Multimed Tools Appl (2016) 75:8469–8492
DOI 10.1007/s11042-015-2766-3

Multi-view fall detection based on spatio-temporal interest points

Songzhi Su[1,2] · **Sin-Sian Wu**[3] · **Shu-Yuan Chen**[3] · **Der-Jyh Duh**[4] · **Shaozi Li**[1,2]

Received: 22 September 2014 / Revised: 6 April 2015 / Accepted: 22 June 2015 /
Published online: 11 July 2015

Abstract Many countries are experiencing a rapid increase in their elderly populations, increasing the demand for appropriate healthcare systems including fall-detection systems. In recent years, many fall-detection systems have been developed, although most require the use of wearable devices. Such systems function only when the subject is wearing the device. A vision-based system presents a more convenient option. However, visual features typically depend on camera view; a single, fixed camera may not properly identify falls occurring in various directions. Thus, this study presents a solution that involves using multiple cameras. The study offers two main contributions. First, in contrast to most vision-based systems that analyze silhouettes to detect falls, the present system proposes a novel feature for measuring the degree of impact shock that is easily detectable with a wearable device but more difficult with a computer vision system. In addition, the degree of impact shock is less sensitive to camera views and can be extracted more robustly than a silhouette. Second, the proposed method uses a majority-voting strategy based on multiple views to avoid performing the tedious camera calibration required by most multiple-camera approaches. Specifically, the proposed method is based on spatio-temporal interest points (STIPs). The number of local STIP clusters is designed to indicate the degree of impact shock and body vibration. Sequences of these features are concatenated into feature vectors that are then fed into a support vector

✉ Shu-Yuan Chen
cschen@saturn.yzu.edu.tw

✉ Shaozi Li
szlig@xmu.edu.cn

[1] School of Information Science and Technology, Xiamen University, Xiamen, Fujian, China

[2] Fujian Key Laboratory of the Brain-like Intelligent Systems, Xiamen University, Xiamen, Fujian, China

[3] Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, Taiwan

[4] Department of Computer Science and Information Engineering, Chien Hsin University of Science and Technology, Taoyuan, Taiwan

machine to classify the fall event. A majority-voting strategy based on multiple views is then used for the final determination. The proposed method has been applied to a publicly available dataset to offer evidence that the proposed method outperforms existing methods based on the same data input.

Keywords Fall detection · Impact shock · Spatio-temporal interest points · Human silhouette · Multiple-view · Foreground segmentation · Camera calibration

1 Introduction

Automatic fall-detection systems can help elderly people live independently while providing continuous and automated access to required assistance [19]. Elderly people who experience a fall event and remain on the ground for an hour or more may suffer from many medical complications, such as dehydration, internal bleeding, or hyperthermia, potentially resulting in serious injury or death. Thus, it is urgent to develop an automatic fall-detection system to help elderly people live independently while ensuring their prompt access to required assistance.

According to Noury et al. [30], automatic methods for fall detection are mostly based on the following five characteristics: lack of significant movement, an impact shock, a lying position, a person lying on the ground, and vertical speed. Rougier et al. [34] added another characteristic, body-shape change, which can be quantified using a camera. In the past two decades, many fall-detection systems have been developed based on different characteristics. Such systems can be divided into four categories based on how the characteristics are measured [28, 41]: wearable [8, 9, 21, 24, 31, 43], ambient [2, 33, 44], camera-based (or vision-based) [3–7, 20, 23, 26, 27, 29, 32, 34, 36, 38, 39, 42], and multimodal [14, 16] devices.

Systems with wearable devices function only when the subject is wearing the device. Monitoring an environment with ambient devices is confined to the installation area. The shortcomings of the wearable and ambient devices could potentially be addressed using vision-based approaches [3–7, 14, 16, 20, 23, 26, 27, 29, 32, 34, 36, 38, 39, 42], in particular through emerging video surveillance systems [15, 25]. Although many vision-based fall-detection systems have been proposed [3–7, 14, 16, 20, 23, 26, 27, 29, 32, 34, 36, 38, 39, 42], most existing methods detect falls by analyzing silhouettes [3, 6, 23, 26, 27, 29, 32, 34, 38, 39, 42], which are easily affected by segmentation noise, and few are specifically designed according to the characteristics of impact shock [27]. This study proposes a novel feature to measure the degree of impact shock. Although the proposed feature is less sensitive to camera views and can be extracted more robustly compared to silhouette, it typically depends on camera angle. Thus, solutions employing multiple cameras were used in this study. However, most multiple-camera systems require camera calibration to compute reliable 3D information [3–5, 38, 42]. To avoid tedious calibration, this study uses majority voting [34].

The remainder of this paper is organized as follows: Section 2 introduces related work. Section 3 presents the dataset for fall detection. In Section 4, the proposed features for fall detection are described. Section 5 describes the SVM-based classifier for fall detection and the proposed majority-voting mechanism based on multiple views. Section 6 provides the experimental results, and Section 7 concludes the study and proposes future work.

2 Related work

Wearable device approaches [8, 9, 21, 24] rely on garments with embedded sensors, such as microwatches, accelerometers, gyroscopes, and spirit levels, to detect the motion and location of a subject's body. Although such devices are inexpensive to produce, they are relatively inconvenient to use. Because of the widespread adoption of mobile phones, researchers have attempted to embed a triaxial accelerometer in a cellphone [43] to reduce the inconvenience of wearable garments. Some studies have directly mounted a wearable camera on the subject's waist for determining fall events from rapid changes in the images of the surrounding environment [31].

Ambient devices [2, 33, 44] install vibration or pressure sensors in the floor or under the bed to overcome the inconvenience of wearable devices. Although these devices are unobtrusive to the user, the monitoring environment is confined to the installation area. Alwan et al. [2] designed a floor-vibration-based fall detector. The system proposed by Zigel et al. [44] was also based on floor vibrations but included sound sensing to reduce the incidence of false alarms. Rimminen et al. [33] used near-field imaging to detect the locations and motion patterns of people for fall detection.

A vision-based system presents a more convenient option. These approaches can be further divided into single-view [6, 23, 26, 27, 29, 32, 36, 39] and multiple-view [3–5, 7, 14, 16, 20, 34, 38, 42], according to the number of cameras deployed. Most single-view approaches perform fall analysis based on a human silhouette [6, 23, 27, 29] or the bounding box enclosing a human silhouette [26, 32, 39]. The 2D-image velocity of people has also been used to detect falls [23, 36]. In addition to visual images, some fall-detection systems use various sensing technologies, such as omnidirectional imaging [26, 29, 39] for wide-range views, infrared imaging [6, 36] for nighttime fall detection, and an optical-stereo signal or depth imaging [6, 7, 20], to prevent perspective problems.

Nait-Charif and Mckenna [29] used wide-angle cameras to track the subject in an ellipse. The resulting trajectory was analyzed to detect unusual inactivity. Miaou et al. [26] used the video from an omnidirectional camera mounted on the ceiling. The aspect ratios of the bounding boxes of people were used to detect fall events. Toreyin et al. [39] used wavelet transform to represent the periodic characteristic of the aspect ratio of the bounding box for fall detection. Sixsmith et al. [36] used infrared integrated systems to sense and track moving heat sources, with the size, location, and speed information of the hot "blob" being used to detect fall events. Banerjee et al. [6] collected data from various sensors, standard web cameras under normal illumination, web cameras with infrared lighting, and the inexpensive Microsoft Kinect camera to detect falls both day and night.

A single, fixed camera does not properly handle falls that occur in various directions. In general, there are three ways to resolve a perspective problem. First, some researchers found invariant features [27]; second, multiple cameras were set up, each for a different perspective [3–5, 14, 16, 34, 38, 42]; lastly, 3D depth data was directly used to prevent perspective problems [6, 7, 20]. Mirmahboub et al. [27] proposed a view-invariant feature to solve this problem by using a simple background separation method to find the silhouette. Variations in the silhouette area, obtained from a single camera, were used to detect fall events.

Most multiple-camera systems require camera calibration to compute reliable 3D information [3–5, 14, 16, 38, 42], and some of them use majority voting [34] to avoid tedious calibration. Thorme et al. [38] performed motion analysis on the basis of posture features from two camera views through metric image rectification. Auvinet et al. [4, 5] used eight-camera calibration

information to reconstruct the 3D shape of a human body. Fall events were then detected by analyzing the volume distribution along the vertical axis. Anderson et al. [3] analyzed the states of a voxel person obtained from two cameras by using a two-level fuzzy hierarchy. Yu et al. [42] also extracted video features from a voxel person with a single-class minimax probability machine to provide fall detection results. Rougier et al. [34] combined two characteristics, namely human shape deformation during a fall, followed by a lack of significant movement to classify the fall, and normal activities. The final decision was determined according to voting results from the four camera views.

Multimodal fall-detection systems [14, 16] focus on system integration to build a hybrid healthcare system. Fleury et al. [16] built a Grenoble Health Smart Home, which is a residential flat installed with infrared presence sensors, door contacts, temperature and moisture sensors in the bathroom, and microphones. A wearable kinematic sensor also provides information on postural transitions and walking periods. The data collected from the various sensors are then used to detect a loss of autonomy. Doukas et al. [14] also built a system to capture video, audio, and motion data from the patient's body sensors and the surrounding environment with overhead cameras and microphone arrays to detect emergency situations.

3 Database for fall detection

The proposed method is designed based on the multiple-camera fall dataset made publicly available by Auvinet et al. [5]. The dataset contains 24 scenarios, from which 22 were selected for experiments by Mirmahboub et al. [27]. In each scenario, an actor engages in several activities such as falling, sitting on a sofa, walking, and pushing objects. The 22 scenarios include 23 fall activities and 91 non-fall activities, each of which was shot simultaneously by using eight cameras, resulting in 912 video clips. Figure 1 shows the configuration of the cameras. The dataset is divided into training (comprising nine scenarios with 272 video clips) and test (comprising 13 scenarios with 640 video clips) sets, as in Mirmahboub et al. [27]. Table 1 shows a summary of the dataset.

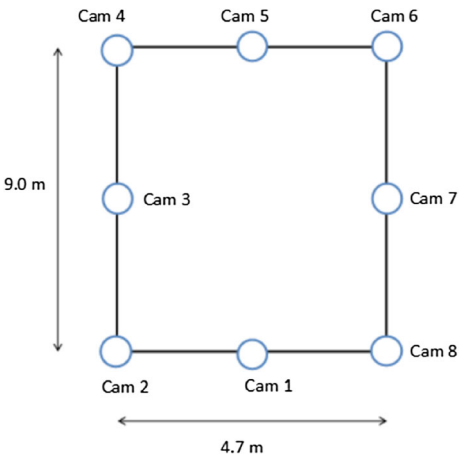


Fig. 1 Configuration of eight cameras [5]

Table 1 Summarization of the dataset

Subset	Scenarios	Video clips	Description	Total clips
Training	9	10×8=80 24×8=192	Fall activities Non-fall activities	272
Test	13	13×8=104* 67×8=536	Fall activities Non-fall activities	640

*Only 6×8=48 fall video clips were tested by Mirmahboub et al. [27]

As mentioned by the dataset creators [4, 5, 34], capturing real-life situations where people fall is impossible; consequently, they have designed scenarios of falls were performed by an actor in their laboratory with appropriate protection. However, each scenario in the dataset [5] was performed a single time by the actor and approved by the local institutional review board authority. In addition, the actor in the videos was one of the dataset creators, a clinician who was well aware of the different features of the real falls of elderly people and performed the simulated falls with appropriate precautions. Noury et al. [30] made the same claims that although the goal of a fall sensor is to detect the fall of elderly people, it is actually impracticable to test the fall situations with them. Thus the fall situations may be simulated by younger persons, or even athletes, and the normal activities may be tested on elderly in the risk age group for falls.

4 Feature extraction

4.1 STIP detection

Activity recognition methods based on spatio-temporal interest points (STIPs) have achieved considerable success in recent years. In this method, a video of an activity is represented by a set of STIPs that indicate rapid changes in the space-time domain. Local STIPs are commonly detected by the 3D Harris detector proposed by Laptev [22], Cuboid detector proposed by Dollar et al. [13], Hessian-matrix-based feature detector proposed by Willems et al. [40], and 3D SIFT detector proposed by Scovanner et al. [35]. The 3D Harris detector, extended from Harris corner detection [18] to 3D space by Laptev [22], with the STIP tool developed by Laptev [22] was adopted for this study. Figure 2 shows the STIP detection results with different threshold values. The default threshold value of the STIP tool was 1.00E-9 [22]. However, the results with a threshold value of 1.00E-7 had significant STIPs while less noise. Thus, the threshold value was set empirically to 1.00E-7 in this study.

4.2 Feature vector of shock

According to Noury et al. [30], automatic methods for fall detection are mostly based on the following characteristics: lack of significant movement, an impact shock, a lying position, a person lying on the ground, and vertical speed. Impact shock is easily detectable with an accelerometer or vibration detector but more difficult with a computer vision system [30, 34]. To the best of our knowledge, few vision-based methods [27] use the characteristics of impact shock for fall detection. This study proposed a novel feature to measure the degree of impact shock and lack of significant movement for fall detection.

(a)

(b)

(c)

(d)

(e)

◀ **Fig. 2** Results of STIP detection with different threshold values. (**a**),(**b**),(**c**),(**d**), and (**e**) are the results with threshold values of 0, 1.00E-9, 1.00E-7, 1.00E-6, and 1.00E-5, respectively. A green circle indicates a STIP

People generally do not engage in violent movements during their daily activities; thus, STIPs are typically concentrated, though minimally dense, in particular limbs of the body. However, when people suddenly engage in strenuous activity, particularly when falling, the shock and vibration may cause a sudden increase in the number of STIPs in the limbs. The stronger the impact, the denser the local STIP cluster. The local clustering of STIPs, represented by the number of STIPs in a specific proximity, was derived to be a new feature for indicating the degree of impact shock and is described in the remainder of this section. Notably the feature is designed to measure the strong motion with acceleration caused by the impact shock rather than the impact itself which yields to the new STIPs.

First, a spatial proximity graph G for a frame F is constructed with each STIP as a node and each pair of STIPs with a distance of less than a threshold value, T_d, being connected by an edge. Figure 3a and b shows an example of a frame with 17 STIPs and the corresponding spatial proximity graph, respectively. A connected component of the graph G is then defined as a maximal connected subgraph in which any pair of nodes can be connected by a path (i.e., an alternating sequence of nodes and edges between the pair of nodes). Notably, a graph G that is not connected has two or more connected components. In addition, each connected component corresponds to a cluster of STIPs. For example, there are five connected components, $H_1,\cdots,H_5$, in Fig. 3b, and five corresponding clusters, $C_1,\cdots,C_5$, in Fig. 3a.

Hence, the number of nodes in each connected component indicates cluster size. The maximum number of nodes is then designated as the degree of shock for the frame. For example, the number of nodes for $H_1,\cdots,H_5$ is 1, 2, 3, 4, and 7, respectively; thus, the degree of shock for the frame is 7. Figure 4 shows the degree of shock from different views for falls and normal walking activity. The radius of the white circle indicates the degree of density of the STIP clusters and the degree of shock feature. The longer the radius is, the denser the local STIP cluster and the higher the degree of shock. Clearly, the degree of shock for a fall event is higher than that for a normal walking activity. In this study, T_d was set at 5 empirically because the size of a human in a frame with a resolution of 320×240 is approximately 10×100. The threshold value T_d is insensitive to camera views because the fixed value of 5 is feasible for different views, as shown in Fig. 4. However, the value T_d should be adapted to the video

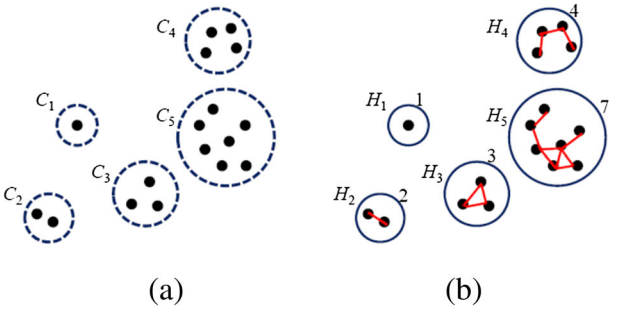


Fig. 3 An example of derivation of degree of shock. **a** A frame with 17 STIPs and 5 clusters; **b** spatial proximity graph G with 5 connected components

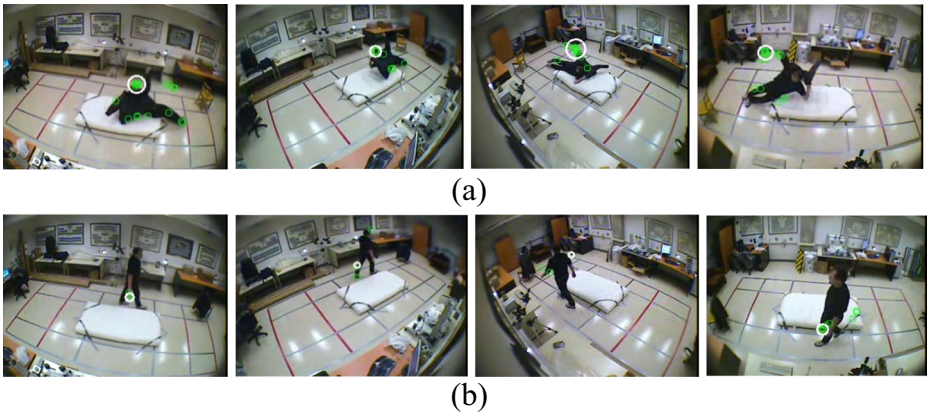


Fig. 4 Degree of shock under different views. **a** Fall; **b** normal activity of walking

resolution. The problem can be solved through proper initial setup settings because the environment of the fall-detection system is controllable.

We let the degree of shock for the frame at time t be h_t. For a continuous 200 frames, the corresponding degrees of shock are then concatenated as feature vector of shock, $\mathbf{h}=(h_t)_{t=1,\ldots,200}$, as in Mirmahboub et al. [27]. Although the number of frames is set to 200 in this study, the proposed method functions satisfactorily with 150 to 250 frames. Notably, 1 to 2 s should be enough to correctly detect fall event since the frame rate is 120 frames per second in the dataset used in this study. However, if the frame rate of the video changes drastically, the number of frames can be adapted to the frame rate and set initially during setup. Figure 5 shows the feature vector of shock under different views for falls and normal walking activity. Clearly, the feature vectors of shock for the fall event have peaks and then rapidly decline to zero. The zero tail indicates the fall is characterized by a lack of significant movement, as in Mirmahboub et al. [27] and Rougier et al. [34].

5 Fall detection

The SVM [1, 11, 12] is a supervised learning method frequently used in classification and regression analyses and is used in the proposed method for fall detection. Sample videos labeled either positive (fall) and negative (non-fall) are fed into the SVM for learning and training. The fall feature vectors are extracted from all of the training video samples including the positive fall videos and negative non-fall videos. These vectors are then fed into the SVM to train SVM-based classifiers. The output of each trained SVM is a model file composed of support vectors. Subsequently, the model is used to evaluate the score determining whether a video portrays a fall event. A high score indicates a high confidence of a fall event having occurred.

5.1 SVM model training

The quality of the SVM training module is crucial for classification accuracy; thus, selecting discriminative training data is critical to producing a satisfactory model. In this study, bootstrap training and elitist selection are used to achieve this goal.

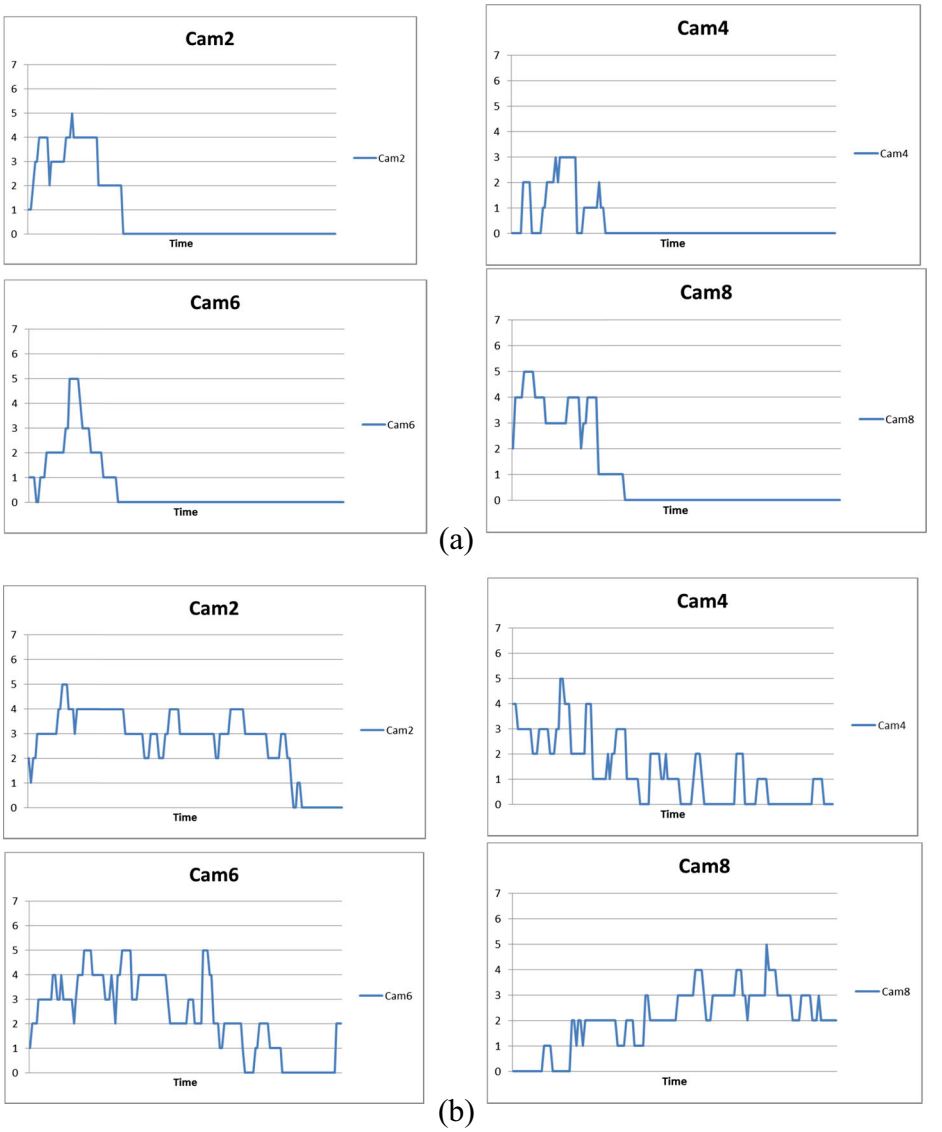


Fig. 5 Feature vectors of shock under different views. **a** Fall; **b** normal activity of walking

5.1.1 Bootstrap training

In this study, the bootstrap approach [37] is adopted to train the SVM. First, we train a naïve SVM by using the initial training set and then run the naïve SVM on all of the videos in the initial training set. Next, we collect all of the videos that are incorrectly classified by the naïve SVM according to two types of errors: falls incorrectly classified as normal activity or normal activity incorrectly classified as falls. Finally, we remove all of the misclassified videos from the initial training set to obtain a final training set and train a new SVM by using the final training set to obtain a retrained SVM.

5.1.2 Elitist selection

The concept of the proposed elitist selection is similar to that of random forests [10]. The training set is further partitioned into three parts, two for training and one for validation (i.e., a training subset and a validation subset). As mentioned in Section 3, the training set contained nine scenarios; thus, there were 84 possible scenario combinations. For each combination, one retrained SVM can be obtained using the bootstrap training procedure, as described in Section 5.1.1. All of the 84 retrained SVMs can be tested on the corresponding validation subset, resulting in 84 classification rates. A number of SVMs with the highest classification rates are selected as elitist SVMs, namely E. The elitist selection is performed for the proposed degree of shock feature, leading to E elitist SVMs; that is, $SVM_h^{(r)}$, with the corresponding rank $r=1,\cdots,E$. Notably, the classification rate is defined as the sensitivity value as described in Section 6.3. The classification rates of $SVM_h^{(r)}$ are assumed to be the corresponding rates of $R_h^{(r)}$. Clearly, the rates of $R_h^{(r)}$ are in descending order with the increasing rank r. For comparison, an elitist selection is also performed for the other two features, the degree of scatter and area, individually leading to E elitist SVMs for each feature; that is, $SVM_s^{(r)}$ and $SVM_a^{(r)}$, $r=1,\cdots,E$. Similarly, the corresponding classification rates of $SVM_s^{(r)}$ and $SVM_a^{(r)}$, are $R_s^{(r)}$ and $R_a^{(r)}$, respectively. Notably, only the first 10 SVMs are selected as elitists, i.e., $E=10$, because the classification rate has a break at Rank 10, as described in Section 6.3.

5.2 Majority-voting mechanism

A majority-voting mechanism based on multiple views was used for the final determination of fall events. The dataset used in this study was produced using eight cameras [5], thus the number of views, V, is set to eight in this study. Figure 6 shows the scheme of the proposed majority-voting mechanism. Two voting strategies were proposed: un-weighted and weighted majority voting. The difference between the two strategies is that un-weighted majority voting treats the elitist SVMs equally, whereas weighted majority voting trusts the SVMs to perform more accurately alone.

The formula of the un-weighted voting mechanism, D_F, is defined using Eq. (1)

$$D_F = \begin{cases} 1 & \text{if } \dfrac{1}{E \times V} \displaystyle\sum_{r=1}^{E} \sum_{i=1}^{V} \mathrm{SVM}_F^{(r)}(i) \geq T_F \\ 0 & \text{otherwise} \end{cases}, \quad F = h, s, a \tag{1}$$

Here, E and V denote the number of elitist SVMs and camera views, respectively, and set to 10 and 8 in this study. The output of 1 for $\mathrm{SVM}_F^{(r)}(i)$ indicates that the SVM $\mathrm{SVM}_F^{(r)}$ classifies the video of Camera i as a fall event, and the output of 1 for D_F indicates that the fall event eventually occurs if the pros over a threshold value T_F. The h, s, and a of symbol F denote, respectively, the feature types of shock, scatter, and area.

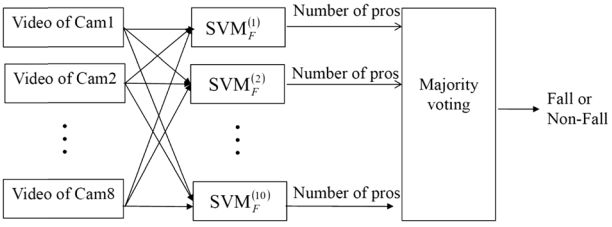


Fig. 6 Scheme of the proposed voting mechanism

The formula of the weighted voting mechanism, WD_F, is defined using Eq. (2).

$$WD_F = \begin{cases} 1 & \text{if } \frac{1}{E \times V} \sum_{r=1}^{E} \mathrm{W}_F^{(r)} \sum_{i=1}^{V} \mathrm{SVM}_F^{(r)}(i) \geq T_F \\ 0 & \text{otherwise} \end{cases}, F = h, s, a$$
$$\mathrm{W}_F^{(r)} = \frac{R_F^{(r)}}{\sum_{r=1}^{E} R_F^{(r)}} \tag{2}$$

with $R_F^{(r)}$ as specified in Section 5.1.2. However, experimental results show that the performances of un-weighted and weighted majority voting are almost even as described in Section 6.3. Thus, the simple version of un-weighted voting is adopted in this study.

In addition, a fusion of two features is also included in this study to compare detection accuracy. The formula of the unweighted voting mechanism, $D_{F_1+F_2}$, with the fusion feature is defined by Eq. (3)

$$D_{F_1+F_2} = \begin{cases} 1 & \text{if } \frac{1}{2 \times E \times V} \sum_{r=1}^{E} \sum_{i=1}^{V} \left[\mathrm{SVM}_{F_1}^{(r)}(i) + \mathrm{SVM}_{F_2}^{(r)}(i) \right] \geq T_F \\ 0 & \text{otherwise} \end{cases}, F_1, F_2 = h, s, a, F_1 \neq F_2 \tag{3}$$

Notably, the threshold value T_F in Eqs. (1) through (3) can be set automatically at $\frac{E \times V + 1}{2 \times E \times V}$ because if the number of votes is $2 \times E \times V$ then the pros should be majority (i.e., over $E \times V$) to decide that a fall event is occurring. In this study, E and V are set to 10 and 8, respectively, thus T_F can be derived as 0.5625.

6 Experimental results

6.1 Experiment setup

The proposed method was implemented on the Microsoft Windows 7 (professional version) operating system. The program was developed in the C++ language and compiled in Microsoft Visual Studio 2010 with an open source OpenCV library (Version 2.4.4) for handling video input and output and running the support vector machine LIBSVM with a linear kernel proposed by Chang and Lin [11]. The STIP tool was developed by Laptev [22].

Notably, the test sets used by Mirmahboub et al. [27] were different from ours, although the dataset was the same (the multiple-camera fall dataset made publically available by Auvinet et al. [4, 5, 34]). Mirmahboub et al. tested only 6×8=48 video clips of falls [27], whereas the present study tested 13×8=104 video clips of falls, thus producing different experimental results from those of Mirmahboub et al. [27].

6.2 Feature vectors of scatter and area for comparison

To offer evidence that the proposed degree of shock feature is discriminative and robust, this study also designs another feature of degree of scatter to represents silhouettes rather than the

Table 2 A Summary of three features, degree of shock, degree of scatter, and area

Features	Degree of shock	Degree of scatter	Area
Characteristics	Shock	Silhouette	Silhouette
Extraction	STIP-based	STIP-based	Segmentation

area, as in Mirmahboub et al. [27]. Similar to the proposed degree of shock feature, the degree of scatter is extracted based on the STIPs, whereas the area is extracted through segmentation. A summary of the differences between the three features is shown in Table 2. The experimental results in Section 6.3 offer evidence that the degree of shock is more discriminative than silhouette representation, whereas STIP-based extraction is more robust than segmentation extraction.

Similar to the degree of shock, when people suddenly engage in strenuous activity, particularly when a fall occurs, the number of STIPs suddenly increases throughout the whole body (Fig. 7). The faster the speed of falling down, the greater the STIPs spread. Thus, the distribution of STIPs (i.e., the degree of scatter) is measured according to the distance between all STIPs and the center of their distribution, to indicate the degree to which the STIPs have spread.

More specifically, the degree of scatter at time t, s_t, is calculated according to the distribution of STIPs in the frame at time t. We assume that there are n_t STIPs, $\mathbf{p}_t^i$, $i=1,\cdots,n_t$, each with coordinates (x_t^i, y_t^i). The degree of scatter, s_t, indicating the variance of the coordinates of the n_t STIPs, can then be derived by Eq. (4).

$$\begin{aligned} s_t &= \sum_{i=1}^{n_t} \left(\mathbf{p}_t^i - \boldsymbol{\mu}_t\right)^T \left(\mathbf{p}_t^i - \boldsymbol{\mu}_t\right) \\ \boldsymbol{\mu}_t &= \frac{\sum_{i=1}^{n_t} \mathbf{p}_t^i}{n_t} \end{aligned} \tag{4}$$

Figure 7 shows the degree of scatter from different views for falls and normal walking activity. The radius of the white circle indicates the degree of scatter, and the cross indicates the

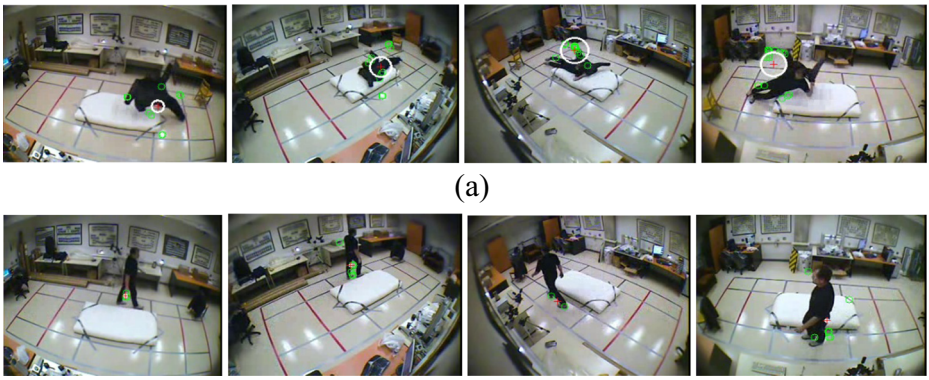


Fig. 7 Degree of scatter under different views. **a** Fall; **b** normal activity of walking

center of the STIPs (i.e., μ_t). Clearly, the scatter value for a fall event is higher than that for a normal walking activity.

Similarly, the feature vector of scatter, $\mathbf{s}=(s_t)_{t=1,\ldots,200}$, can be obtained. In addition, the feature vector of area, $\mathbf{a}=(a_t)_{t=1,\ldots,200}$, is also defined because the feature area was used in Mirmahboub et al. [27], to which the proposed method is compared in Section 6. Notably, the area is obtained by separating a stationary background from a moving foreground and is defined as the area of the foreground silhouette [27]. Figure 8 shows

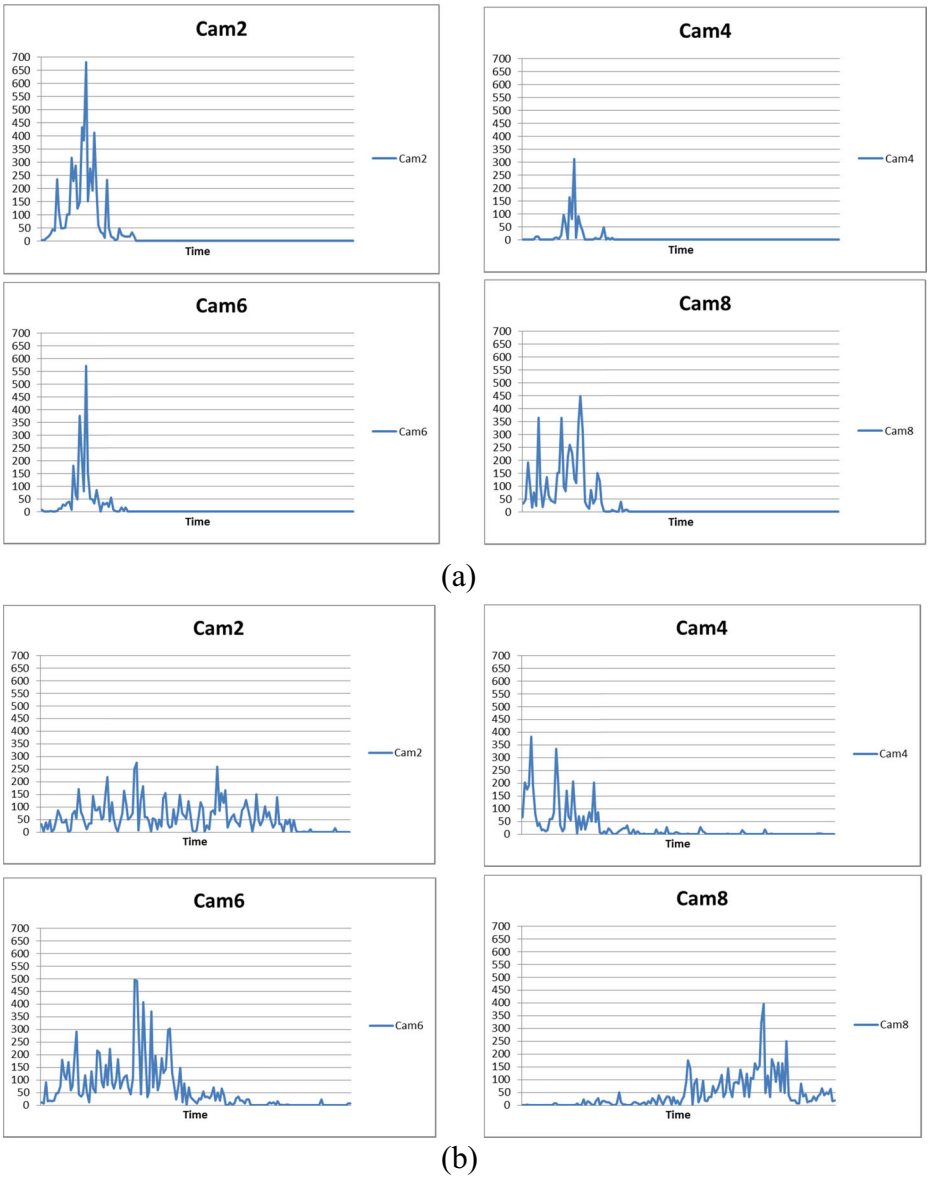


Fig. 8 Feature vectors of scatter under different views. **a** Fall; **b** normal activity of walking

Table 3 A summary of performance measures

Measures	Sensitivity	Selectivity	Specificity	Accuracy
Equations	$\frac{TP}{TP+FN}$	$\frac{TP}{TP+FP}$	$\frac{TN}{TN+FP}$	$\frac{TP+TN}{TP+FN+TN+FP}$
Remark	Recall	Precision		

the feature vectors of scatter from different views for falls and normal walking activity. Although the feature vectors of scatter for the fall event have peaks and then rapidly decline to zero, the characteristic is less significant than that of the feature vectors of shock from all of the views. For example, the feature vector of scatter for normal activity, as shown in Fig. 8b, with Camera 4 is relatively similar to that for the fall event for Camera 8, as shown in Fig. 8a. Thus, the degree of impact shock is less sensitive to camera views than the degree of shock, representing silhouette in this study. The experimental results in Section 6.3 offer more evidence that the degree of shock has the most discriminative power.

6.3 Performance evaluation

Performance is evaluated using the measures of sensitivity (Sen), specificity (Spe), and accuracy (Acc), as in Mirmahboub et al. [27], but with the addition of selectivity (Sel) in this study because Sel is also an informative measure for performance comparisons. The following parameters are used to define the performance measures:

(a) True positive (TP): the number of fall events detected correctly.
(b) True negative (TN): the number of normal activities detected correctly.
(c) False positive (FP): the number of normal activities detected as fall events.
(d) False negative (FF): the number of fall events detected as normal activities.

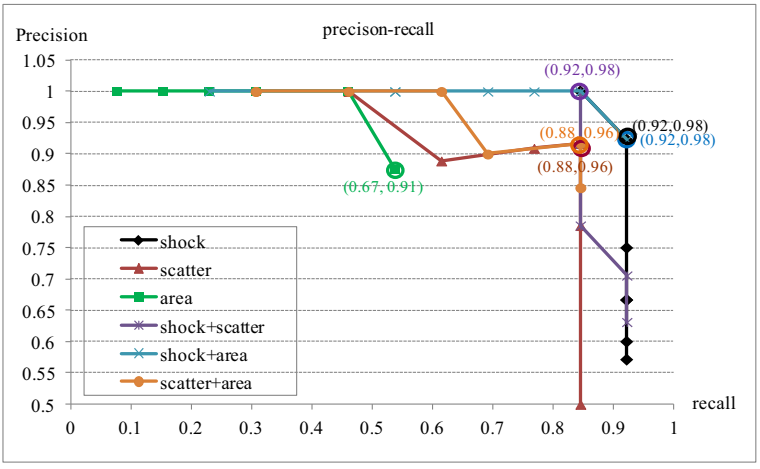
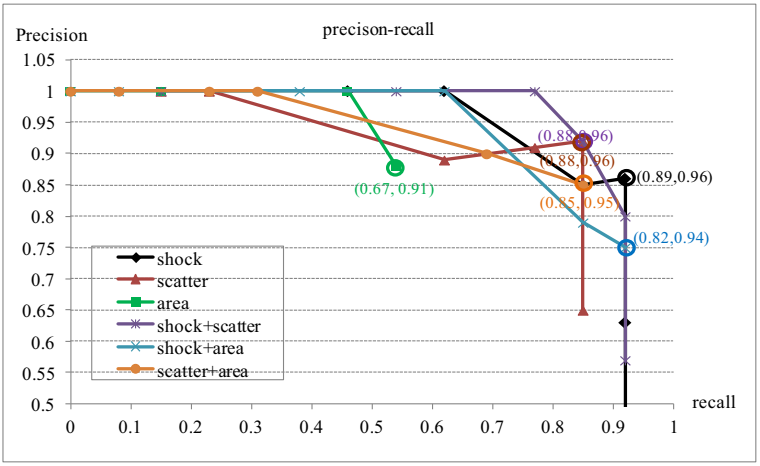


Fig. 9 Precision-recall curves from using various features in eight-view environment. The (x, y) is added beside each circle to denote corresponding maximum F-score (x) and accuracy values (y), respectively

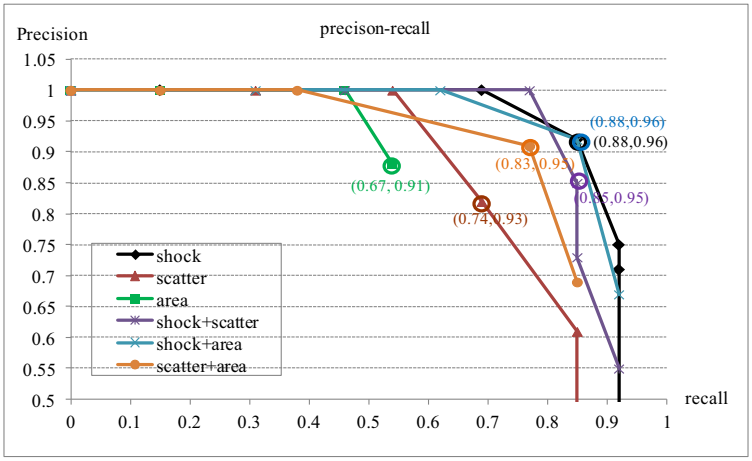
Table 4 Maximum F-score values of using different features with corresponding performance values

Feature	Shock	Scatter	Area	Shock+Scatter	Shock+Area	Scatter+Area
F-Score	**0.92**	0.88	0.67	**0.92**	**0.92**	0.88
Sen	**0.92**	0.85	0.54	0.85	**0.92**	0.85
Sel	0.92	0.92	0.88	**1.00**	0.92	0.92
Spe	0.99	0.99	0.99	**1.00**	0.99	0.99
Acc	**0.98**	0.96	0.91	**0.98**	**0.98**	0.96

The highest values for each measure are marked as bold



(a)



(b)

Fig. 10 Precision-recall curves from using various features in four-view environments. The (x, y) is added beside each circle to denote corresponding maximum F-score (x) and accuracy values (y), respectively. **a** Cameras 1, 3, 5, 7; **b** Cameras 2, 4, 6, 8

The Sen value is then defined as the ratio of the number of fall events detected correctly (TP) to the total number of fall events (TP+FN). The Sel value is defined as the ratio of the number of fall events detected correctly (TP) to the number of detected fall events (TP+FP). The Spe value is defined as the ratio of the number of normal activities detected correctly (TN) to the total number of normal activities (TN+FP). The Acc value is defined as the ratio of the number of fall events and normal activities detected correctly (TP+TN) to the total number of fall events and normal activities (TP+FN+TN+FP). The measures of Sen and Sel correspond to recall and precision, respectively [17]. Table 3 shows a summary of the performance measures. Moreover, the proposed method also uses the evaluation measure F-score [17], which is univariable combination of recall (R) and precision (P) and defined using Eq. (5).

$$F_b = \frac{\left(b^2+1\right) \times R \times P}{b^2 \times R + P} \tag{5}$$

with b=1.

Three experiments were conducted in this study. The first experiment was performed to show that the degree of shock had the most discriminative power and was even better than that of feature fusion. The second experiment was performed to show that the degree of shock had the most discriminative power, even in four-view and single-view environments. Finally, the third experiment was conducted to confirm that the proposed strategies were feasible including that bootstrap training was necessary, only the first 10 SVMs being selected as elitists was reasonable, and un-weighted majority voting was sufficient.

There were five conclusions were drawn in the first experiment as shown in Figure 9 and Table 4. Figure 9 shows the precision-recall (P-R) curves from using different features in an eight-view environment. In this figure, the maximum F-score values are also marked with a circle for respective features. In addition, (x, y) is added beside each circle to denote corresponding maximum F-score (x) and accuracy values (y), respectively. Table 4 shows the maximum F-score values from using different features with corresponding performance meaure values. First, regarding single feature, the highest accuracies—0.98, 0.96, and 0.91—were obtained for the

Table 5 Maximum F-score values of using different features with corresponding performance values in four-view environments

Feature	Shock	Scatter	Area	Shock+Scatter	Shock+Area	Scatter+Area
(a)						
F-Score	**0.89**	0.88	0.67	0.88	0.82	0.85
Sen	**0.92**	0.85	0.54	0.85	**0.92**	0.85
Sel	0.86	0.92	0.88	**0.92**	0.75	0.85
Spe	0.97	0.99	**0.99**	**0.99**	0.94	0.97
Acc	**0.96**	0.96	0.91	**0.96**	0.94	0.95
(b)						
F-Score	**0.88**	0.74	0.67	0.85	**0.88**	0.83
Sen	**0.85**	0.69	0.54	0.85	**0.85**	0.77
Sel	0.92	0.82	0.88	**0.85**	0.92	0.91
Spe	**0.99**	0.97	**0.99**	0.97	**0.99**	**0.99**
Acc	**0.96**	0.93	0.91	0.95	**0.96**	0.95

(a) Cameras 1, 3, 5, 7; (b) Cameras 2, 4, 6, 8. The highest values for each measure are marked as bold

Table 6 A summary of performance values from using different features in single-view environment

Feature	Shock	Scatter	Area	Shock+Scatter	Shock+Area	Scatter+Area
Sen	**0.77**	0.58	0.32	**0.83**	0.81	0.65
Sel	**0.56**	0.54	0.55	0.46	**0.50**	0.47
Spe	0.87	0.90	**0.90**	0.81	0.81	**0.83**
Acc	**0.86**	0.85	0.81	**0.81**	0.81	0.80

The highest values of each measure are marked as bold for the cases of singe feature and feature fusion, respectively

degree of shock (shock), degree of scatter (scatter), and area, respectively. Clearly, shock had the highest recognition accuracy of a singly employed feature. Second, regarding feature fusion, the highest accuracies—0.98, 0.98, and 0.96—were obtained for the fusion of shock and scatter (shock+scatter), shock and area (shock+area), and scatter and area (scatter+area), respectively. Although shock+scatter and shock+area had the same highest accuracies, the area under the *P-R* curve of shock+scatter was less than that of shock+area. Thus, shock+area had the highest performance for the case of feature fusion.

Third, the proposed feature of degree of shock was superior, even better to that of feature fusion because the effectiveness of shock and fusion of shock and area (shock+area) were approximately the same, whereas shock is a single feature and shock+area employs two features. The reasons are that shock is less sensitive to camera views rather than silhouette easily affected by the camera views in close-up, long range, frontal or lateral views. Thus, involving the noisy features (degree of scatter or area) does not boost the accuracy of using degree of shock alone. Fourth, the experimental results offer evidence that degree of shock was more discriminative than the silhouette either represented by degree of scatter or area, and STIP-based extraction (shock and scatter) was more robust than segmentation extraction (area). Thus, shock had the best performance, even comparable with the feature fusion, whereas area had the lowest performance. Fifth, the results showed that the proposed method performed with an accuracy of 0.98, higher than that of 0.95 of the newest method [27] or that of 0.96 of the most effective [34] of existing methods on the same publicly available dataset [5]. Notably, the 0.95 and 0.96 values were quoted from [27] and [34] and rounded off to two significant digits, respectively.

The results of second experiment were shown in Fig. 10 and Tables 5 and 6. The *P-R* curves from using different features in a four-view environment are shown in Fig. 10a (Cameras 1, 3, 5, and 7) and (b) (Cameras 2, 4, 6, and 8). Table 5(a) and (b) shows a summary of the maximum F-score values from using different features with corresponding performance

Table 7 A summary of performance values produced by naive or retrained SVMs using different features

Feature	Shock		Scatter		Area	
Type	Naïve	Retrained	Naïve	Retrained	Naïve	Retrained
Sen	0.77	**0.92**	0.77	**0.85**	**0.54**	**0.54**
Sel	**1.00**	0.92	0.91	**0.92**	**1.00**	0.88
Spe	**1.00**	0.99	**0.99**	**0.99**	**1.00**	0.99
Acc	0.96	**0.98**	0.95	**0.96**	**0.93**	0.91

The highest values of each measure are marked as bold for respective features

Table 8 Maximum F-score values of using different voting mechanisms with corresponding performance values for degree of shock

Mechanism	Un-weighted	Weighted
F-Score	**0.923**	0.917
Sen	**0.923**	0.846
Sel	0.923	**1.000**
Spe	**0.985**	0.955
Acc	**0.975**	**0.975**

The highest values for each measure are marked as bold

meaure values. Summaries of the performance values from using different features in a single-view environment are shown in Table 6. Conclusions similar to those drawn from Fig. 9 and Table 4 were drawn from Fig. 10, Tables 5 and 6. Clearly, the degree of shock provided the highest performance; even the fusion of shock and scatter or shock and area did not improve the performance of the degree of shock alone. In addition, the results offer evidence that the degree of shock was less sensitive to camera views rather than silhouette (degree of scatter or area) because the former had less variation in performance values between the environments of Cameras 1, 3, 5, and 7 and Camera 2, 4, 6, and 8 than that of the latter, even less than that of feature fusion. Finally, performance did not decay significantly when fall detection was tested in a four-view environment but decayed drastically in a single-view environment, thus indicating that multiple views can improve the accuracy of fall detection.

The third experiment was concluded with Tables 7 and 8 and Figs. 11 and 12. Table 7 shows a comparison of the performance of the proposed method with and without bootstrap training. The accuracies of the retrained SVMs were 0.98 and 0.96 for shock and scatter, respectively, which exceeded the accuracies of 0.96 and 0.95 achieved using the respective naïve SVMs. However, regarding area, a conclusion may not be drawn because the effectiveness of the feature area with and without bootstrap was approximately the same. As shown in Fig. 11, the classification rates for the degree of shock, degree of scatter, and area have breaks at Rank 10, respectively. The *P-R* curves from using different voting mechanisms for the feature of the degree of shock are shown in Fig. 12. Table 8 shows a summary of the maximum F-score values from using different voting mechanisms with corresponding performance values for the degree of shock. The results showed that the performance of weighted voting

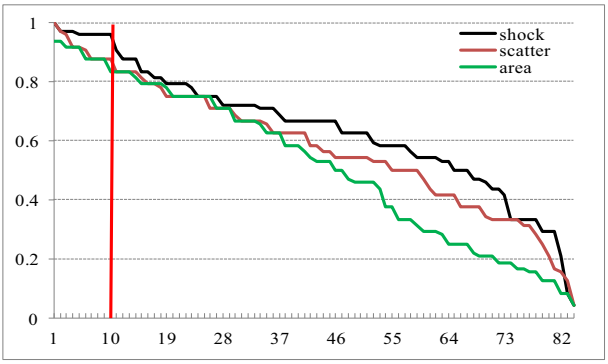


Fig. 11 Classification rates achieved by SVMs with different ranks for respective features

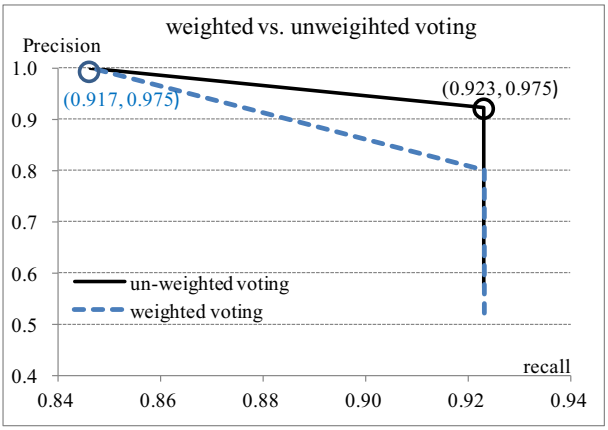


Fig. 12 Precision-recall curves from using different voting mechanisms for degree of shock. The (x, y) is added beside each *circle* to denote corresponding maximum F-score (x) and accuracy values (y), respectively

was no better than and even less than that of un-weighted a little. Thus, the simple version of un-weighted voting is adopted in this study.

Finally, Fig. 13 shows examples of video clips misclassified by the proposed method. As shown in Fig. 13a, the activity in the clip classified as a false negative is a fall, but one starting from a sitting position; thus, the impact shock was too weak to be detected. As shown in Fig. 13b and c, the activities in the clips classified as false positives are moving down and lying on a sofa, which were defined by the dataset creators as confounding events with characteristics similar to falls [4, 5, 34].

7 Conclusions and future work

The proposed method with eight views has an accuracy of 0.98, higher than the results obtained using the newest (0.95) and most effective (0.96) of existing methods. The proposed degree of shock feature based on the STIPs is discriminative and robust, and the proposed majority-voting mechanism based on multiple views facilitates fall detection from various views. However, the proposed method has some limitations; for example, our method does not work in various light conditions, especially those of extremely low light and extremely bright

Fig. 13 Examples of misclassified video clips. (**a**) False negative; (**b**) and (**c**) false positive

light. Thus, employing the multimodal approach for building a hybrid fall-detection system is an effective solution. Future studies can focus on:

(a) Extending the proposed method to resolve occlusion problems that are a continuing challenge.
(b) Extending the proposed method to detect fall events in 3D video sequences.
(c) Building a hybrid fall-detection system for day and night by combing the proposed method using visual and infrared imaging devices together with wearable devices

Acknowledgments The authors would like to thank E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau for providing "Multiple cameras fall dataset," and anonymous reviewers for the valuable and insightful comments on the earlier version of this manuscript. This work was supported by the National Science Council of Taiwan, Republic of China (NSC-103-2221-E-155-033-), the Nature Science Foundation of China (No. 61202143), and the Natural Science Foundation of Fujian Province (No. 2013J05100).

References

1. Alpaydin E (2010) Introduction to machine learning. MIT Press
2. Alwan M, Rajendran PJ, Kell S, Mack D, Daldl S, Wolfe M, Felder R (2006) A smart and passive floor-vibration based fall detector for elderly. In: Proc. IEEE Int. Conf. Information and Communication Technologies, pp. 1003–1007
3. Anderson D, Luke RH, Keller JM, Skubic M, Rantz M, Aud M (2009) Linguistic summarization of video for fall detection using voxel person and fuzzy logic. Comput Vis Image Underst 113(1):80–89
4. Auvinet E, Multon F, St-Arnaud A, Rousseau J, Meunier J (2011) Fall detection with multiple cameras: an occlusion-resistant method based on 3-D Silhouette vertical distribution. IEEE Trans Inf Technol Biomed 15(2):290–300
5. Auvinet E, Rougier C, Meunier J, St-Arnaud A, Rousseau J (2010) Multiple cameras fall dataset. DIRO - Universit'e de Montr'eal, Tech. Rep. 1350
6. Banerjee T, Keller JM, Skubic M, Stone E (2014) Day or night activity recognition from video using fuzzy clustering techniques. IEEE Trans Fuzzy Syst 22(3):483–493
7. Belbachir AN, Litzenberger M, Schram S, Hofstätter M, Bauer D, Schön P, Humenberger M, Sulzbachner C, Lunden T, Merne M (2012) CARE: a dynamic stereo vision sensor system for fall detection. In: Proc. IEEE Int. Symp. Circuits and Systems, pp. 731–734
8. Bianchi F, Redmond S, Narayanan M, Cerutti S, Lovell N (2010) Barometric pressure and triaxial accelerometry-based falls event detection. IEEE Trans Neural Syst Rehabil Eng 18(6):619–627
9. Boissy P, Choquette S, Hamel M, Noury N (2007) User-based motion sensing and fuzzy logic for automated fall detection in older adults. Telemed e-Health 13(6):683–693
10. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
11. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm
12. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
13. Dollar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: Proc. Int. Conf. Visual Surveillance and Performance Evaluation of Tracking and Surveillance
14. Doukas CN, Maglogiannis I (2011) Emergency fall incidents detection in assisted living environments utilizing motion, sound, and visual perceptual components. IEEE Trans Inf Technol Biomed 15(2):277–289
15. Fleck S, Strasser W (2008) Smart camera based monitoring system and its application to assisted living. Proc IEEE 96(10):1698–1714
16. Fleury A, Vacher M, Noury N (2010) SVM-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results. IEEE Trans Inf Technol Biomed 14(2):274–283

17. Goutte C, Gaussier E (2005) A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Proc. Europe Conf. Information Retrieval Research, pp. 345–359
18. Harris C, Stephens M (1988) A combined corner and edge detector. In: Proc. Alvey Vision Conference
19. Hawley-Hague H, Boulton E, Hall A, Rfeiffer K, Todd C (2014) Older adults' perceptions of technologies aimed at falls prevention, detection or monitoring: a systematic review. Int J Medical Inform 83(6):416–426
20. Huang S-H, Pan Y-C (2013) Learning-based fall detection using RGB-D cameras. In: Proc. IEEE Int. Conf. Machine Vision Applications, pp. 439–442
21. Karantonis DM, Narayanan MR, Mathie M, Lovell NH, Celler BG (2006) Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. IEEE Trans Inf Technol Biomed 10(1):156–167
22. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2–3):107–123
23. Lee T, Mihailidis A (2005) An intelligent emergency response system: preliminary development and testing of automated fall detection. J Telemed Telecare 11(4):194–198
24. Lin C-S, Hsu HC, Lay Y-L, Chiu C-C, Chao C-S (2007) Wearable device for real-time monitoring of human falls. Measurement 40(9–10):831–840
25. Liu H, Chen S, Kubota N (2013) Intelligent video systems and analytics: a survey. IEEE Trans Ind Inf 9(3): 1222–1233
26. Miaou SG, Sung PH, Huang CY (2006) A customized human fall detection system using omni-camera images and personal information. In: Proc. Int. Conf. Distributed Diagnosis Home Healthcare
27. Mirmahboub B, Samavi S, Karimi N, Shirani S (2013) Automatic monocular system for human fall detection based on variations in silhouette area. IEEE Trans Biomed Eng 60(2):427–436
28. Mubashir M, Shaon L, Seed L (2013) A survey on fall detection: principles and approaches. Neurocomputing 100:144–152
29. Nait-Charif H, McKenna SJ (2004) Activity summarisation and fall detection in a supportive home environment. In: Proc. Int. Conf. Pattern Recognition, pp. 323–326
30. Noury N, Fleury A, Rumeau P, Bourke AK, Laighin GO, Rialle V, Lundy JE (2007) Fall detection – principles and methods. In: Proc. Int. Conf. Engineering in Medicine & Biology Society, pp. 1663–1666
31. Ozcan K, Mahabalagiri AK, Casares M, Velipasalar S (2013) Automatic fall detection and activity classification by a wearable embedded smart camera. IEEE J Emerging Sel Top Circuits Syst 3(2): 125–136
32. Qian H, Mao Y, Xiang W, Wang Z (2008) Home environment fall detection system based on a cascaded multi-SVM classifier. In: Proc. Int. Conf. Control, Automation, and Robot Vision
33. Rimminen H, Lindstrom J, Linnavuo M, Sepponen R (2010) Detection of falls among the elderly by a floor sensor using the electric near field. IEEE Trans Inf Technol Biomed 14(6):1475–1476
34. Rougier C, Meunier J, St-Arnaud A, Rousseau J (2011) Robust video surveillance for fall detection based on human shape deformation. IEEE Trans Circuits Syst Video Technol 21(5):611–622
35. Scovanner P, Ali S, Shah M (2007) A 3-dimensional sift descriptor and its application to action recognition. In: Proc. Int. Conf. Multimedia
36. Sixsmth A, Johnson N (2004) A smart sensor to detect falls of the elderly. IEEE Trans Pervasive Comput 3(2):42–47
37. Sung KK, Poggio T (1998) Example-based learning for view-based human face detection. IEEE Trans Pattern Anal Mach Intell 20(1):39–51
38. Thome N, Miguet S, Ambellouis S (2008) A real-time, multiview fall detection system: a LHMM-based approach. IEEE Trans Circuits Syst Video Technol 18(11):1522–1532
39. Toreyin BU, Dedeoglu Y, Cetin AE (2006) HMM based falling person detection using both audio and video. In: Proc. IEEE Signal Processing and Communications Applications
40. Willems G, Tuytelaars T, Van Gool L (2008) An efficient dense and scale-invariant spatio-temporal interest point detector. In: Proc. European Conf. Computer Vision
41. Yu X (2008) Approaches and principles of fall detection for elderly and patient. In: Proc. Int. Conf. E-health Network. Application Service
42. Yu M, Naqvi SM, Rhuma A, Chambers J (2012) One class boundary method classifiers for application in a video-based fall detection system. IET Comput Vis 6(2):90–100
43. Zhang T, Wang J, Liu P, Hou J (2006) Fall detection by embedding an accelerometer in cellphone and using KFD algorithm. Int J Comput Sci Netw Secur 6(10):277–284
44. Zigel Y, Litvak D, Gannot I (2009) A method for automatic fall detection of elderly people using floor vibrations and sound—proof of concept on human mimicking doll falls. IEEE Trans Biomed Eng 56(12): 2858–2867

Song-Zhi Su received the B.S. degree in Computer Science and Technology from Shandong University, China, in 2005. He received M.S. and Ph.D degree in Computer Science in 2008 and 2011, both from Xiamen University, Fujian, China. He joined the faculty of Xiamen University as an assistant professor in 2011. His research interests include pedestrian detection, time-of-flight camera based human action recognition and image/video retrieval.

Sin-Sian Wu received the B.S. degree in Computer Science from I-Shou University, Kaohsiung, Taiwan, in 2011 and the M.S. degree in Computer Science and Engineering from Yuan Ze University, Taoyuan, Taiwan, in 2014. Since 2014, he has served in military. His research interests include network management, image processing and pattern recognition.

Shu-Yuan Chen received the B. S. degree in Electrophysics in 1980, the M. S. and Ph.D. degrees both in Computer Engineering in 1982 and 1990, all from National Chiao Tung University, Hsinchu, Taiwan. Dr. Chen joined the faculty of Yuan Ze University (YZU), Taoyuan, Taiwan, in 1994 and has been an YZU Professor in the Department of Computer Science and Engineering from 2002. At YZU, Professor Chen has been the Head of the Department of Computer Science and Engineering from 2002 through 2004. Professor Chen's major research interests include image processing, pattern recognition, intelligent transportation systems, and image/video retrieval.

Der-Jyh Duh received the B.S. degree in Electrophysics from National Chiao Tung University in 1981, the M.S. degrees in Electrical Engineering from Tatung University in 1983 and the Ph.D. degree in Computer Science and Engineering from Yuan Ze University in 2005. He has been working as a senior system engineering at the Information and Communication Research Division of Chung Shan Institute of Science and Technology from 1983 to 2008. Dr. Duh joined the faculty of Chien Hsin University of Science and Technology, Taoyuan, Taiwan in 2008 and has been an assistant professor in the Department of Computer Science and Information Engineering from 2008. His research interests include image/video processing, pattern recognition, image compression and content retrieval.

Shao-Zi Li received the B.S. degree from the Computer Science Department, Hunan University in 1983, and the M.S. degree from the Institute of System Engineering, Xi'an Jiaotong University in 1988, and the Ph.D. degree from the College of Computer Science, National University of Defense Technology in 2009. He currently serves as the Professor and Chair of Cognitive Science Department of Xiamen University, the Vice Director of Fujian Key Lab of the Brain-like Intelligence System, and the Vice Director and General Secretary concurrently of the Council of Fujian Artificial Intelligence Society. His research interests cover Artificial Intelligence and Its Applications, Moving Objects Detection and Recognition, Machine Learning, Computer Vision, Natural Language Processing and Multimedia Information Retrieval, Network Multimedia and CSCW Technology and others.