Compressed-Domain Fall Incident Detection for Intelligent Home Surveillance

Chia-Wen Lin, Zhi-Hong Ling, Yuan-Cheng Chang
Dept. Computer Science & Information Engineering
National Chung Cheng University
Chiayi 621, Taiwan
cwlin@cs.ccu.edu.tw

Chung J. Kuo
Component Business Group
Delta Electronic, Inc.
Taoyuan 333, Taiwan
chung.kuo@delta.com.tw

***Abstract*—This paper presents a compressed-domain fall incident detection scheme for intelligent home surveillance applications. For object extraction, global motion parameters are estimated to distinguish local object motions and camera motions so as to obtain a rough object mask. Then, we perform change detection and/or background subtraction on the DC+2AC images extracted from the incoming coded bitstream to refine the object mask. Subsequently, an object clustering algorithm is used to automatically extract the individual video objects iteratively. After detecting the moving objects, compressed-domain features of each object are then extracted for identifying and locating fall incident. Our experiments show that the proposed method can correctly detect fall incidents in real time.**

I. Introduction

Electronic video surveillance systems can be classified under a technological perspective into three successive generations [2]. New-generation video surveillance systems can benefit from new advances in digital video communication (video compression, bandwidth reduction, and convenient networking), digital video processing, and broadband access network infrastructures. For example, digital video compression allows efficient transmission and recording of video events. Video enhancement algorithms can be implemented to enhance the quality of video under poor illumination conditions. Video streaming and real-time video networking can provide flexible and ubiquitous video monitoring from remote locations. Automatic alarms can be generated and sent through networks or pagers to notify the users of abnormal situations. Research work on distributed real-time video processing techniques for robust video transmission, color-video processing, event-based attention focusing, model-based sequence understanding in surveillance applications is expected to provide more and more interesting features, thanks to the availability of low-cost high-performance computers, and mobile and fixed multimedia communications. In an intelligent video surveillance system, it would be very helpful to provide features of automatically detecting unusual event, such as, fall incident detection, human object detection and tracking, and fire/smoke detection.

In the case of elderly people living on their own, there is a particular need for monitoring their behavior, such as a fall, or a long period of inactivity. Falls amongst the elderly are particularly serious and may lead to injury, restricted activities, fear, or death. It is shown in [2] that 28-34% elderly people in the community experience at least one fall every year, and 40-60% of the falls lead to injury. The main reasons elders become bedridden are apoplectic ictus, decrepitude, falls, and fractures [4]. Fall-related injuries have also been among the five most common causes of death amongst the elderly population [5]. The early detections and recording of fall incidents can help the elderly to obtain in-time medical treatments as well as help identify reasons of incidents while sustaining a fall.

Most of existing fall detection schemes described in [3]-[5] propose to use specially designed sensors and circuitry. Our work is based on vision-based methods since networked video cameras have been widely deployed for surveillance and homecare applications. To reduce the computational complexity, we focus on compressed-domain schemes. The first task for vision-based fall incident detection is to detect human objects. There have been some research works for video object segmentation in the compressed domain [7][8]. For example, the method in [7] proposes an EM approach to estimate the camera parameters so as to generate the object masks. Similarly, the method in [8] also proposes to extract object by applying the EM algorithm. The above two methods both utilize the motion vectors to segment object. However, the motion vectors (MVs) are irregular and coarsely sampled, due to the use of “non-sophisticated” block matching motion estimation algorithm in generating the MV field, so the results of object segmentation may not be precise enough for the use of event detection.

In this paper, we propose a vision-based compressed-domain fall detection schemes for intelligent home surveillance applications. The proposed scheme can detect and track moving objects from the incoming compressed video in the compressed domain, without the need of decoding the incoming video into pixel values. In addition to the motion information, we propose to utilize DC+2AC image to perform change detection and/or background subtraction to refine the object mask. After detecting the moving objects, compressed-domain features of each object are then extracted for identifying and locating fall incident. The proposed system can also differentiate fall-down and squatting events.

II. PROPOSED FALL-INCIDENT DETECTION SCHEME

The block diagram of the proposed compressed-domain fall incident detection scheme is given in Fig. 1. The proposed scheme involves two steps: compressed-domain object extraction and fall-down detection. At first, the MVs and the DC+2AC image [10] of each video frame are extracted from the incoming bitstream for the subsequent processing. The MVs extracted from the incoming bitstream are fed into the Global Motion Estimation (GME) module to estimate the global motion (GM) parameters. As a result, the global motion and local object motion(s) are separated, and then those macroblocks (MBs) with significant local motions are grouped together to obtain a rough object mask.

If the video shot contains a global motion, the GM-compensated Change Detection operation is performed to refine the object mask. Otherwise, the Change Detection module is used to refine the object mask. For frames that contain more than a single object, the object clustering operation is performed to separate the object mask into multiple individual object sub-masks.

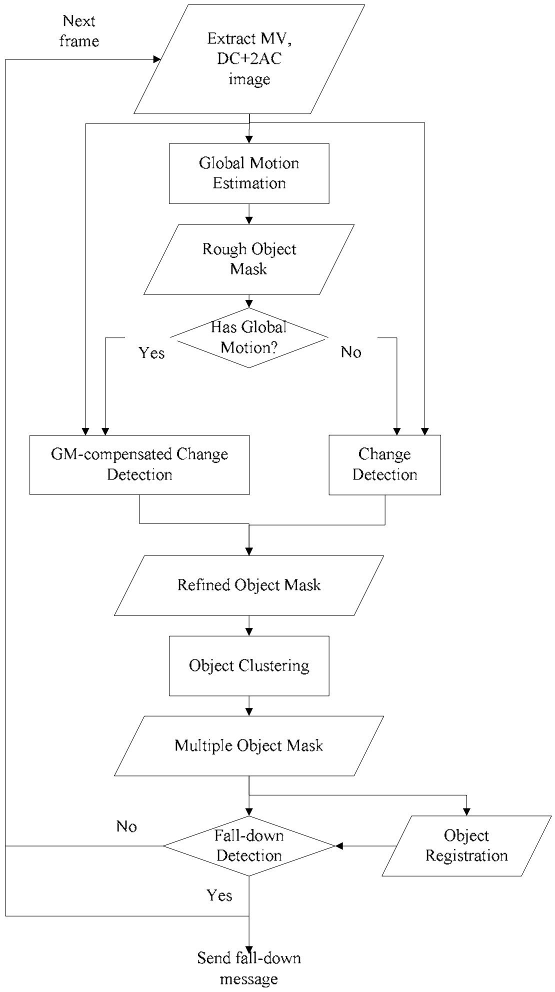


Fig. 1. Block diagram of the proposed compressed-domain fall-down detection scheme.

After extracting the video object, the fall-down detection module uses three feature parameters: the centroid of a human object, the vertical projection histogram value, and the duration of an event to identify and locate fall-down events. Object tracking is activated in our method when the video has more than one object. The Object Registration module is used to find the correspondence of video objects between two consecutive frames so as to obtain the associated feature parameters of each object.

III. COMPRESSED-DOMAIN OBJECT EXTRACTION

A. *Initial Object Segmentation*

In order to separate motion and local object motions, the global motion needs to be estimated. In this work, we modify the GME method proposed in [9] to estimate the GM parameters between two consecutive video frames in the compressed domain. In our method, the incoming MVs are first filtered using a 2-D median filter with a 3×3 mask to remove the noise due to the inaccurate motion estimation performed in video encoding. Then the global motion is obtained by minimizing the fitting error between the input MVs and the MVs generated from the estimated motion model using the Newton-Raphson method with outlier rejections [9]. The six-parameter affine model shown in (1) is adopted to estimate the GM parameters.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ a_4 & a_5 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_3 \\ a_6 \end{bmatrix} \qquad (1)$$

where (x',y') and (x,y) represent the pixel coordinates in the reference frame and the current frame, respectively; [a_1, a_2, a_3, a_4, a_5, a_6] are the GM parameters.

After the GME step, MBs with MVs significantly different from the global motion are classified as belonging to local moving objects using the following rule.

$$\mathrm{MB}^i = \begin{cases} \textit{Object} & \text{if } \left|\mathrm{MV}_x^i - \mathrm{MV}_{\mathrm{GM},x}^i\right| > TH_{\mathrm{obj}} \text{ or } \left|\mathrm{MV}_y^i - \mathrm{MV}_{\mathrm{GM},y}^i\right| > TH_{\mathrm{obj}} \\ \textit{Background} & \text{otherwise} \end{cases}$$

where MB^i is the segmentation mask of the ith MB; (MV_x^i, MV_y^i) is the incoming MV of the ith MB; TH_{obj} is a threshold, which is set empirically; ($\mathrm{MV}_{\mathrm{GM},x}^i$, $\mathrm{MV}_{\mathrm{GM},y}^i$) is the MV of the ith MB mapped from the GM parameters as calculated by (2).

$$\begin{cases} \mathrm{MV}_{\mathrm{GM},x}^i = \left(a_1 x^i + a_2 y^i + a_3\right) - x^i \\ \mathrm{MV}_{\mathrm{GM},y}^i = \left(a_4 x^i + a_5 y^i + a_6\right) - y^i \end{cases} \qquad (2)$$

where (x^i, y^i) is the displacement of the ith MB.

B. *Refinement of Object Segmentation Masks*

After the classification, we obtain a rough object mask with granularity of MB size (i.e., 16×16 pixels). The granularity of MB size, however, may be too coarse to represent the object shape with enough accuracy for the subsequent fall-down event detection. To achieve finer granularity, we propose to refine the segmentation result by using the change detection masks (CDMs) of DC+2AC images [10]. The CDM-based refinement procedure is divided into two parts: one performs change detection by taking the previous frame as the reference frame, while the other performs background subtraction that takes the background frame as the reference frame. Using the previous frame as the reference frame for change detection usually performs well when there are significant object movements. However, should there be no significant object movement, the change detection scheme may fail; instead, the background subtraction scheme can be used to cope with such situation. According to our observations, if an object has significantly movement, its corresponding object sizes in CDMs of the current frame ($SIZE^{n}_{\text{CDM}}$) and previous frame tend to be close. Otherwise $SIZE^{n}_{\text{CDM}}$ would be very small and much less than $SIZE^{n-1}_{\text{CDM}}$. The following rule is used to determine whether the video objects have moved or not.

```
if ($SIZE^{n}_{\text{CDM}} > K_{\text{SIZE}} \times SIZE^{n-1}_{\text{CDM}}$) && ($SIZE^{n}_{\text{CDM}} > TH_{\text{SIZE}}$)
    Use the CDM and background subtraction for object refinement
else
    Perform background subtraction, and use the result for refinement
```

where K_{SIZE} and TH_{SIZE} are two parameters obtained empirically. The CDM obtained from the above procedure is used to refine the object masks. The extracted background information is subsequently used to update the background frame memory for use in processing the following frames.

For a video clip containing global motion, the compensation of global motion should be performed prior to the CDM-based refinement. Otherwise, most part of non-still background may be misclassified as object movements. Before extracting the DC+2AC image, we have to compute all the DCT coefficients of the current frame using the DCT-domain motion compensation (DCT-MC) scheme [11]. After extracting the DC+2AC image from the GM-compensated DCT coefficients, The CDM obtained by using the previous frame as the reference frame is then used to refine the object segmentation masks.

The CDM-based refinement procedure is described below. First, the CDM is refined to the granularity of 8×8 block size (SEG_{CD}), while the rough object mask obtained from the GME module is also enlarged to the same granularity (SEG_{GME}). If the objects move significantly, we consider both the two results SEG_{CD} and SEG_{GME} reliable enough. Otherwise, only SEG_{CD} is considered reliable. In the case of no significant object movements, the MVs of object and background MBs are almost the same, thus SEG_{GME} may not be reliable. The following rule is used to obtain the refined object mask:

$$SEG_{\text{final}} = \begin{cases} SEG_{\text{CD}} & \text{if } \dfrac{1}{N_{\text{obj}}} \displaystyle\sum_{MB_i \in \text{object}} \left(\left|\text{MV}^{i}_{x}\right| + \left|\text{MV}^{i}_{y}\right| \right) < TH_{\text{MV}} \\ SEG_{\text{GME}} \;\&\; SEG_{\text{CD}} & \text{otherwise} \end{cases}$$

where N_{obj} is the number of object MBs; TH_{MV} is the threshold to compare with the average magnitude of MVs of object MBs.

Since a video frame may have multiple moving objects, after the above refinement procedure, an iterative object clustering algorithm is performed to automatically separate individual objects by clustering the foreground MBs with distinct local motions from the refined segmentation mask. In this algorithm, the connected component labeling is first performed on the object mask to label connected MBs. The local motion of a clustered MB group with the same label is used to determine whether the object group has more than one objects or not. Object MBs with homogeneous local motions and spatial locations are grouped as an object iteratively until all the objects are extracted.

C. Experimental Results

Three CIF (352×288) test sequences: "Coast-guard" (one object without GM), "Pamphlet" (two objects without GM) and "Hall" (one object with GM) are encoded using an MPEG-4 encoder as the inputs to evaluate the proposed compressed-domain segmentation scheme. We compare the extracted object masks with the ground-truth masks to calculate three performance indices for each frame: the number of object blocks, the number of missing blocks, and the number of false alarm blocks. The following objective metric is also used to evaluate the ratio of correctness of object segmentation:

$$d(M^{\text{ref}}_{t}, M^{\text{seg}}_{t}) = 1 - \frac{\sum_{(x,y)} M^{\text{ref}}_{t}(x,y) \oplus M^{\text{seg}}_{t}(x,y)}{\sum_{(x,y)} M^{\text{ref}}_{t}(x,y)} \quad (3)$$

where M^{ref}_{t} is the ground-truth mask of the tth frame, and M^{seg}_{t} is the extracted object masks of the tth frame. (x,y) is the index of block.

From Table I, we can observe that the segmentation results for the sequence containing no GM are better than those for the sequence with GM, and for the sequence containing a single object are better than the results of the sequence containing multiple objects. The simulations are performed on an AMD Athlon 1GHz PC. The processing speed is about 13-18 CIF fps, depending on the characteristics of sequences.

TABLE I. PERFORMANCE EVALUATION OF THE PROPOSED COMPRESSED-DOMAIN OBJECT SEGMENTATION METHOD

	Average # of object blocks	Average miss-rate	Average false-alarm rate	Ratio of correctness
Pamphlet	559	3.8%	3.4%	0.927
Hall	87	14.9%	8.0%	0.74
Coastguard	220	16.4%	8.6%	0.75

IV. VISIO-BASED FALL INCIDENT DETECTION FROM THE OBJECT MASK

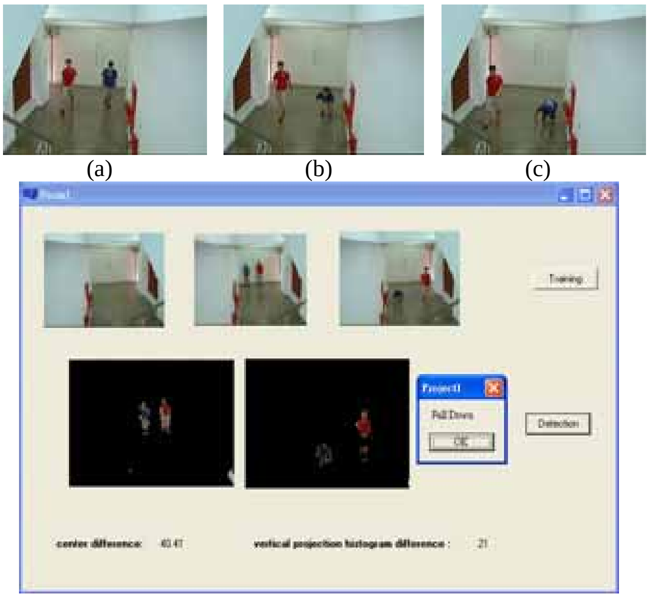
To identify and locate a fall-down event of a person, we observe that three features can be used to characterize fall-down events according to our experiments. First, a fall-down usually occurs in a short time period with a typical range of 0.4s~0.8s. Second, a person's centroid will change significantly when he/she falls down. Third, the change rate of vertical projection histogram is a useful feature for detecting a fall-down event, because a standing or falling-down human object will have different vertical projection histogram values.

In order to obtain the three feature values: the centroid of a human object, the vertical projection histogram value, and the duration of an event detected, the human objects need to be extracted first. Fig. 2 illustrates two snapshots of the object segmentation results using the proposed compressed-domain object segmentation method. After extracting the foreground object, the vertical projection histogram is computed as follows.

$$H(x,y) = \begin{cases} 1 & \text{if } (x,y) \text{ is an object pixel} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$V(x) = \sum_{y} H(x,y) \quad (5)$$

We can then find the centroid of human object. Fig. 3 compares the centroid locations and vertical projection histograms of a normal-walking person and a falling-down persons. In this example the average centroid locations before and after falling down are 137 and 184, respectively. The vertical projection histogram values before and after falling down are 98 and 73, respectively. The duration of the event is 0.56 s which is within the typical time range of fall-down event. Our system will thus issue a fall-down alarm for such situation as shown in Fig. 2. The changing rate of centroid locations can be used to differentiate the fall-down (fast change rate) and normal squatting (slow change rate) events.

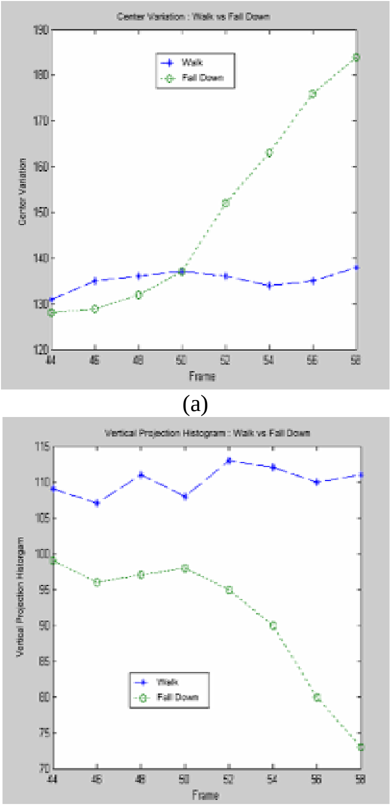

(a) (b) (c)

center difference: 40.47

vertical projection histogram difference : 21

(d)

Fig. 2. (a)-(c): three snapshots of a fall-down event with two persons that only one of them falls down; (d) detection result.



(a)

(b)

Fig. 3. (a) the position of object center versus time; (b) the vertical projection histogram of object versus time.

V. Conclusion

We have presented a vision-based compressed-domain fall-down detection scheme for intelligent surveillance applications. The proposed scheme involves two steps: compressed-domain object extraction and fall-down detection. In the object extraction step, the MVs and the DC+2AC image of each frame are firstly extracted. GME is then performed to distinguish moving object MBs from background MBs to obtain a rough object segmentation mask. The CDM is then used to refine the object mask. Should the video shot contain GMs, the GM compensation is performed prior to the change detection operation. Finally, object clustering is performed to separate the object mask into multiple individual objects. In the second step, three feature values: the centroid of a human object, the vertical projection histogram value, and the duration of an event detected are used to identify and locate fall-down events. The proposed object segmentation method can extract moving objects with or without cameral motions, thereby being useful for video surveillance applications equipped with still or pan-tilt cameras. The experimental results show that the proposed method can correctly detect fall-down events in real-time.

References

[1] C. Regazzoni, V. Ramesh, and G. L. Foresti, “Scanning the issue/technology,” *Proc. IEEE*, vol. 89, no. 10, pp. 1355-1367, Oct. 2001.

[2] J. Teno, D. Kiel, and V. Mor, “Multiple strumbles: a risk factor for falls in community-dwelling elderly,” *J. America Geriatrics Society*, vol. 38, no. 12, pp. 1321-1325, 1990.

[3] N. Noury *et al*., “Monitoring behavior in home using a smart fall sensor and position sensors,” in *Proc. IEEE Int. Conf. EMBS*, pp. 607-610, Oct. 2000, Lyon, France.

[4] T. Tamura *et al*., “An ambulatory fall monitor for the elderly,” in *Proc. IEEE Int. Conf. EMBS*, pp. 2608-2610, July 2000., Chicago, IL.

[5] G. Williams *et al*., “A smart fall & activity monitor for telecare applications,” in *Proc. IEEE Int. Conf. EMBS*, vol. 20, no. 3, pp. 1151-1154, 1998.

[6] K. Yoon, D.F. Dementhon, and D. Doermann, “Event detection from MPEG video in the compressed domain,” in *Proc. IEEE ICPR*, Barcelona, Spain, 2000.

[7] R. Wang, H.-J. Zhang and Y.-Q. Zhang, “A confidence measure based moving object extraction system built for compressed domain,” in *Proc. IEEE ISCAS*, vol. 5, pp.21 -24, May 2000, Geneva, Switzerland.

[8] R. V. Babu, K. R. Ramakrishnan , “Compressed domain motion segmentation for video object extraction,” in *Proc. IEEE ICASSP*, vol. 4 , pp. 3788 –3791, 2002.

[9] Y. Su, M.-T. Sun, and V. Hsu, “Global motion estimation from coarsely sampled motion vector field and the applications,” in *Proc. IEEE ISCAS*, vol. 2, pp. 628-631, Mar. 2003, Bangkok, Thailand.

[10] B. L. Yeo, *Efficient Processing of Compressed Images and Video*, Ph.D. thesis, Princeton University, Jan.1996.

[11] S. F. Chang and D. G. Messerschmitt, “Manipulation and compositing of MC-DCT compressed video,” *IEEE J. Select. Areas Commun.*, pp. 1-11, 1995.