電気学会論文誌 C（電子・情報・システム部門誌）
IEEJ Transactions on Electronics, Information and Systems
Vol.133 No.1 pp.117–127 DOI: 10.1541/ieejeiss.133.117

Paper

The Estimation of Heights and Occupied Areas of Humans from Two Orthogonal Views for Fall Detection

Dao Huu Hung[*a)] Non-member, Hideo Saito[*] Member

(Manuscript received April 12, 2012, revised Aug. 29, 2012)

In this paper, we present a video-based method of detecting fall incidents of the elderly living alone. We propose using the measures of humans' heights and occupied areas to distinguish three typical states of humans: standing, sitting, and lying. Two relatively orthogonal views are utilized, in turn, simplifying the estimation of occupied areas as the product of widths of the same person, observed in two cameras. However, the feature estimation based on sizes of silhouettes varies across the viewing window due to the camera perspective. To deal with it, we suggest using Local Empirical Templates (LET) that are defined as the sizes of standing people in local image patches. Two important characteristics of LET are: (1) LET in unknown scenes can be easily extracted by an automatic manner, and (2) by its nature, LET hold the perspective information that can be used for feature normalization. The normalization process is not only to cancel the perspective but also to take the features of standing people as the baselines. We realize that heights of standing people are greater than that of sitting and lying people. People in standing states also occupy smaller areas than whom in sitting and lying states. Thus, three humans' states fall into three separable regions of the proposed feature space, composing of normalized heights and normalized occupied areas. Fall incidents can be inferred from time-series analysis of human state transition. We test the performance of our method on 24 video samples in *Multi-view Fall Dataset* [(1)] leading to high detection rates and low false alarms, which outperform the state-of-the-art methods [(2)(3)] tested on the same benchmark dataset.

Keywords: Fall detection, orthogonal views, local empirical templates, normalized height, normalized occupied area, and time-series analysis of human state transition

1. Introduction

Nowadays, senior residents account for an increasing high percentage of the population, particularly in developed countries. Aging population is raising various social problems, in particular, health care services for the elderly. A recent study [(4)] shows that majority of the elderly live on their own and they are considered as an "at-risk" group. They appear to be associated with higher risks of accidental falls that are reported as the most common cause of injury for the elderly [(5)]. The instant treatment for injuries of fallen people is very critical, especially for the elderly. The degree of injury is proportional to the delay time in receiving medical treatments. Hence, we should detect the fall as soon as possible since accidental falls seem to be unavoidable and hard to be predicted. Timely responses help fallen people not worsen the injuries.

In the health care industry, there is a tremendous demand for supportive products and technologies to improve the safety at home [(5)]. One typical example is Personal Emergency Response System (PERS) [(6)] composing of a small radio transmitter, a console connecting to the user's telephone, and an emergency response center that monitors these types of calls. In emergency cases, users press the "help" button to contact with emergency response centers to receive immediate assistance. One weakness of PERS, which makes it inapplicable in assisting the elderly living alone is that users have to carry the "help" button 24 hours a day. But the elderly may easily forget to carry it all the time due to the dementia or the deterioration of cognitive ability. Moreover, the impact of shock after accidental falls may force fallen people to experience unconscious states of mind as well as physical pain. Pressing the "help" button to call for emergency assistance seems to be inappropriate in practice. Thus, the second generation of PERS [(6)] that is capable of providing automatic sensing of emergencies is in demand for the health care of the elderly. In this paper, we describe a video-based method of automatic fall-incident detection that plays as a central part in the second generation of PERS. Fall incidents are automatically detected for triggering the console to make an instant notification or contact with the emergency response center.

To develop an effective method of fall detection, it is essential to understand the concept of a fall as well as important characteristics of various kinds of fall. In general, a person is detected to be fallen if he/she changes from upright posture to the almost lengthened one in a

a) Correspondence to: Dao Huu Hung. E-mail: hungdaohuu@hvrl.ics.keio.ac.jp
* Department of Information and Computer Science
Keio University
3-14-1, Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan

rapid manner and subsequently remains relatively immobile in the latter posture for a while due to the shock impact of the fall [7]. Based on this definition, the fall detection method must

(1) identify three typical states of humans: standing, sitting, and lying.

(2) analyze the state transition to detect whether the person changes states directly from standing to lying. Normally, if the elderly want to take a rest on a bed or a sofa, they will firstly sit down and subsequently lie on the bed or on the sofa in a gentle way.

(3) extract the speed of the state transition. In contrast to the gentle manner of doing daily activities of the elderly, a fall often occurs in a very fast pace. Quantitatively, the fall usually lasts between 1 and 3 seconds [5].

(4) and finally verify the relative immobility in the lying state of fallen people. The human body can lie either on the ground floor or on some objects since in some cases he/she falls to the furniture, for example, a sofa or a bed, etc.

There are two typical scenarios of fall occurrences [5].

(1) People fall from sleeping (or bed and sofa) or sitting (or chair). In these two cases, when they try to get up or stand up, respectively, a fall happens probably due to the dizziness or syncope.

(2) People fall from standing or walking perhaps on account of loss of balance. This fall often occurs when people perform daily activities, i.e. carrying objects and doing housework, etc.

In daily life, fall incidents occasionally happen. Fall incidents must be carefully discriminated from like-fall events, i.e. sitting down brutally on a sofa, and kneeling on the ground, etc. Challenges of indoor video surveillance like dynamic lighting conditions, low contrast between humans' appearance and background, and occlusion by the furniture also pose considerable difficulties. Moreover, the initialization process of the method should be simple so that users without technical knowledge are able to set up and run the system easily.

Our contributions in this paper are three-folds.

(1) First, we propose using the measures of humans' heights and occupied areas to distinguish three typical states: standing, sitting, and lying. We realize the following relationship between the features and these states. Heights of standing people are greater than that of sitting and lying people. Moreover, people in standing states occupy smaller areas than whom in sitting and lying states.

(2) Two cameras whose fields of view are relatively orthogonal are utilized, in turn, to simplify the estimation of occupied areas roughly as the product of widths of the same person, observed in two views. This simplification facilitates the real-time performance of the proposed method. However, the heights and occupied areas estimated by the analysis of silhouettes' sizes vary across the viewing windows due to the camera perspective.

(3) To handle the camera perspective, we suggest using Local Empirical Templates (LET), which have demonstrated to be effective in dealing with this issue [8]. By definition, LET are the sizes of foregrounds typically made by a standing person in local image patches. Therefore, the automatic LET extraction for unknown scenes is straightforward [8][9]. We observe that LET in local image patches far from the camera are smaller than LET in patches close to the camera. By its nature, LET hold the perspective information that can be used for perspective normalization of estimated heights and occupied areas. The normalization process serves two purposes: (1) to cancel the perspective and (2) to take the features of standing people as the baselines. It creates the distance measures between features of detected people and the appropriate LET (in standing states). In consideration of the above feature-state relationship, three states of humans fall into three separable regions of the proposed feature space, composing of normalized heights and normalized occupied areas, which can be classified by using support vector machines.

In performance evaluation and comparison, a recommended methodology is to test the method on a common benchmark dataset [9]. This facilitates a fair comparison with other existing methods. In this paper, we choose *Multi-view fall dataset*," recently released in public for scientific communities by Université de Montréal [1] in Canada for the evaluation of our method. Our proposed method outperforms two state-of-the-art methods [2][3] tested on the same benchmark dataset.

We continue this paper with the provision of related works in Section 2. Section 3 describes the details of our proposed method. A brief introduction to the benchmark dataset, experiments and performance comparison are reported in Section 4. Finally, we make concluding remarks and future works in Section 5. The early version of this paper appeared in FCV2012 [10].

2. Related Works

In this section, existing fall detection methods are reviewed to delineate the context of our proposed method. There are three kinds of technologies used to develop fall detection methods: wearable devices, ambient devices and video processing. Making a review of wearable and ambient devices-based methods goes beyond the scope of this paper. Please refer to two survey papers [5][7] for your special interests in these methods. Only video-based methods are taken into consideration in this Section. We classify them into two categories of using 2D and 3D information.

In the former category, the early work of Anderson *et al.* [11] analyzed the sizes of human body silhouettes. The width-to-height ratios or aspect ratios of humans in standing and lying states are large and small, respectively. However, this observation may not be true in consideration of the effect of human body upper limb activities. To eliminate this effect, Liu *et al.* [12] used a sta-

tistical scheme to remove peaks in vertical histograms of silhouette images. They proposed K-NN classifier working with a feature space composing of the aspect ratios and the difference between height and width of silhouettes. Huang *et al.*[13] introduced the combination of speed and aspect ratios of silhouettes to discern fall incidents. Both three methods are prone to false alarms with a fall in parallel to the optical axis of cameras. Occlusion caused by other objects, i.e., the furniture, is also not taken into account. They merely reported the experiments with cameras placed sideways. In practice of indoor surveillance, the camera is preferred to be in oblique settings for wider views and occlusion avoidance. Shoaib *et al.*[14] presented a context model to learn the head and floor planes from the foregrounds of a moving person in the scene. Distance measures between detected heads and referenced heads, provided by the context model, are adopted as a discriminative feature. In general, it is able to distinguish bending and sitting actions from a fall, except the one in parallel to the optical axis.

Lee and Mihailidis[15] labeled furniture areas in the image captured from a top-view camera as inactivity zones, i.e. chair, sofa, and bed, etc. The speed of silhouettes' centroids is extracted and analyzed by applying special thresholds for different inactivity zones. In the similar experimental scenarios, Charif and McKenna[16] argued that there are few places in a room in which people are relatively inactive most of the time for relaxing activities, i.e., watching television, reading newspaper, and drinking tea, etc. They are tracked and checked whether they are inactive in a known inactivity zone. Their immobility outside known inactivity zones is more likely caused by fall occurrences. However, both systems expose several limitations. Firstly, fall occurrences in inactivity zones are not taken into consideration. Secondly, the speed estimation of 2D silhouettes is highly sensitive to cluttered background and daily activities of humans. Finally, using top-view cameras seems to be inappropriate for the problem of fall detection since crucial clues from the vertical motion of human body to recognize a fall is not available.

Motion History Image is adopted to quantify the motion of human body[17]. Large motion is more likely caused by fall incidents. Silhouettes are approximated by eclipse models whose orientation angle and the ratio of major to minor semi-axes are utilized to discriminate fall incidents from other events, including the like-fall ones, i.e., sitting down brutally and kneeling. Similarly, integrated spatiotemporal energy map is used for the calculation of motion activity coefficients to detect large motion events[18]. Orientation angle, displacement, and major-to-minor-semi-axes ratio of human eclipse models are analyzed in the framework of Bayesian Belief Networks to recognize fall and slip-only events. Chen *et al.*[19] presented a combination of distance map of two sampling human skeletons and variation analysis of eclipse human models. Rougier *et al.*[2] supposed that human shape should change progressively and slowly during usual activities, and drastically and rapidly during a fall. Hence, shape-matching costs during a fall and a usual activity are high and low, respectively. The method is reported to work with the frame rate of 5 fps due to the expense of high computational cost.

Apparently, methods in the former category have several limitations in terms of camera viewpoints, occlusion, and falls in parallel to the optical axis, etc. The reason is that the combination of 3D spatial features and temporal structures of actions, which is powerful in representing and recognizing human activities are not made use of. In the latter category, Cucchiara *et al.*[20] used a caliberated camera to train probabilistic projection maps for each posture, i.e. standing, crouching, sitting, and lying. They suggested using a tracking algorithm with a state-transition graph to handle occlusion, in turn, leading to reliable classification results. In their latter work[21], partial occlusion is detected and compensated by a wrapping method from multiple cameras. A Hidden Markov Model (HMM) is trained for obtaining more robust recognition results. Posture recognition is carried out by using 3D human centroid distance from the floor plane, extracted from a calibrated camera, and the orientation of the body spine[23]. Thome *et al.*[24] applied the metric image rectification to derive the 3D angle between vertical line and principal axis of eclipse human models. Decisions made independently by multiple cameras are fused in a fuzzy context to classify postures. Layer HMM is hand designed to make event inference. Anderson *et al.*[25] introduced a framework of fall detection in the light of constructing voxel person. Linguistic aspect of the hierarchy of fuzzy logic used in this research for fall inference makes this framework extremely flexible, allowing for user customization based on their knowledge of cognition and physical ability. Recently, Auvinet *et al.*[3] discussed a method of reconstructing 3D human shape from a network of cameras. They proposed the idea of Vertical Volume Distribution Ratio since volumes of standing and lying-down person are vertically distributed significantly differently. The method is able to handle occlusion since the 3D reconstructed human shape is contributed from multiple cameras.

3. Our Proposed Method of Fall Detection

To develop an effective method of fall detection, we take advantages of the latter category of approaches by combining 3D spatial information and temporal structure of actions. The measures of humans' heights and occupied areas form a feature vector to classify three typical humans' states: standing, sitting, and lying. Fall incidents are discriminated from other usual activities by a time-series analysis of human state transition. Fig. 1 shows the flowchart of our proposed method.

In order to ensure the real-time performance, we suggest using two orthogonal views to simplify the computation of occupied areas. The two cameras are in oblique viewpoint settings and their fields of view are relatively orthogonal. The video sequences are processed by Gaussian Mixture Models (GMM)[26] to segment foregrounds for detecting people. The sizes of people are extracted

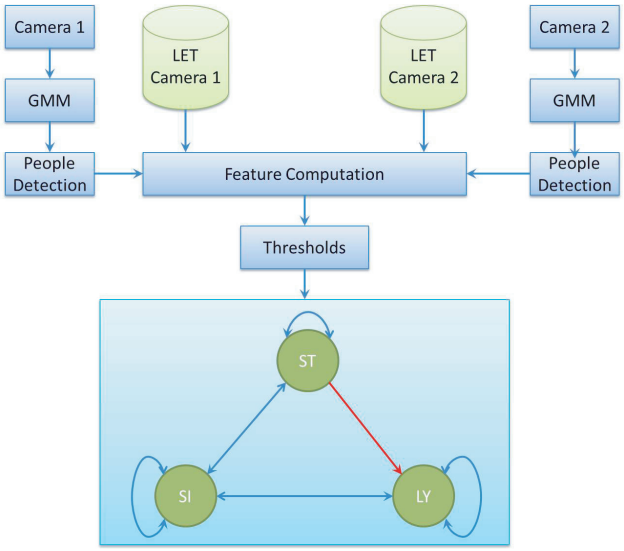


Fig. 1. The flowchart of our proposed method

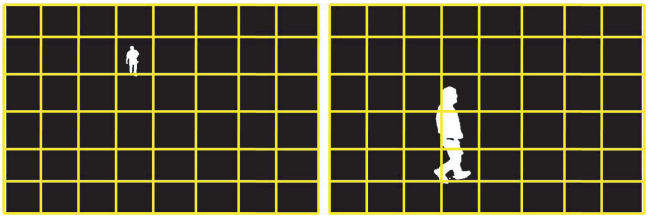


Fig. 2. Local Empirical Templates (The sizes of cells in this figure are for demonstration purpose)

from two cameras and fused to compute the feature vector for discriminating humans' states. In the followings, we describe the key modules of our proposed method in details.

3.1 Local Empirical Templates Local Empirical Templates (LET) are important in perspective normalization process and the people detection algorithm. In this section, we introduce the definition, important characteristics of LET and an automated way of extracting LET for unknown scenes. As indicated in the flowchart, LET of the working scene must be available before the whole method works. The process of LET extraction can be considered as the initial setup of our proposed method. However, it is straightforward and can be done in an automated way [8] [9].

We divide the working scene into many cells, as shown in Fig. 2. LET is defined as the sizes of standing people in local image patches or local cells. There is one LET reflecting the typical size of standing people in each cell. Our observations in Fig. 2 are that the size of the man in the left image is small since he is far from the camera. Meanwhile, his size in the right image is bigger since he appears close to the camera. These definition and observations lead to two important characteristics of LET. (1) LET, by its nature, hold the information of the camera perspective. (2) LET extraction for unknown scenes is straightforward and can be done automatically since LET are merely the sizes of standing people in local image patches [8] [9].

Suppose that the scene is divided into $M \times N$ cells so

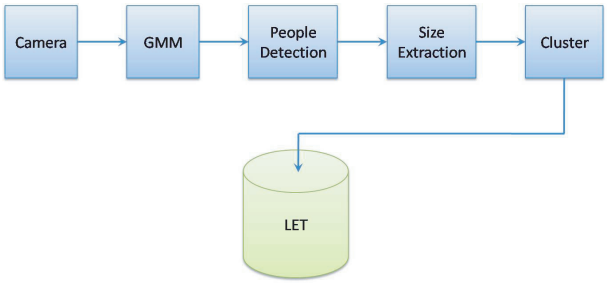


Fig. 3. The flowchart of LET extraction process

that the sizes of people must be nearly constant in each cell. The number of cells depends upon the resolution and the viewpoints of cameras. It is common in practice that LET does not appear fully in one cell but in several cells as shown in Fig. 2. Thus, we define the LET for the cell (i, j) as the following:

$$T(i,j) = \{W_T(i,j), H_T(i,j)\} \cdots\cdots\cdots\cdots\cdots\cdots (1)$$
$$i \in [1, M], j \in [1, N]$$

where, $T(i,j)$ the LET whose head appears in the cell (i,j), $W_T(i,j)$ and $H_T(i,j)$ width and height of the LET, respectively.

The fall detection method is dedicated to a specific elderly person. LET should be the sizes of the monitored elderly person to improve the accuracy of the method. In this paper, we adopt the automated way of LET extraction for unknown scenes, presented by Hung *et al.* [8] [9] To do that, we capture the foregrounds and trajectories of the monitored elderly person moving around the scenes. The sizes of foregrounds are extracted and kept in each cell buffer for clustering to generate an appropriate LET for the cell. The initial setup is straightforward and can be performed in an automated manner [8] [9]. It allows users without technical knowledge to customize the fall detection system for the different elderly and under various camera viewpoints. The flowchart of LET extraction process is shown in Fig. 3.

3.2 People Detection Foreground, segmented by GMM [26], is enhanced by applying morphological operators such as open and close to eliminate pepper noise before being labeled by connected component algorithms (CCA). Isolated foreground regions labeled by CCA are so-called blobs. After these preprocessing steps, a pool of N blobs $\{B_1, B_2, ..., B_N\}$ is created for the algorithm of people detection.

We search in the pool of blobs to find a head candidate and then group blobs in the neighborhood of the head candidate to form a person. The common labeling order of CCA is from top to bottom and subsequently from left to right of images. People are supposed to be in upright poses. Consequently, the blob with smallest label is more likely the head candidate. LET of the cell in which the head candidate appears provides the tentative size of detected person $\{W_T, H_T\}$ or the tentative area in which the person appears. All blobs whose centroids satisfy the spatial constraint posed by LET more likely belong to the person. They are grouped together for accumulating their densities and extracting the bound-

```
LOOP
    IF N > 0
        Head_Candidate ← Blob with smallest index = B_si
        Density ← Density(B_si)
        P ← Position(B_si) = (m, n)
        LET ← T(m, n) = {W_T, H_T}
        Spatial_Constraint ← (m, n, W_T, H_T)
        IF N > 1
            sum ← 0
            LOOP in N blobs
                IF B_i satisfies Spatial_Constraint
                    Select B_i for grouping
                    Update the boundaries of detected person
                    Density ← Density + Density(B_i)
                    Remove B_i from the pool of blobs
                    sum ← sum + 1
                END
            END LOOP
            N ← N − sum
            Update the pool of blobs
        END
        Take density ratio D by Eq. 2
        IF D > Threshold
            Confirm 'A person is detected'
            Mark a rectangular box for detected person
        END
    ELSE
        Exit LOOP
    END
END
```

Fig. 4. The algorithm of detecting people from the pool of blobs

aries. We take the ratio of the total density to the size of the appropriate LET by the following formula [8].

$$D = \frac{Total\_Density}{W_T \times H_T} \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (2)$$

We confirm a detection if the density ratio exceeds a particular threshold. Please refer to Table 1 in the study of Hung *et al.* [8] for selecting the threshold of 0.3. We update the pool of blobs by removing blobs of detected people. In the next search, the head candidate is associated with the blob with the smallest label remaining in the pool. The process of searching for head candidates and grouping blobs in the neighborhood of head candidates is continued until there is no blob remaining in the pool. The algorithm of people detection is summarized as pseudo code in Fig. 4.

3.3 Feature Computation In this paper, we propose using the feature vector composing of the measures of humans' heights and occupied areas to discriminate three typical states: standing, sitting, and lying. We realize that it is not necessary to estimate exactly how many squared meters a person is in. An approximate estimation is good enough for this application, in turn, facilitating the real-time performance. To this end, we suggest using two orthogonal views to simplify the estimation of humans' occupied areas. Two cameras are in oblique viewpoint settings whose fields of view are relatively orthogonal, as shown in Fig. 5. The occupied areas are roughly estimated by the product of the widths of a same person observed in the two orthogonal views. Suppose that the person appears in the cell (m, n) in the first view with the size of $\{W_1(m,n), H_1(m,n)\}$. We also observe this person in the cell (p, q) in the second view

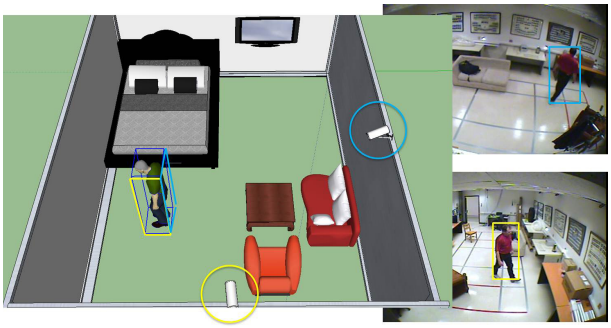


Fig. 5. Two orthogonal views for the estimation of humans' occupied areas

with the size of $\{W_2(p,q), H_2(p,q)\}$. The occupied area is estimated as the following.

$$OA(m,n,p,q) = W_1(m,n) \times W_2(p,q) \cdots\cdots\cdots (3)$$

However, the estimated occupied areas by Eq. 3 vary across the viewing window because of the camera perspective. As discussed in Section 3.1, LET hold the perspective information and can be used for the perspective normalization. We extract the LET $T_1(m,n) = \{W_{T1}(m,n), H_{T1}(m,n)\}$ in the cell (m,n) in the first view and $T_2(p,q) = \{W_{T2}(p,q), H_{T2}(p,q)\}$ in the cell (p,q) in the second view. The occupied area of LET can be estimated by the following formula.

$$OA_{LET}(m,n,p,q) = W_{T1}(m,n) \times W_{T2}(p,q) \cdots (4)$$

We take the ratio of the occupied area of detected person to that of an appropriate LET for perspective normalization, leading to a promising feature, so-called normalized occupied area NOA.

$$NOA = \frac{OA(m,n,p,q)}{OA_{LET}(m,n,p,q)} = \frac{W_1(m,n) \times W_2(p,q)}{W_{T1}(m,n) \times W_{T2}(p,q)} \cdots\cdots\cdots\cdots (5)$$

It is noted that LET are defined as the sizes of a standing person appearing in the vicinity of detected person. The normalization in Eq. 5 is not only to cancel the perspective but also to take the features of standing people as the baselines. In other words, the normalization measures the distance between the features of detected people and the appropriate LET (in standing states). Therefore, NOA is both lower and upper bounded and does not depend on the cell index. The cell-index notation of NOA in Eq. 5 are removed for simplicity. NOA is also highly relevant to the three typical states because of the feature-state relationship. A person lying on the ground occupies a larger area than standing and sitting. The occupied area in sitting state is, in general, larger than that in standing states.

$$NOA_{Standing} < NOA_{Sitting} < NOA_{Lying} \cdots (6)$$

However in practice, poor foreground segmentation, occlusion, and human body upper limb activities, might cause the estimation of NOA in standing and sitting

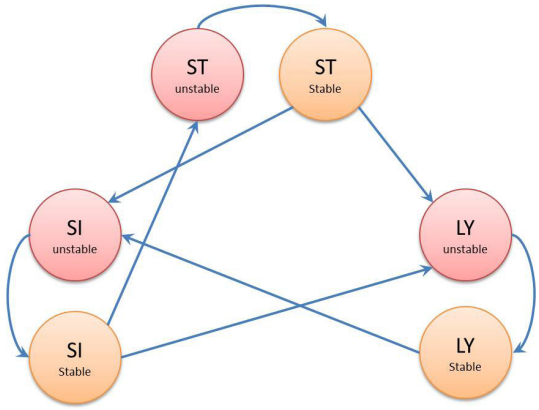


Fig. 6. Time-series human state transition

states by Eq. 5 to be quite similar. Fortunately, the humans' heights are significantly different and can be used to discriminate standing states from sitting and lying states.

The estimation of humans' heights is highly sensitive to the occlusion that is frequently happened in the indoor environments mainly by the furniture. However, under two relatively orthogonal views, we realize that people are partially occluded in one view but likely visible in the other one. The height of people should be the maximum of the height estimations from the two views. Like the feature of occupied area, the height must be normalized to that of an appropriate LET for perspective cancellation. The estimation of normalized height is summed up by the following formulae.

$$NH_1 = \frac{H_1(m,n)}{H_{T1}(m,n)}$$
$$NH_2 = \frac{H_2(p,q)}{H_{T2}(p,q)}$$
$$NH = Max(NH1, NH2) \quad \cdots\cdots\cdots\cdots\cdots\cdots (7)$$

In summary, we have the feature space composing of normalized heights and normalized occupied areas that is separable for three typical states of humans. We will discuss and demonstrate this property in Section 4.2.

3.4 Fall Event Inference It is impossible to recognize human actions in a single frame or few frames since actions have temporal structures. Hence to make the fall event inference, we eye on the states of the elderly person in a period of time. In this paper, three typical states Standing (ST), Sitting (SI) and Lying (LY) are taken into consideration. A time-series analysis of human state transition shown in Fig. 6 that is similar to the state transition graph in the study of Cucchiara *et al.* [20] is adopted. Table 1 sums up all actions, which can be inferred from the time-series analysis of human state transition. In general, all state transitions are allowed. However, for the specific application dedicated to the elderly, the direct transition from LY to ST states is quite improbable. The elderly often make the transitions in a gentle way from LY to SI and then to ST states.

Suppose that we keep states of the monitored elderly person in N frames for making event inference in a probabilistic manner. The *instant state* classified in each frame is not reliable for detecting state transitions. Therefore, we suggest using *stable states* and *unstable states*, instead of *instant states*. Only one out of three states, i.e., ST, SI, and LY, appearing in the window of N frames with the highest probability, is the stable state. The others are defined as unstable states.

$$Stable\_State = argmax_x\{P(x); ST, SI, LY\} \quad \cdots\cdots\cdots\cdots\cdots\cdots (8)$$

where, $P(x)$ the probability of observing the state x in the window of N frames, evaluated by frequentist paradigm, with $x \in \{ST, SI, LY\}$. Direct transitions between two stable states are not allowed. A state transition must undergo an unstable state before reaching its corresponding stable state, as illustrated in Fig. 6. When a state transition is in progress, the probability of observing the current stable state gradually decreases. Meanwhile, the probability of observing one of the other unstable states slightly increases. The state transition is confirmed upon the generation of a new stable state by Eq. 8.

In this paper, we are interested in detecting fall incidents rather than other events. In consideration of the definition and characteristics of a fall as discussed in Section 1, a fall event can be inferred by a direct transition from standing to lying states and subsequently an observation of staying in the lying state in some moments. Therefore, we dedicate a special attention on the aftermath of such state transitions to confirm a fall by

Table 1. Actions can be inferred from the time-series analysis of human state transition

Current States	Next States		
	ST	SI	LY
ST	Standing or Walking	Sitting down	**Falling**
SI	Standing up	Sitting	Lying down
LY	*NA*	Getting up	Lying

START
  Update the pool of N states
    Delete the oldest state
    Add the latest state
  $stable\_state = argmax_x\{P(x); ST, SI, LY\}$
  IF $(current\_stable\_state == ST)\&(stable\_state == LY)$
    A Fall probably happened
    $start\_counter \leftarrow true$
    $counter \leftarrow 0$
    $current\_stable\_state \leftarrow stable\_state$
  END
  IF $start\_counter == true$
    IF $(stable\_state == LY)\&(current\_stable\_state == LY)$
      $counter \leftarrow counter + 1$
    END
  END
  IF $counter > Threshold$
    Confirm the Fall
    $start\_counter \leftarrow false$
  END
  IF Other state transitions happen
    $current\_stable\_state \leftarrow stable\_state$
  END
END

Fig. 7. The time-series analysis of human state transition

(a) falling forward (b) falling backward (c) carrying object (d) putting off the coat

(e) losing balance (f) failing to sit on a chair (g) lying on a sofa (h) sitting

(i) falling to the sofa (j) falling from the sofa (k) doing housework (l) crouching with occlusion

Fig. 8. Examples of typical fall incidents and confounding events

verifying the duration of staying in the lying state after the state transition happened. The time-series analysis of human state transition to make inference of fall incidents is summarized as pseudo code in Fig. 7.

4. Performance Evaluation and Comparison

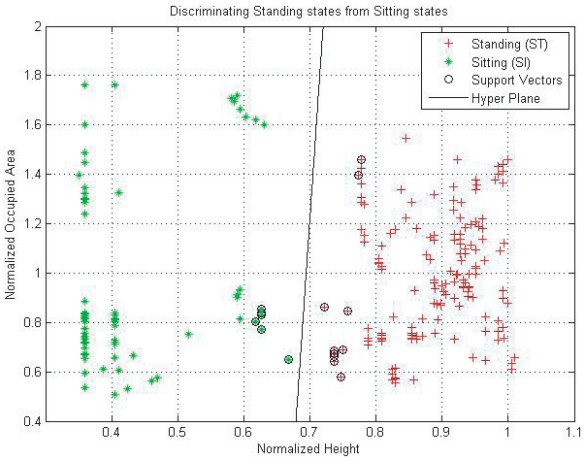
4.1 Multi-view Fall Dataset For fair comparisons with existing methods of fall detection, the recommended methodology of performance evaluation is to conduct the experiments on the same dataset [9]. In this paper, we use the "Multi-view fall dataset" recently released by Auvinet *et al.* [1], which was adopted in the experiments of two latest studies [2] [3]. Consequently, it is fair to compare the performance between our proposed method and the two methods.

For the effort of making this dataset available in public for research purposes, real fall situations are not applicable due to the issue of privacy protection. Simulated falls were performed by an experienced clinician in the field of health care for the elderly, and were spotted simultaneously from eight inexpensive IP cameras with a wide angle to cover all the room. Consequently, the images are highly distorted. The dataset consists of 24 scenarios showing 24 fall incidents and 24 confounding events (11 crouching, 9 sitting, and 4 lying on a sofa). Various kinds of falls are demonstrated, i.e. falling forward, falling backward, losing balance, and falling to furniture. Confounding events include crouching, kneeling, carrying objects, and doing housework, etc. Camera settings and spatial arrangement, information of multi-camera synchronization, calibration parameters, and event annotation for all scenarios are provided in their study [1]. Examples of typical simulated fall incidents and confounding events are shown in Fig. 8. Video sequences from Cameras 2 and 5 are used in this paper since these two views are relatively orthogonal.

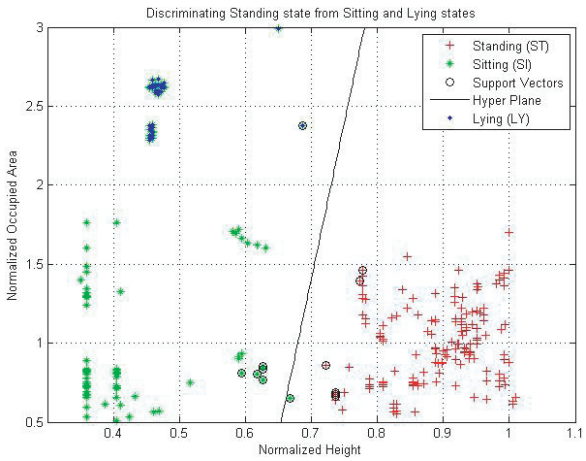
4.2 Linearly Separable Feature Space It is stated in Section 3.3 that the proposed feature space is separable for three typical states of humans. This section will discuss, demonstrate this statement, and find the decision boundaries for the state classification.

Since LET are defined as the sizes of standing people in local image patches, the normalization process takes the features of standing people as the baselines. It creates the distance measures between the features of detected people and the appropriate LET (in standing states). Thus, normalized heights, NH, of standing people should be approximate 1. For people in sitting and lying states, normalized heights are much smaller than 1. It is possible to distinguish standing states from sitting and lying states only based on the feature of normalized height. To discriminate lying states from sitting states, occupied area is a strong discriminative feature. Apparently, a person in lying states occupies a larger area than in sitting states. As a result, there exist two linear decision boundaries separating the feature space for three states of standing, sitting, and lying.

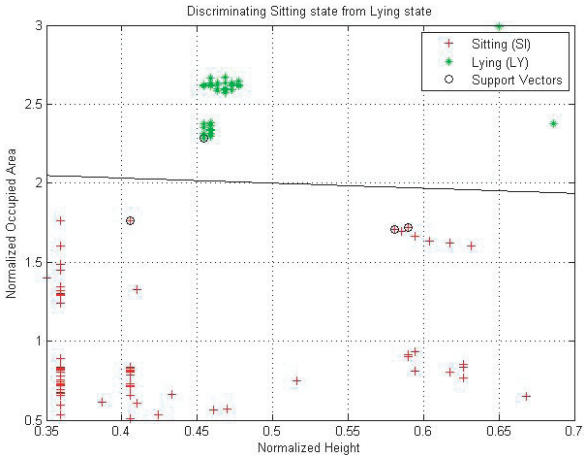
To demonstrate our discussion and to find the decision boundaries, we use the ninth scenario of the dataset for training purpose. In this scenario, the man approaches to the chair after entering the scene. He sits on the chair



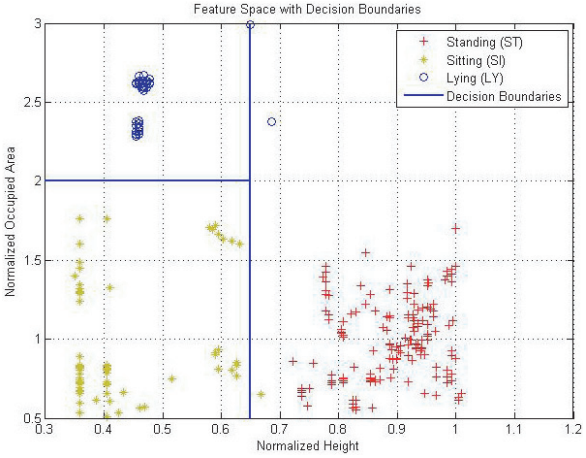
(a) Discriminating ST state from SI state



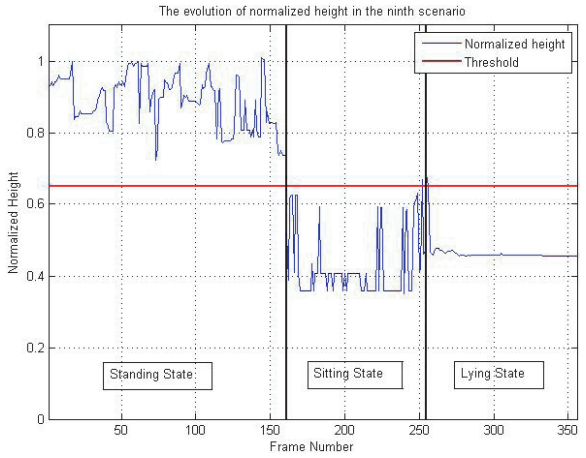
(b) Discriminating ST state from SI and LY states



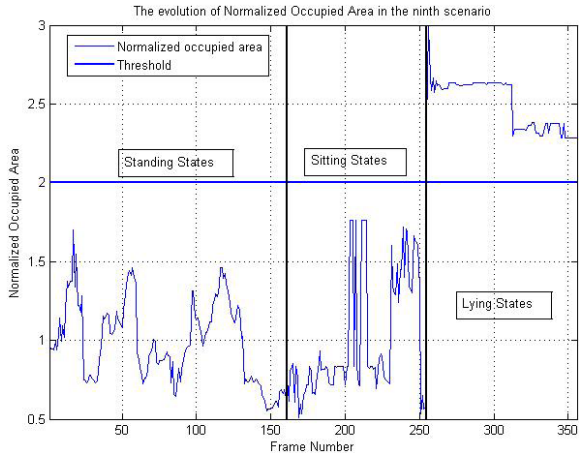
(c) Discriminating SI state from LY state



(d) Feature Space with modified decision boundaries



(e) Time-series evolution of normalized height



(f) Time-series evolution of normalized occupied area

Fig. 9. Feature space of the ninth scenario with decision boundaries found by support vector machines

for a while and stands up before falling to the ground. The annotation of this scenario provides the state label in each frame. We calculate the feature vectors for every frame in combination with the corresponding state labels to create the training data.

Both the training data sketched in the feature space in Fig. 9 (d) and the above discussion show that it can be linearly separated. Therefore in this paper, two-class support vector machines (SVM) are adopted to find the decision boundaries for separating the three states. We make three following experiments in training SVM to find the decision boundaries. Firstly, standing states are separated from sitting states by a nearly vertical line in Fig. 9 (a). Secondly, we combine sitting and lying states as one class. The second class of SVM is the standing state. The decision boundary separating the two classes is given in Fig. 9 (b). Thirdly, the decision boundary for sitting and lying states is found in Fig. 9 (c) as a nearly

horizontal line.

The results of experiments in training SVM quite fit to our above discussion, except the one in Fig. 9 (b). However, it is clear to see some outliers in the training data of lying states, impairing the obtained decision boundary in Fig. 9 (b). The decision boundary in Fig. 9 (a) indicates that normalized heights of people in sitting states cannot be greater than 0.7. This observation is also true for normalized heights of people in lying states. However, the decision boundary in Fig. 9 (b) creates a region in which normalized heights of people in both sitting and lying states are well greater than 0.7. It is not reasonable in practice since the heights of people in sitting and lying states must be much smaller than in standing states. Therefore, the decision boundary for separating standing states from sitting and lying states in Fig. 9 (b) should be a nearly vertical line, like the one in Fig. 9 (a). We make the modification for the obtained decision boundaries based on our prior knowledge of humans' heights, as shown in Fig. 9 (d). It leads to the generation of the thresholds for normalized heights and occupied areas, being 0.65 and 2, respectively. Fig. 9 (e) and Fig. 9 (f) show the time-series evolution of normalized heights and normalized occupied areas with obtained thresholds in the ninth scenario, respectively. In Fig. 10, we provide the visual results of state classification of the first and third scenarios in *Multi-view fall dataset* and the time-series evolution of each feature to further confirm the validation of the obtained thresholds.

Width and height are estimated by the horizontal and vertical sides of rectangular bounding boxes, respectively, not depending on human states, standing, sitting, and lying. Therefore, estimated widths of people in sitting and lying states are usually larger than actual ones, so-called the estimation error. However, such estimation errors only make the estimation of NOA of people in sitting and lying states larger. It fits to our observation in Eq. 6. Hence, such estimation errors do not influence on the performance of our method. In addition, the state classification implemented by using SVM is also to deal with the estimation errors. In experiments on the dataset containing limited challenges in the real world, we do not find any case in which the estimation errors affect to the performance of our method.

4.3 Performance Evaluation and Comparison

To evaluate the performance of our method and to compare it with two state-of-the-art methods [2] [3], tested on the same dataset, we compute *sensitivity* and *specificity* as the follows.

$$Se = \frac{TP}{TP + FN}$$
$$Sp = \frac{TN}{TN + FP} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (9)$$

where

(1) Se, the sensitivity
(2) Sp, the specificity
(3) TP, True Positive, the number of falls correctly detected
(4) FN, False Negative, the number of falls not detected
(5) TN, True Negative, the number of normal activities not detected as a fall
(6) FP, False Positive, the number of normal activities detected as a fall

Table 2. Performance comparison between our method and two state-of-the-art methods [2] [3], tested on the same dataset

	Sensitivity (Se)	Specificity (Sp)
Our method	**95.8%**	**100%**
Auvinet *et al.* [3]	80.6%	100%
Rougier *et al.* [2]	95.4%	95.8%

High sensitivity means that most fall incidents are correctly detected. Similarly, high specificity implies that most normal activities are not detected as fall events. A good fall detection method must achieve high values of sensitivity and specificity.

Our method detects 23 out of 24 fall incidents in the whole dataset. It only fails in the 22nd scenario in which the person is sitting on a chair and suddenly slips to the floor. Our method recognizes it as the lie-down event instead of a fall incident. No normal activity detected as a fall is reported in our experiments. The *sensitivity* and *specificity* are 95.8% and 100%, respectively.

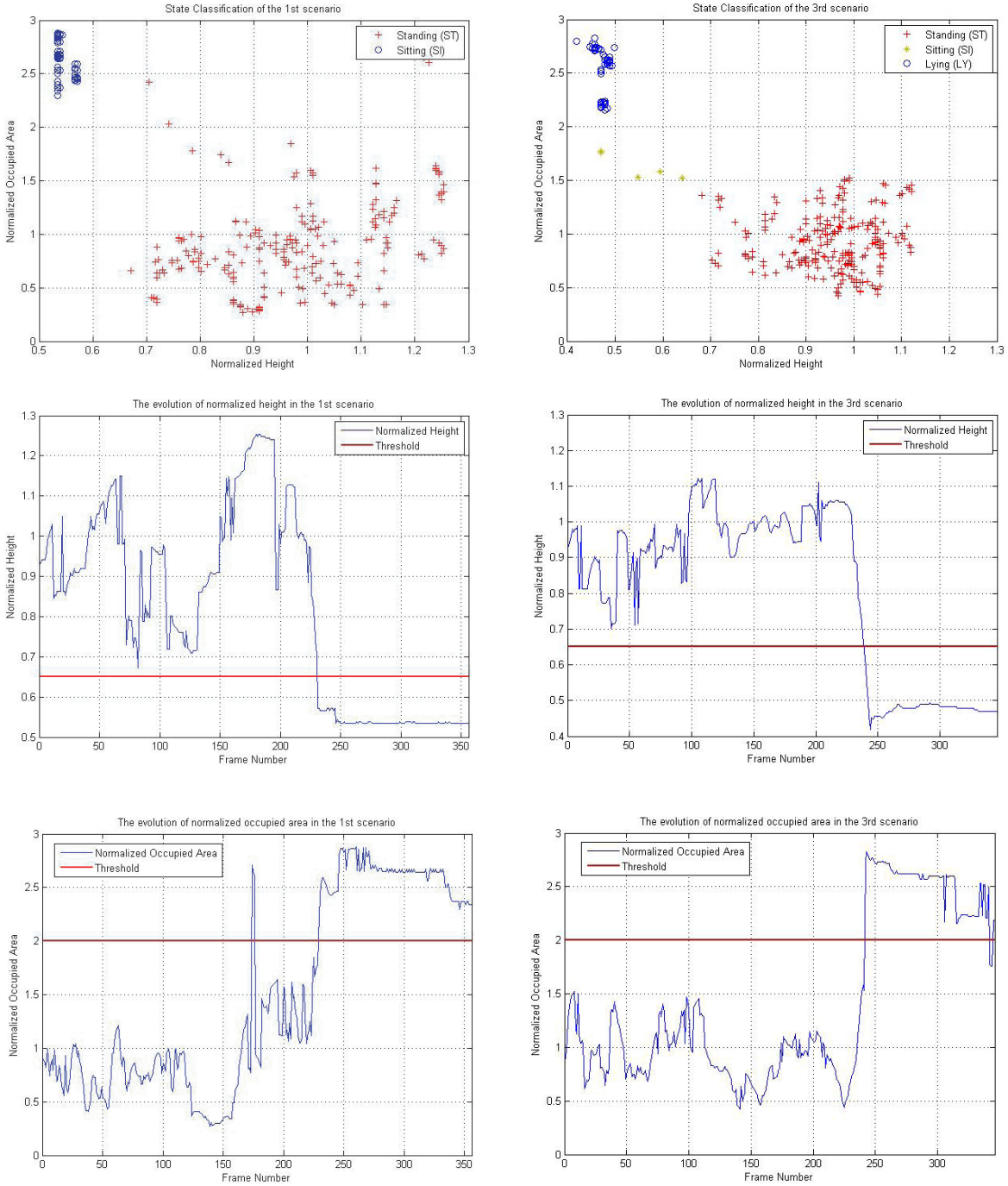
We compare the performance between our method and two state-of-the-art methods [2] [3], tested on the same dataset, in Table 2. It is noted that the results of the method proposed by Auvinet *et al.* [3] are reported with a network of three cameras. The sensitivity can be boosted to 100% if a network of more than four cameras is employed. However, both methods are high computational costs. Rougier *et al.* [2] reports the implementation of 5 fps and argues that this frame rate is sufficient for detecting fall events. Auvinet *et al.* [3] presents the GPU implementation to realize their method in real-time. Meanwhile, our method composing of low-cost modules is implemented in real-time in a common desktop PC [†] and achieves very competitive performance.

5. Conclusions

We have presented a novel method of fall detection that plays as a central part of the second generation of PERS for aiding the elderly living alone. The novelty lies in the feature space composing of humans' heights and occupied areas to discriminate three typical states of humans, i.e. standing, sitting and lying. It is the fact that the heights of people in standing states are greater than in sitting and lying states. Moreover, People in lying states occupy a larger area than in sitting and standing states. Therefore, the proposed feature space is linearly separable for these three states. Fall incidents can be inferred from the time-series analysis of human state transition.

In implementation, our aim is to develop the method with simple but effective modules to achieve both real-time and good discrimination performance. We propose using two orthogonal views: (1) to simplify the estimation of occupied area, and (2) to improve the reliability

[†] CPU: Intel Core i7 950 3.07 GHz, 3 GB Ram



(a) The 1st scenario (b) The 3rd scenario

Fig. 10. State Classification and the time-series evolution of normalized height and occupied area of the 1st and 3rd scenarios

of computing the feature vector based on sizes of silhouettes in the presence of occlusion. People are partially occluded in one view but visible in the other one. The feature vector is normalized by the size of an appropriate LET to cancel the camera perspective and to realize the linear separability of the proposed feature space.

In performance evaluation, a good method of fall detection is associated with high *sensitivity* and *specificity*. We choose *Multi-view fall dataset* that only includes simulated falls by an experienced clinician in the health care for the elderly, to test our method for fair comparison with existing methods. The results of our method reach to 95.8% of *sensitivity* and 100% of *specificity*. It outperforms two state-of-the-art methods [2] [3], tested on the same dataset. In the future work, we will test on real falls of the elderly in real home environments to further validate the performance of our proposed method, especially the influence of the estimation errors. We also take the situation of falling from sitting positions like the one in the scene 22 into account by labeling the furniture areas. If the person is sitting or lying *in the furniture areas*, i.e., sofa or bed, and *suddenly lying on the ground*, such lying-down events should be detected as falls.

References

(1) E. Auvinet, C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau: "Multiple cameras fall data set", Technical Report 1350, DIRO Université de Montréal (2010)

(2) C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau: "Robust video surveillance for fall detection based on human shape deformation", IEEE Trans. on Circuits and Systems for Video Technology, Vol.21, No.5, pp.611-622 (2011)

(3) E. Auvinet, F. Multon, A. St-Arnaud, J. Rousseau, and J. Meunier: "Fall detection with multiple cameras: An occlusion-resistant method based on 3D silhouette vertical distribution", IEEE Trans. on Information Technology in Biomedicine, Vol.15, No.2, pp.290-300 (2011)

(4) K. Kharicha, S. Iliffe, D. Harari, C. Swift, G. Gillmann, and A. E. Stuck: "Health risk appraisal in older people 1: are older people living alone an 'at-risk' group?", British Journal of General Practice, Vol.57, pp.271-276 (2007)

(5) X. G. Yu: "Approaches and principles of fall detection for elderly and patient", Proc. of 10th IEEE Int'l Conf. on e-Health Networking, Applications and Services, pp.42-47 (2008)

(6) K. Doughty, K. Cameron, and P. Garner: "Three generations of telecare of the elderly", Journal of Telemedicine and Telecare, Vol.2, No.2, pp.71-80 (1996)

(7) N. Noury, A. Fleury, P. Rumeau, A. K. Bourke, G. Ó. Laighin, V. Rialle, and J. E. Lundy: "Fall detection: principles and methods", Proc. of 29th IEEE Int'l Conf. on EMBS, pp.1663-1666 (2007)

(8) D. H. Hung, S. L. Chung, and G. S. Hsu: "Local empirical templates and density ratios for people counting", Proc. of 10th Asian Conf. on Computer Vision, Vol.4, pp.90-101 (2010)

(9) D. H. Hung, G. S. Hsu, S. L. Chung, and H. Saito: "Real-time people counting in crowded areas by using local empirical templates and density ratios", IEICE Trans. on Information and Systems, Vol.E95-D, pp.1791-1803 (2012)

(10) D. H. Hung and H. Saito: "Fall detection with two cameras based on occupied areas", Proc. of 18th Japan-Korea Joint Workshop on Frontier in Computer Vision, pp.33-39 (2012)

(11) D. Anderson, J. M. Keller, M. Skubic, X. Chen, and Z. H. He: "Recognizing falls from silhouettes", Proc. of 28th IEEE Int'l Conf. on EMBS, pp.6388-6391 (2006)

(12) C. L. Liu, C. H. Lee, and P. M. Lin: "A fall detection system using k-nearest neighbor classifier", Expert Systems with Applications, Vol.37, pp.7174-7181 (2010)

(13) B. Huang, G. H. Tian, and X. L. Li: "A method for fast fall detection", Proc. of World Congress on Intelligent Control and Automation, pp.3619-3623 (2008)

(14) M. Shoaib, R. Dragon, and J. Ostermann: "View-invariant fall detection for elderly in real home environment", Proc. of the 4th Pacific-Rim Symposium on Image and Video Technology, pp.52-57 (2010)

(15) T. Lee and A. Mihailidis: "An intelligent emergency response system: preliminary development and testing of automated fall detection", Journal of Telemedicine and Telecare, Vol.11, No.4, pp.194-198 (2005)

(16) H. N. Charif and S. J. McKenna: "Activity summarization and fall detection in a supportive home environment", Proc. of IEEE Int'l Conf. on Conf. Pattern Recognition, pp.323-326 (2004)

(17) C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau: "Fall detection from human shape and motion history using video surveillance", Proc. of 21st Int't Workshops on Advanced Information Networking and Applications, pp.875-880 (2007)

(18) Y. T. Liao, C. L. Huang, and S. C. Hsu: "Slip and fall event detection using Bayesian Belief Network", Pattern Recognition, Vol.45, pp.24-32 (2012)

(19) Y. T. Chen, Y. C. Lin, and W. H. Fang: "A hybrid human fall detection scheme", Proc. of IEEE Conf. on Image Processing, pp.3485-3488 (2010)

(20) R. Cucchiara, C. Grana, A. Prati, and R. Vezzani: "Probabilistic posture classification for human-behavior analysis", IEEE Trans. on Systems, Man, and Cybernetics, Vol.35, No.1, pp.42-54 (2005)

(21) R. Cucchiara, A. Prati, and R. Vezzani: "A multi-camera vision system for fall detection and alarm generation", Expert Systems, Vol.24, No.5, pp.334-345 (2007)

(22) C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau: "Monocular 3D head tracking to detect falls of elderly people", Proc. of 28th IEEE Int'l Conf. on EMBS, pp.6384-6387 (2006)

(23) G. Diraco, A. Leone, and P. Siciliano: "An active vision system for fall detection and posture recognition in elderly health care", Proc. of Design, Automation and Test in Europe Conf. and Exhibition, pp.1536-1541 (2010)

(24) N. Thome, M. Serge, and S. Ambellouis: "A real-time multi-view fall detection system: A LHMM-based approach", IEEE Trans. on Circuits and Systems for Video Technology, Vol.18, No.11, pp.1522-1532 (2008)

(25) D. Anderson, R. H. Luke, J. M. Keller, M. Skubic, M. Rantz, and M. Aud: "Linguistic summarization of video for fall detection using voxel person and fuzzy logic", Computer Vision and Image Understanding, Vol.113, pp.80-89 (2009)

(26) C. Stauffer and W. L. R. Grimson: "Learning patterns of activities using real-time tracking", IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.22, No.8, pp.747-757 (2000)

Dao Huu Hung (Non-member) was born in Hanoi, Vietnam. He received B.Sc. from Hanoi University of Technology and M.Sc. from National Taiwan University of Science and Technology, both in Electrical Engineering, in 2007 and 2010, respectively. In 2011, he was with Panasonic R&D Center Vietnam and spent one month at Panasonic Advanced Technology Development Center at Nagoya, Japan as a visiting R&D Engineer. Currently, he is working towards PhD degree at Hyper Vision Research Laboratory, Department of Information and Computer Science, Keio University at Yagami, Yokohama, Japan. He is also performing a research internship in NTT communication R&D Laboratory at Atsugi, Kanagawa, Japan from July 2012 to March 2013. His research interests include image processing, computer vision, pattern recognition, and applications to surveillance systems.

Hideo Saito (Member) received his B.E., M.E., and PhD degrees in Electrical Engineering from Keio University, Japan, in 1987, 1989, and 1992, respectively. He has been on the faculty of Department of Electrical Engineering, Keio University since 1992. From 1997 to 1999, he stayed at Robotics Institute, Carnegie Mellon University as a visiting researcher. Since 2006, he has been a Professor of Department of Information and Computer Science, Keio University. His research interests include computer vision, mixed reality, virtual reality, and 3D video analysis and synthesis. He has been an area chair of ACCV' 09, ACCV' 10, and ACCV'12, general co-chair of ICAT' 06, ICAT' 08, and general chair of MVA' 09. He is a senior member of IEEE and IEICE, Japan.