial network. Yuan received his Ph.D. degree in control theory and control engineering from Northeast University. Contact him at yuanxudong@sia.cn.