Human fall detection in surveillance videos using fall motion vector modeling

Chalavadi Vishnu, Rajeshreddy Datla, Debaditya Roy,
Sobhan Babu, and C Krishna Mohan, *Senior Member, IEEE*

***Abstract*—Representation of spatio-temporal properties of human body silhouette and human-to-ground relationship, significantly contribute to the fall detection process. So, we propose an approach to efficiently model the spatio-temporal features using fall motion vector. First, we construct a Gaussian mixture model (GMM) called fall motion mixture model (FMMM) using histogram of optical flow and motion boundary histogram features to implicitly capture motion attributes in both the fall and non-fall videos. The FMMM contains both fall and non-fall attributes resulting in a high-dimensional representation. In order to extract only the relevant attributes for a particular fall or non-fall videos, we perform factor analysis on FMMM to get a low dimensional representation known as fall motion vector. Using fall motion vector, we are able to efficiently identify fall events in varieties of scenarios, such as the narrow angle camera (Le2i dataset), wide angle camera (URFall dataset), and multiple cameras (Montreal dataset). In all these scenarios, we show that the proposed fall motion vector achieves better performance than the existing methods.**

Keywords - Human fall detection, surveillance videos, Gaussian mixture model, fall motion vector, and factor analysis.

I. Introduction

Human fall is an abnormal activity that occurs due to an abrupt loss of balance by being startled that causes slipping. These falls may cause long term disabilities, and even death due to absence or delay of assistance. As the world population by the year 2050 would consist of 20% of the elderly (over 65 yrs age) people [1], an automatic human fall detection system is required for effective monitoring. The detection systems for human fall is mainly categorized into two types, namely, wearable sensor-based systems and computer vision-based systems. Also, various sensor [2–4] and vision-based methods for human activity recognition are presented in [5]. Wearable sensor-based systems employ different multiple-sensors like the heartbeat [6, 7], gyroscope [8], a comprehensive data acquisition system [9], etc., connected to the body of each person, who is prone to fall. The measurements used in the wearable sensor-based systems are computationally intensive, e.g., frontal area calculation and skeletal joint expectation, to determine an abnormal human activity. Also, these sensor-based systems impose an individual person to wear the sensors. Usually, people may forget or sometimes do not feel comfortable wearing such sensors. In such scenarios, video surveillance systems can be used to continuously monitor persons passively. When an abnormal activity (like the fall event) occurs, computer-vision based systems send an alarm to the concerned care taker for an immediate assistance that helps in advancing the well being of the person. Surveillance videos have become a prominent medium, which avoids the constraint of wearable devices, in monitoring the activities of the people, who are prone to fall.

Surveillance video cameras provide vital information useful for observing the behavior of people during the fall event. These videos are recorded using either single or multiple cameras. Single cameras do not capture all the directions of human fall, such as anterior, posterior, left, and right falls. The view-point changes in camera requires a new training approach [10] to help in searching the candidate regions. Using multiple cameras, the direction of fall can be captured with different fields of view. These videos provide an insight into the changes in human body silhouette, head pose, and the human-to-ground relationship compared to wearable sensors. Visual cues from these video sequences are useful in establishing a human fall detection system. Typically, a human fall includes a sequence of events, such as movement history, abnormal forward, backward, or side-way movements, and the human-to-ground relationship. However, multiple perspectives, shadows, and non-uniform illumination in the surveillance videos pose various challenges in determining the co-occurrences of the events in spatio-temporal domain. Also, deceptive events such as unexpected sitting, picking up an object, by bending down make the detection process further complex.

In literature, fall detection methods in the surveillance videos focus on the extraction of shape and geometry of the person to recognize their irregular movements. However, the performance of these methods are influenced by the shadow of a person and view-point. Deep learning methods using convolutional neural networks (CNN) and long-short time memory (LSTM) networks learn spatio-temporal features automatically from the large amount of data. However, it is difficult to obtain the annotated videos of the fall events, as these are rare events. To circumvent these problems, a 3D-CNN based human fall detection [11] is proposed where kinematic data in the training process is employed to extract the features automatically. Additionally, spatio-temporal information is incorporated using LSTM. In [12], a human fall detection approach is explored that uses information from multiple ultra-wideband (UWB) radars without the need for identification of the person.

In this paper, we propose an approach for fall motion vector

C. Vishnu, Rajeshreddy Datla, Sobhan Babu, and C Krishna Mohan are with Visual Learning and Intelligence Group (VIGIL), Department of Computer Science and Engineering,
Indian Institute of Technology Hyderabad, Kandi, Sangareddy-502285, India e-mail: cs16m18p000001@iith.ac.in

Debaditya Roy is with Agency for Science, Technology and Research (A*STAR), Singapore.

modelling to detect human fall in surveillance videos. We employ both histogram of optical flow (HOF) and motion boundary histogram (MBH) to describe the spatio-temporal characteristics, termed fall motion attributes, in the surveillance videos. These spatio-temporal features are used in the training of a single Gaussian mixture model (GMM) to obtain a fall motion mixture model (FMMM). This FMMM encompasses a large number of mixtures for modelling the fall motion attributes, which are captured implicitly by estimating probability density function (pdf) that generates a particular fall event. We obtain a high-dimensional vector by concatenating the mean adaptation of the mixtures of FMMM, which contains the redundant attributes from all the videos. Hence, we perform factor analysis on this high-dimensional vector to obtain a low-dimensional fall motion vector (FMV), which retains only the important attributes relevant for the detection of the fall event. The main contributions of the proposed work are summarized as: (i) A fall motion mixture model (FMMM) is constructed to learn the fall motion attributes implicitly from the surveillance videos. (ii) An efficient low-dimensional representation called fall motion vector is obtained for fall and non-fall activities. (iii) The efficacy of the proposed method is demonstrated on the surveillance videos consisting of the narrow angle camera (Le2i dataset), wide angle camera (URFall dataset), and multiple cameras (Montreal dataset). The rest of this paper is organized as follows. Section 2 presents the literature of fall detection methods in videos. The proposed approach for the detection of human fall events is described in Section 3. In Section 4, we discuss experimental results of the proposed method. Finally, conclusion and future directions are presented in Section 5.

II. Related work

This section presents the existing computer-vision based methods for fall event detection in the surveillance videos. The human fall detection process becomes difficult in surveillance videos, mainly due to the deceptive events, such as picking up an object from the ground, sitting due to tiredness, tying shoes, etc. The existing methods for fall detection primarily focus on separating the foreground from its background in each frame of the video to obtain cues. These cues help the detection process in determining the type of an event. A set of algorithms for background-subtraction are analyzed in [13], to obtain spatio-temporal information in determining a fall event from a video sequence. This analysis is helpful in finding the background-subtractor algorithm and optimized their parameters using genetic algorithm. This improves the performance in detecting the human falls, especially in the night-time environment. Motion history image (MHI) is used in [14] to assess the fall behaviours by separating the foreground from background. Further, the fall activity is determined with the help of acceleration and angular acceleration fall features.

In addition to foreground and background separation, the perception of a quick change in the posture of a human is also examined in distinguishing the fall events from other events. A Gaussian mixture model (GMM) based adaptive background subtraction method is used for object detection. And, a set of features such as aspect ratio, horizontal and vertical gradient values of an object, and fall angle are used to describe a fall model. A two-state finite state machine (FSM) was implemented to continuously monitor human activity. However, this method detects the fall activity of a single person. The fall detection method in [15] computes the measurements such as distance and the angle between the lines joining three key points, in representing the human posture. Subsequently, a fall event is classified by the change of posture state. Stone and Skubic proposed a two-stage system [16] for fall detection. In the first stage, the vertical state of a segmented object is characterized and their time-series data is used to identify on-ground events. The second stage determines the confidence of a fall event by combining the decision trees and features extracted from the on-ground event.

The shape features used in the action detection methods have also been extended for fall event detection. The features describing the shape variation and motion history of a person are explored in [17, 18] for video based fall detection. In [17], the shape variation is quantified as timed motion history image (tMHI) by approximating the person with an ellipse using moments and orientation of the ellipse. Eventually, the standard deviation of difference between maximum values of histograms of the horizontal and vertical projection are used to identify the fall activity. In [18], the integrated spatio-temporal energy (ISTE) map is used to measure the intensity of human motion. The causality of the post and pre-events of the slip-only and fall events are modelled using Bayesian Belief Network (BBN).

Some methods for fall detection also explored the features derived from neural networks and machine learning approaches. In [19], feature learning methods are applied over the training samples constructed using ViBe [20], which extracts humans in a specific resolution. A feature vector for human fall detection is described in [21], which combines histograms of oriented gradients (HOG), local binary pattern (LBP) and deep features of the video frames. A classifier based on K-Nearest Neighbor is used in [22], over the features such as orientation angle, ratio of fitted ellipse, motion coefficient, and silhouette threshold to detect falls. In [23], the method uses Gaussian mixture model and principle component analysis (PCA) to identify the fall events. Also, they mentioned the sensitivity of the technique, which cause false detection due to the change in aspect ratio and angle of major axis in every frame. Thus, a consecutive-frame voting is introduced in [24] to improve the fall detection accuracy.

In the fall detection process, the perception of a quick change in the posture of a human in timely manner helps in the detection of the fall event accurately. A monitoring scheme using a multivariate exponentially weighted moving average (MEWMA) [25] captures even small changes, which is used to detect falls effectively. Further, SVM based classification is applied over detected sequences to differentiate gestures that resembles fall activity. This methodology was validated on the University of Rzeszow fall detection dataset (URFD) and the fall detection dataset (FDD). A stereo-vision based method [26] for human fall detection estimates the human pose in 2D based on deep learning approach. In addition,

both human key points and ground plane in 3D are achieved using depth information. These measures are used to outline the reasons for concluding a human fallen activity. In [27], the method provides a human readable summarization of activity & the detection of human unusual inactivity. A maximum a posteriori estimation of Gaussian mixture model is used to capture inactivity zones and entry zones with their spatial context. The unusual inactivity detection helped as cue to demonstrate the fall detection.

The ability of 3D-CNNs is exploited in [28] to obtain an effective representation from the videos by modelling both the appearance and motion simultaneously. A deep auto encoder network [29] with the combination of 3D-CNN and Convolution Gated Recurrent Units (GRU) is employed to learn both local and global spatial features in the spatio-temporal dimension. Recently, a fall detection method based on LSTM [30] has been proposed and evaluated on publicly available fall detection datasets and a complex-scene fall event dataset, introduced in [30]. This method initially uses YOLO v3 model to detect objects and Deep-Sort tracking algorithm to track the detected multi-objects, especially during occlusion. Next, an attention guided LSTM model is employed to obtain spatio-temporal features, which helps to distinguish three disturbing events from three kinds of falls, namely, forward, backward, and sideways directions.

However, the existing methods for fall detection in surveillance videos involve combination of various tasks such as foreground and background separation, object detection, and tracking. In contrast, our proposed method models fall events by training a fall motion mixture model (FMMM) using histogram of optical flow (HOF) & motion boundary histogram (MBH) features, to capture the fall motion attributes implicitly from all videos. These attributes include spatio-temporal properties of the human body silhouette, abnormal forward & backward, and side-way movements. A low-dimensional fall motion vector is obtained to provide efficient representation of fall and non-fall events.

III. Proposed Work

Figure 1 presents the block diagram of the proposed fall detection method with various stages, such as feature extraction, fall motion mixture model construction, and fall motion vector extraction. These stages are carried out in sequence to obtain an efficient representation for fall event detection, which are explained in detail in the following subsections.

A. Feature extraction

A fall event is composed of a sequence of spatio-temporal properties corresponding to the human body silhouette. We employ both histogram of optical flow (HOF) & motion boundary histogram (MBH) descriptors from surveillance videos. The feature points that are densely sampled at multiple spatial scales are tracked across consecutive frames by using dense optical flow. These feature points of consecutive frames are concatenated to form a trajectory. Usually, trajectories drift away from its initial locations while tracking, so the length of the trajectory is set to L_n = 15 frames [31–33].

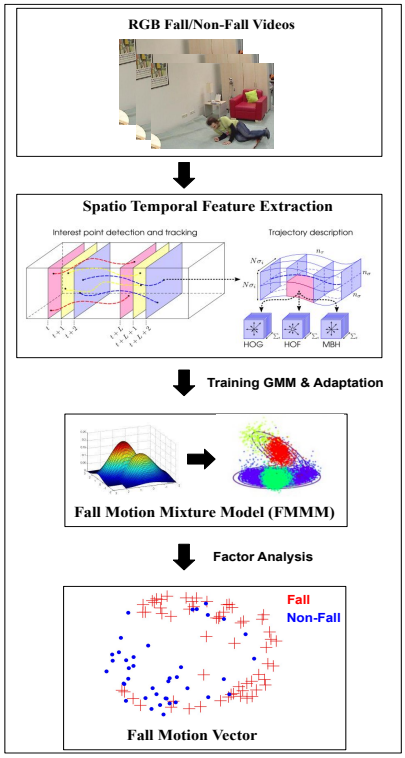

Fig. 1: The block diagram of fall motion vector modeling.

In order to extract motion information of the fall event, local descriptors like histogram of optical flow (HOF) & motion boundary histogram (MBH) are computed around the trajectory within a spatio-temporal volume (of size $N \times N$ pixels and L_n frames). The volume is subdivided into grids of size $n_h \times n_w \times n_t$, where n_h, n_w, and n_t are height, width, and temporal segment lengths (typically $n_h = n_w = 2$, and $n_t = 3$), respectively, to attain the structure information. The orientations of HOF descriptor are quantized into 9 bins resulting in a dimension of 108 ($2 \times 2 \times 3 \times 9$). Further, MBH helps to remove the background camera motion induced by the optical flow with the computation of its spatial derivatives across x & y directions. The orientation of obtained derivatives is quantized into a histogram of 8 bins i.e., MBHx of 96 ($2 \times 2 \times 3 \times 8$) & MBHy of 96 ($2 \times 2 \times 3 \times 8$) dimension. Parameters considered for the space-time volume size i.e., $2 \times 2 \times 3$ are found to be optimal after cross-validating on the training set of our datasets. And it is observed that further increase in the number of cells beyond $2 \times 2 \times 3$ does not improve the performance similar to [32]. The obtained HOF and MBH descriptors are used in the fall motion vector modelling to analyze the dynamics of fall and non-fall events.

B. Fall motion mixture Model (FMMM)

Each video is considered as a random process, assumed to be a Gaussian probability distribution function (*pdf*). To compute the *pdf* of these random processes, the parameters are estimated by training a Gaussian mixture model (GMM) [33] for fall and non-fall videos, where the number of mixtures 32, 64, 128, 256, and 512 are chosen empirically for construction

of fall motion mixture model (FMMM). This FMMM is a GMM which captures the fall motion attributes represented as

$$p(\mathbf{x}_l) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x}_l|\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k), \tag{1}$$

where weights of the mixture w_k satisfy the constraint $\sum_{k=1}^{K} w_k = 1$. The mean and covariance for the mixture k of the FMMM are given by $\boldsymbol{\mu_k}$ and $\boldsymbol{\sigma_k}$, respectively. A feature $\mathbf{x}_l$ is the component of a video $\mathbf{x}$ expressed as a set of feature representations $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L\}$. Then, we train a separate FMMM for HOF, MBH, and concatenated features of both HOF & MBH descriptors using expectation maximization (EM) algorithm. In order to estimate the *pdf* of a clip, a maximum a posteriori (MAP) adaptation is performed using the features of the fall and non-fall video clips [33] to improve the contribution of the features existing in that video clip.

C. Fall motion vector (FMV)

The posterior probability $p(k|\mathbf{x}_l)$ of the mixture k and the feature vector $\mathbf{x}_l$ (of size $d \times 1$) for a video clip $\mathbf{x}$ is given by

$$p(k|\mathbf{x}_l) = \frac{w_k p(\mathbf{x}_l|k)}{\sum_{k=1}^{K} w_k p(\mathbf{x}_l|k)}, \tag{2}$$

where $p(\mathbf{x}_l|k)$ is the likelihood of $\mathbf{x}_l$ coming from a mixture k. Using the posterior $p(k|\mathbf{x}_l)$, the Baum-Welch statistics of each video clip $\mathbf{x}$ are computed by

$$n_k(\mathbf{x}) = \sum_{l=1}^{L} p(k|\mathbf{x}_l), \tag{3a}$$

and

$$\mathbf{H}_k(\mathbf{x}) = \frac{1}{n_k(\mathbf{x})} \sum_{l=1}^{L} p(k|\mathbf{x}_l)\mathbf{x}_l. \tag{3b}$$

We adapt the weights and means of every mixture in the fall motion mixture Model (FMMM) for a particular video clip $\mathbf{x}$ as

$$\hat{w}_k = \alpha n_k(\mathbf{x})/L + (1-\alpha)w_k \tag{4a}$$

and

$$\hat{\boldsymbol{\mu}}_k = \alpha \mathbf{H_k}(\mathbf{x}) + (1-\alpha)\boldsymbol{\mu}, \tag{4b}$$

respectively. We concatenate the adapted means of K mixtures to form a feature vector of dimension $Kd \times 1$. The attained feature vector is called a fall motion vector (FMV) represented by $\mathbf{f}(\mathbf{x}) = [\hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_2 \cdots \hat{\boldsymbol{\mu}}_K]^t$. This fall motion vector is of high-dimension and contains redundant features that do not contribute to particular event. So, we obtain the low-dimensional representation of FMV using factor analysis on the high-dimensional fall motion vector $\mathbf{f}$, which is decomposed as

$$\mathbf{f} = \mathbf{v} + \mathbf{Tq}, \tag{5}$$

where $\mathbf{v}$ represents the mean of the fall motion mixture model (FMMM), $\mathbf{T}$ represents a rectangular variability matrix of $Kd \times r$ dimension, & $\mathbf{q}$ is a vector of r-dimension, assuming the Gaussian distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [33]. Here, *fall motion vector (FMV)* is determined by posterior probability $P(\mathbf{q}|\mathbf{x})$ after noticing the video $\mathbf{x}$ indicated as $P(\mathbf{q}|\mathbf{x}) \propto P(\mathbf{x}|\mathbf{q})\mathcal{N}(\mathbf{0}, \mathbf{I})$

$$\begin{aligned} &\propto \exp\left(\mathbf{q^t T^t \Sigma^{-1}} \tilde{\boldsymbol{f}}(\mathbf{x}) - \frac{1}{2}\mathbf{q}\mathbf{T}^t\mathbf{D}(\mathbf{x})\boldsymbol{\Sigma}^{-1}\mathbf{T}\mathbf{q} - \frac{1}{2}\mathbf{q}^t\mathbf{q}\right), \\ &= \exp\left(\mathbf{q}\mathbf{T^t}\boldsymbol{\Sigma}^{\mathbf{-1}}\tilde{\boldsymbol{f}}(\mathbf{x}) - \frac{1}{2}\mathbf{q}^t\mathbf{M}(\mathbf{x})\mathbf{q}\right), \\ &= \exp\left(-\frac{1}{2}(\mathbf{q}-\mathbf{N}(\mathbf{x}))^t\mathbf{M}(\mathbf{x})(\mathbf{q}-\mathbf{N}(\mathbf{x}))\right) \times constant. \end{aligned} \tag{6}$$

where $\boldsymbol{\Sigma}$ represents the diagonal covariance of $Kd \times Kd$ dimension. Also, the matrix $\mathbf{N}(\mathbf{x}) = \mathbf{M}^{-1}(\mathbf{x})\mathbf{T}^t\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}}(\mathbf{x})$. Here, $\tilde{\boldsymbol{f}}(\mathbf{x})$ is centered fall motion vector due to the posterior probability of $\mathbf{q}$ conditioned on the statistics of Baum-Welch algorithm of the video and centered over means of FMMM. The first order statistics of Baum-Welch algorithm of FMMM can be computed using $\tilde{\mathbf{H}}_k(\mathbf{x}) = \sum_{l=1}^{L} p(k|\mathbf{x}_l)(\mathbf{x}_l - \boldsymbol{\mu}_k)$. Here, $\tilde{\boldsymbol{f}}(\mathbf{x})$ is given by concatenating the obtained first order statistics $\tilde{\boldsymbol{f}}(\mathbf{x}) = [\tilde{\mathbf{H}}_1(\mathbf{x})\tilde{\mathbf{H}}_2(\mathbf{x})\cdots\tilde{\mathbf{H}}_K(\mathbf{x})]^t$. The matrix $(\mathbf{M}(\mathbf{x}))$ can be written as $\mathbf{M}(\mathbf{x}) = \mathbf{I} + \mathbf{T}^t\boldsymbol{\Sigma}^{-1}\mathbf{D}(\mathbf{x})\mathbf{T}$. Here, $\mathbf{D}(\mathbf{x})$ is $Kd \times Kd$ diagonal matrix with diagonal blocks of $n_k(\mathbf{x})\mathbf{I}$, for $k = 1, ..., K$ & $\mathbf{I}$ is $d \times d$ identity matrix. Following are the mean & covariance matrix of the posterior probability:

$$E[\mathbf{q}(\mathbf{x})] = \mathbf{M}^{-1}(\mathbf{x})\mathbf{T}^t\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{f}}(\mathbf{x}) \tag{7a}$$

$$Cov(\mathbf{q}(\mathbf{x}), \mathbf{q}(\mathbf{x})) = \mathbf{M}^{-1}(\mathbf{x}). \tag{7b}$$

In Expectation Maximization algorithm, the E step estimates the posterior mean & covariance iteratively and in M step, $\mathbf{T}$ & $\boldsymbol{\Sigma}$ are updated using the same statistics. The mean and covariance of fall motion mixture model (FMMM) are defined as $\mathbf{v}$ and $\boldsymbol{\Sigma}$. The initial matrix is computed by taking the matrix $\mathbf{T}$ and a suitable rank r. Then Equations 7a & 7b are used to compute $E[\mathbf{q}(\mathbf{x})]$ and $Cov(\mathbf{q}(\mathbf{x}), \mathbf{q}(\mathbf{x}))$. The matrix $\mathbf{T}$ is computed in M-step as the solution of

$$\sum_{\mathbf{x}} \mathbf{D}(\mathbf{x})\mathbf{T}E[\mathbf{q}(\mathbf{x})\mathbf{q}^t(\mathbf{x})] = \sum_{\mathbf{x}} \tilde{\boldsymbol{f}}(\mathbf{x})E[\mathbf{q}^t(\mathbf{x})], \tag{8}$$

resulting in r linear equations. Here $\tilde{\boldsymbol{f}}(\mathbf{x})$ accounts for the total number of features in the video. Since $\mathbf{T}$ is same for all the videos, the left hand side is weighed by $\mathbf{D}(\mathbf{x})$ which accounts for the number of features in the video.

For every mixture $k = 1, 2, \cdots, K$, the covariance matrix $\boldsymbol{\Sigma}$ is estimated as

$$\boldsymbol{\Sigma}_k = \frac{1}{n_k(\mathbf{x})}\left(\sum_{\mathbf{x}} \tilde{J}_k(\mathbf{x}) - \mathbf{M}_k\right), \tag{9}$$

where $\mathbf{M}_k$ represents the k^{th} diagonal block of the $Kd \times Kd$ matrix $\frac{1}{2}\sum_{\mathbf{x}} \tilde{\boldsymbol{f}}(\mathbf{x})E[\mathbf{q}^t(\mathbf{x})]\mathbf{T}^t + \mathbf{T}E[\mathbf{q}(\mathbf{x})]\tilde{\boldsymbol{f}}^t(\mathbf{x})$. The second-order Baum-Welch statistics of the video $\tilde{\boldsymbol{J}}_k(\mathbf{x})$ is computed as

$$\tilde{\boldsymbol{J}}_k(\mathbf{x}) = diag\left(\sum_{l=1}^{L} p(k|\mathbf{x}_l)(\mathbf{x}_l - \boldsymbol{\mu}_k)(\mathbf{x}_l - \boldsymbol{\mu}_k)^t\right). \tag{10}$$

The fall motion vector (FMV) for each fall and non-fall video is estimated after the last iteration of M-step using

$$\mathbf{q}(\mathbf{x}) = (\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}\mathbf{D}(\mathbf{x})\mathbf{T})^{-1}\mathbf{T}^t\mathbf{\Sigma}^{-1}\tilde{\boldsymbol{f}}(\mathbf{x}). \quad (11)$$

This process of decomposing the high-dimensional fall motion vector to low-dimensional is called factor analysis. The $\mathbf{T}$-matrix consists of the eigen vectors of the largest r eigenvalues of the covariance matrix. It is hypothesized that these large eigenvalues come from the Gaussian mixtures, that model the motion profile in the video. Now, the fall motion vector (FMV) can be projected onto a r-dimensional motion profile vector using $\mathbf{T}$. Subsequently, a polynomial support vector machine is employed over fall motion vectors to categorize the fall/non-fall videos. The computational complexity of the MAP estimation [34] for fall motion vector ($\mathbf{q}$) given in Eq. 7a is $\mathcal{O}(Kdr + Kr^2 + r^3)$, where K is the number of GMM mixtures in FMMM, d & r represent the dimension of feature vector and fall vector, respectively.

IV. Experimental results

In this section, we analyze the performance of the proposed method on the three varieties of benchmark fall detection datasets, namely, Le2i [35], URFall [36], and Montreal [37].

A. Datasets used

The video datasets are composed of several simulated normal daily activities and fall events involving one or more persons recorded from single & multiple cameras. Normal daily activities in these videos include human walk in various directions, housekeeping activities, crouching down, sitting down, standing up, etc. Whereas, videos of simulated falls consist of forward, backward, left, and right directions along with unexpected sitting, loss in balance, etc.

Le2i dataset- Le2i [35] is a standard RGB based fall detection dataset consisting of 130 fall and non-fall scenes recorded at 25 frames per second by 17 actors using a single camera. The average duration of all videos is about 1 minute, where the person either falls or continues performing a daily activity. The videos are recorded in different locations such as "Home", "Coffee room", "Office", and "Lecture room" with high variance between fall and non-fall events. This helps in accurately simulating realistic video sequences that can be found in home environments.

URFall dataset- URFall [36] is another conventional fall detection dataset containing 100 RGB videos, which consists of 60 fall events and 40 daily living activities. These fall videos are recorded with two Kinect cameras and its accelerometric data. Activities of daily life videos are recorded with only one camera and an accelerometer.

Montreal dataset- Montreal dataset [38] contains 24 scenarios recorded with 8 video cameras. The first 22 scenarios contain both fall and non-fall events. The last 2 scenarios contain only non-fall events.

In all the above three datasets, videos are recorded at 25 frames per second with the resolution of 320 X 240 pixels on an average. Few sample frames of these datasets are shown in Figures 2, 3, and 4. Both the normal daily activities and the simulated falls in these three datasets are segregated into non-fall and fall events, respectively. We split each dataset into 70%-30% training and testing ratio in the experimental setup.

TABLE I: Classification performance (in %) of the proposed method on Le2i, URFall, and Montreal datasets.

	Le2i		URFall		Montreal	
	3D-CNN	**HOF+MBH**	**3D-CNN**	**HOF+MBH**	**3D-CNN**	**HOF+MBH**
32	**69.57**	**78.50**	71.00	32.00	98.21	99.13
64	68.14	63.84	70.00	47.00	**99.30**	**99.82**
128	62.59	47.69	**74.00**	78.00	98.64	98.70
256	58.23	53.84	71.00	**80.00**	96.85	97.19

B. Analysis of fall motion vector on three datasets

We have trained 6 fall motion mixture models, 2 each on Le2i, URFall, and Montreal datasets using histogram of optical flow (HOF) with motion boundary histogram (MBH) features and 3D convolutional neural networks (3D-CNN) features separately. In our experiments, we use the features of conv5 layer from the pre-trained ResNet-101 backbone with 3D-CNN architecture. We have considered 3D-CNN and HOF + MBH features separately to train the fall motion mixture model (FMMM) due to their state-of-the-art performance in action recognition tasks [33, 39]. The obtained fall motion vector is passed to polynomial support vector machine (SVM) [40] for the classification of fall and non-fall videos.

Le2i dataset: Table I presents the classification performance of fall and non-fall videos using the proposed fall motion vector modeling on Le2i dataset. It can be noted that the proposed method gives better performance using histogram of optical flow (HOF) with motion boundary histogram (MBH) features compared to 3D-CNN features. The fall motion vector of dimension 200 with 32 mixture components is able to capture the fall motion attributes effectively. The visualization of fall motion attributes of fall and non-fall videos is shown in Figure 5. It can be observed from the figure that the proposed method clearly distinguishes between fall and non-fall videos. Also, there are some misclassification of fall events because the narrow angle view creates confusion between fall and non-fall events.

URFall dataset: The classification performance of the proposed method on URFall dataset is presented in Table I. It can be observed that the proposed method gives better performance using histogram of optical flow (HOF) with motion boundary histogram (MBH) features compared to 3D-CNN features. The fall motion vector of dimension 200 with 256 mixture components is able to capture the attributes of fall/non-fall videos. Figure 6 depicts the visualization of fall motion attributes on URFall dataset in the t-SNE plot. As can be seen from the figure that there is a overlap of fall motion attributes causing confusion between fall and non-fall events.

Montreal dataset: The classification performance of the proposed method on Montreal dataset is given in Table I. Videos in this dataset include redundant scene information from multiple cameras. It is to be noted that the proposed method with HOF and MBH features exhibits slightly better classification performance compared with 3D-CNN features. The fall motion vector (FMV) of size 200 with 64 mixture components using HOF and MBH features is able to retain only the significant attributes of fall/non-fall videos. The

Fig. 2: Le2i dataset

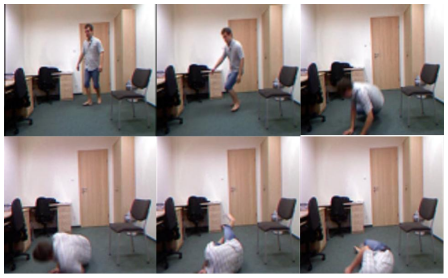

Fig. 3: URFall dataset

Fig. 4: Montreal dataset

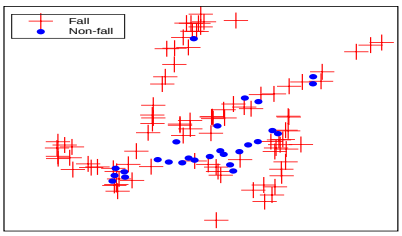

Fig. 5: 3D t-SNE plot for Le2i dataset.

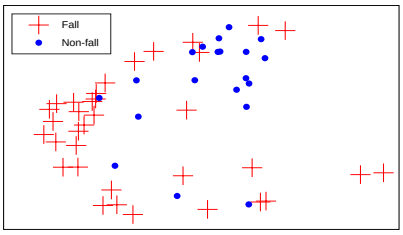

Fig. 6: 3D t-SNE plot for URFall dataset.

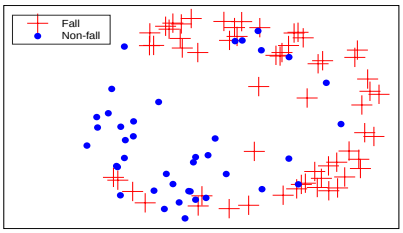

Fig. 7: 3D t-SNE plot for Montreal dataset.

visualization of fall motion attributes on the Montreal dataset is shown in Figure 7. It can be observed that there is a confusion between some fall and non-fall events resulting into misclassification. The reason for these misclassifications may be due to multiple camera views that will cause confusion of human fall attributes.

Since our objective is to get best possible classification performance, we chose the fall motion mixture model (FMMM) trained on HOF+MBH even though it has more mixture components than the 32 mixture FMMM using 3D-CNN features.

C. Performance Comparison with Existing Approaches

In this section, we present the performance comparison of our proposed approach for human fall detection with existing methods using metrics such as recall, precision, specificity, and F1-score. The number of correctly predicted fall events are quantified in terms of precision. Recall is the ability to classify the correct fall as a fall, and the specificity is the ability to classify the non-fall correctly as non-fall. The trade-off between precision and recall is determined using F1-score.

Le2i dataset - Existing computer vision-based methods explored various properties of human movements in Le2i dataset in order to determine the human fall events. Charfi et al. considered a set of properties of the human bounding box for classifying the fall and non-fall videos using support vector machine (SVM). These properties include height, width, orientation, the trajectory of the bounding box in order to track the human body silhouette. Chamle et al. employed gradient boosting classifier [41] to determine the fall event based on the features, such as fall angle, aspect ratio, and silhouette height. Poonsri et al. [24] determined the events by extracting the aspect ratio, orientation, and area ratio from the human silhouette using the principal component analysis (PCA). Due to the single camera recording, the axis aligned human movements in Le2i dataset are not captured properly using existing methods. Whereas, our method learns the fall motion attributes implicitly in order to capture the changes in human body silhouette. Our proposed method achieves better performance on Le2i dataset as compared to the existing three vision-based approaches given in Table II.

TABLE II: Comparison of the proposed method with the existing methods on Le2i dataset.

Method	Average Precision	Average Recall	F1 score
Charfi et al. [37]	0.990	0.980	0.985
Chamle et al. [41]	0.794	0.843	0.818
Poonsri et al. [24]	0.891	0.931	0.911
Ours (3D-CNN)	0.815	0.930	0.868
Ours (HOF+MBH)	**0.995**	**0.989**	**0.991**

URFall dataset - Smriti et al. [42] used optical flow and Harris corner detector to obtain the interest points from the fall and non-fall videos of URFall dataset. These interest points are passed to SVM in order to classify the fall/non-fall videos. Feng et al. [30] used the combination of spatial and temporal features in order to detect the human and also used the Deep-Sort algorithm for subsequent tracking during occlusions. Due to wide-angle camera, the vignetting effect would influence the spatio-temporal features in the video. Table III shows the performance comparison of the proposed method of two vision-based approaches [30, 42] on URFall dataset.

Montreal dataset - Shengke et al. [19] extracts the histogram of oriented gradients (HOG) features from each video of Montreal dataset. The quantized HOG features based on PCA-Net are used to classify the events into fall or non-fall with support vector machine (SVM). Kun et al. [21]

TABLE III: Comparison of the proposed method with the existing methods on URFall dataset.

Method	Average Precision	Average Recall	F1 score
Smriti et al. [43]	0.935	0.966	0.950
Feng et al. [30]	0.948	0.914	0.931
Nabil et al. [44]	-	-	0.960
Ours (3D-CNN)	0.884	0.766	0.821
Ours (HOF+MBH)	**0.969**	**0.975**	**0.971**

uses a combination of HOG and local binary pattern (LBP) features in addition to deep features for the classification of fall/non-fall videos. Feng et al. uses attention guided LSTM to capture spatio (attention module) and temporal (LSTM module) information. Table IV gives the performance comparision of the proposed method with two existing computer vision-based approaches [30, 42] on Montreal dataset. The proposed fall motion vector model (FMV) is able to discriminate well the human fall events from non-fall events on three different human fall detection datasets.

TABLE IV: Comparison of the proposed method with the existing methods on Montreal dataset.

Method	Sensitivity	Specificity
Shengke et al. [19]	0.889	0.989
Kun et al. [21]	0.937	0.920
Feng et al. [30]	0.935	0.916
Ours (3D-CNN)	0.991	0.948
Ours (HOF+MBH)	**0.993**	**1.000**

In [11], the classification performance of 3D-CNN is evaluated on Montreal dataset by considering different frame intervals. This method exhibits an average classification performance of 99.73% on Montreal dataset with an interval of one frame. Also, it can be seen that the increase of frame intervals decreases the true positive rate (TPR) due to the absence of frames relevant to fall events. It is also mentioned that sampling of frame intervals help in reducing the computation time, but does not improve the activity recognition performance. This is evident from the experimental results of the proposed method as given in Table I.

V. Conclusion

In this paper, we presented an approach for human fall detection in surveillance videos using a fall motion mixture model (FMMM) representing fall and non-fall events. To retain relevant attributes of a particular fall or non-fall videos from fall motion mixture model, factor analysis is employed on fall motion mixture model to get a low dimensional representation known as fall motion vector. The efficacy of the proposed method is demonstrated on varieties of surveillance video datasets consisting of narrow angle camera, wide angle camera, and multiple camera views. Also, the proposed method exhibits improvement in the human fall detection in surveillance videos over exiting methods. Some non-fall events are classified as fall events because of the presence of similar visual cues. So, the proposed method is unable to deal with subtle variations of some fall and non-fall events. In future, we would like to explore methods to handle variable-length patterns of human fall videos.

References

[1] U. Nations, "World population ageing:highlights," 2019.

[2] J. Clemente, F. Li, M. Valero, and W. Song, "Smart seismic sensing for indoor fall detection, location, and notification," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 2, pp. 524–532, 2020.

[3] S. Yu, H. Chen, and R. A. Brown, "Hidden markov model-based fall detection with motion sensor orientation calibration: A case for real-life home monitoring," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 6, pp. 1847–1853, 2018.

[4] Z. Liu, M. Yang, Y. Yuan, and K. Y. Chan, "Fall detection and personnel tracking system using infrared array sensors," *IEEE Sensors Journal*, vol. 20, no. 16, pp. 9558–9566, 2020.

[5] L. M. Dang *et al.*, "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition Letters*, vol. 108, p. 107561, 2020.

[6] J. Cheng, X. Chen, and M. Shen, "A framework for daily activity monitoring and fall detection based on surface electromyography and accelerometer signals," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 1, pp. 38–45, 2013.

[7] L. Kau and C. Chen, "A smart phone-based pocket fall accident detection, positioning, and rescue system," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 44–56, 2015.

[8] T. Theodoridis, V. Solachidis, N. Vretos, and P. Daras, "Human fall detection from acceleration measurements using a recurrent neural network," in *Precision Medicine Powered by pHealth and Connected Health*. Springer, 2018, pp. 145–149.

[9] M. Saleh, M. Abbas, and R. L. B. Jeannès, "Fallalld: An open dataset of human falls and activities of daily living for classical and deep learning applications," *IEEE Sensors Journal*, pp. 1–1, 2020.

[10] M. Saeidi and A. Ahmadi, "A novel approach for deep pedestrian detection based on changes in camera viewing angle," *Signal, Image and Video Processing*, pp. 1–9, 2020.

[11] N. Lu, Y. Wu, L. Feng, and J. Song, "Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 1, pp. 314–323, 2019.

[12] J. Maitre, K. Bouchard, and S. Gaboury, "Fall detection with UWB radars and CNN-LSTM architecture," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020.

[13] M. Alonso, A. Brunete, M. Hernando, and E. Gambao, "Background-subtraction algorithm optimization for home camera-based night-vision fall detectors," *IEEE Access*, vol. 7, pp. 152 399–152 411, 2019.

[14] C. Lin, S. Wang, J. Hong, L. Kang, and C. Huang, "Vision-based fall detection through shape features," in *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, April 2016, pp. 237–240.

[15] C. Jia-Luen, C. Yoong Choon, and L. Wee Keong, "A simple vision-based fall detection technique for indoor video surveillance," *Signal, Image and Video Processing*, vol. 9, pp. 623–633, 2015.

[16] E. Stone and M. Skubic, "Fall detection in homes of older adults using the microsoft kinect," *IEEE journal of biomedical and health informatics*, vol. 19, pp. 290–301, 2015.

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

[18] Y. Liao, C. Huang, and S.-C. Hsu, "Slip and fall event detection using bayesian belief network," *Pattern Recognit.*, vol. 45, pp. 24–32, 2012.

[19] S. Wang, L. Chen, Z. Zhou, X. Sun, and J. Dong, "Human fall detection in surveillance video based on PCANet," *Multimedia tools and applications*, vol. 75, pp. 11 603–11 613, 2016.

[20] B. Olivier and D. Marc Van, "Vibe: A universal background

subtraction algorithm for video sequences," *IEEE Trans Image Process*, vol. 20, no. 6, pp. 1709–1724, 2011.

[21] K. Wang, G. Cao, D. Meng, W. Chen, and W. Cao, "Automatic fall detection of human in video using combination of features," *in: IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1228–1233, 2016.

[22] K. G. Gunale and P. Mukherji, "Fall detection using k-nearest neighbor classification for patient monitoring," in *2015 International Conference on Information Processing (ICIP)*, Dec 2015, pp. 520–524.

[23] A. Poonsri and W. Chiracharit, "Fall detection using Gaussian mixture model and principle component analysis," in *2017 9th International Conference on Information Technology and Electrical Engineering (ICITEE), Thailand*, Jan 2017, pp. 1–4.

[24] A. Poonsri and W. Chiracharit, "Improvement of fall detection using consecutive-frame voting," in *2018 International Workshop on Advanced Image Technology (IWAIT)*, Jan 2018, pp. 1–4.

[25] F. Harrou, N. Zerrouki, Y. Sun, and A. Houacine, "Vision-based fall detection system for improving safety of elderly people," *IEEE Instrumentation & Measurement Magazine*, vol. 20, pp. 49–55, 2017.

[26] M. D. Solbach and J. K. Tsotsos, "Vision-based fallen person detection for the elderly," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2017, pp. 1433–1442.

[27] H. Nait-Charif and S. J. McKenna, "Activity summarisation and fall detection in a supportive home environment," in *Proceedings of the 17th International Conference on Pattern Recognition*, vol. 4, 2004, pp. 323–326.

[28] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[29] X. Wang, W. Xie, and J. Song, "Learning spatiotemporal features with 3DCNN and ConvGRU for video anomaly detection," *2018 14th IEEE International Conference on Signal Processing (ICSP)*, pp. 474–479, 2018.

[30] Q. Feng, C. Gao, L. Wang, Y. Zhao, T. Song, and Q. Li, "Spatio-temporal fall event detection in complex scenes using attention guided LSTM," *Pattern Recognition Letters*, vol. 130, pp. 242–249, 2020.

[31] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *International Conference on Computer Vision Pattern Recognition*, 2011, pp. 3169–3176.

[32] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558.

[33] D. Roy, K. S. R. Murty, and C. K. Mohan, "Unsupervised universal attribute modeling for action recognition," *IEEE Transactions on Multimedia*, vol. 21, no. 7, pp. 1672–1680, July 2019.

[34] H. Aronowitz and O. Barkan, "Efficient approximated i-vector extraction," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4789–4792.

[35] I. Charfi, J. Mitéran, J. Dubois, M. Atri, and R. Tourki, "Optimised spatio-temporal descriptors for real-time fall detection: comparison of SVM and Adaboost based classification," *Journal of Electronic Imaging (JEI)*, vol. 22, no. 2013, Oct. 2013.

[36] B. Kwolek and M. Kepski, "Human fall detection on embedded platform using depth maps and wireless accelerometer," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 3, pp. 489–501, Dec. 2014.

[37] I. Charfi, J. Miteran, M. Dubois, J.and Atri, and R. Tourki, "Definition and performance evaluation of a robust SVM based fall detection solution," *in: IEEE Eighth International Conference on Signal Image Technology and Internet Based Systems (SITIS)*, pp. 218–224, 2012.

[38] M. Li, G. Xu, B. He, X. Ma, and J. Xie, "Pre-impact fall detection based on a modified zero moment point criterion using data from kinect sensors," *IEEE Sensors Journal*, vol. 18, no. 13, pp. 5522–5531, July 2018.

[39] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and Imagenet?" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6546–6555.

[40] J. A. Suykens and J. Vandewalle, "Least squares support vector machine classifiers," *Neural processing letters*, vol. 9, no. 3, pp. 293–300, 1999.

[41] M. Chamle, K. G. Gunale, and K. K. Warhade, "Automated unusual event detection in video surveillance," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, vol. 2, Aug 2016, pp. 1–4.

[42] A. Abobakr, M. Hossny, H. Abdelkader, and S. Nahavandi, "RGB-D fall detection via deep residual convolutional LSTM networks," in *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2018, pp. 1–7.

[43] S. Bhandari, N. Babar, P. Gupta, N. Shah, and S. Pujari, "A novel approach for fall detection in home environment," *in:IEEE 6th Global Conference on Consumer Electronics (GCCE)*, pp. 1–5, 2017.

[44] N. Zerrouki, F. Harrou, Y. Sun, and A. Houacine, "Vision-based human action classification using adaptive boosting algorithm," *IEEE Sensors Journal*, vol. 18, no. 12, pp. 5115–5121, 2018.

Chalavadi Vishnu received his B.Tech (CSE) from JNTU Hyderabad in 2016 and M.Tech (CSE) from IIT Hyderabad, in 2018, in India. Currently he is pursuing PhD (CSE) at IIT Hyderabad. His research interests include learning representations of video activities and drone data, and autonomous vehicles.

Rajeshreddy Datla received B.Tech.(2002) and M.Tech.(2012) in CSE from JNTU Hyderabad, India. He is currently pursuing PhD (CSE) at IIT Hyderabad (India). His research interests include satellite image processing and machine learning.

Debaditya Roy (Member IEEE) received B.Tech. from the West Bengal University of Technology, in 2011, M.Tech. from the NIT Rourkela, in 2013, and PhD (CSE) from IIT Hyderabad, in 2018, in India. He is currently a Research Scientist at IHPC, A*STAR, Singapore. His research interests include action recognition and surveillance analysis.

Dr. Sobhan Babu received B.E from University of Madras, in 1999, M.Tech (CSE) from IIT Bombay, in 2001 and PhD (CSE) from Institute of Technology Bombay (IITB), in 2007, in India. He is currently an Associate Professor in CSE at IIT Hyderabad, India. His research interests include big data analytics, graph theory and algorithms.

Dr. C. Krishna Mohan received M.Tech (SACA) from NIT Surathkal in 2000, and PhD (CSE) from IIT Madras in 2007, in India. He is currently Professor in the dept. of CSE, IIT Hyderabad (India). His research interests include video content analysis and machine learning. He is a senior member of IEEE.