Engineering Applications of Artificial Intelligence 77 (2019) 21–45

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A review of state-of-the-art techniques for abnormal human activity recognition

Chhavi Dhiman [a], Dinesh Kumar Vishwakarma [b,*]

[a] Department of Electronics and Communication Engineering, Delhi Technological University, Delhi, 110042, India
[b] Department of Information Technology, Delhi Technological University, Delhi, 110042, India

ARTICLE INFO

Keywords:
Two-dimensional anomaly detection
Three-dimensional anomaly detection
Crowd anomaly
Skeleton based fall detection
Ambient Assistive Living

ABSTRACT

The concept of intelligent visual identification of abnormal human activity has raised the standards of surveillance systems, situation cognizance, homeland safety and smart environments. However, abnormal human activity is highly diverse in itself due to the aspects such as (a) the fundamental definition of anomaly (b) feature representation of an anomaly, (c) its application, and henceforth (d) the dataset. This paper aims to summarize various existing abnormal human activity recognition (AbHAR) handcrafted and deep approaches with the variation of the type of information available such as two-dimensional or three-dimensional data. Features play a vital role in an excellent performance of an AbHAR system. The proposed literature provides feature designs of abnormal human activity recognition in a video with respect to the context or application such as fall detection, Ambient Assistive Living (AAL), homeland security, surveillance or crowd analysis using RGB, depth and skeletal evidence. The key contributions and limitations of every feature design technique, under each category: 2D and 3D AbHAR, in respective contexts are tabulated that will provide insight of various abnormal action detection approaches. Finally, the paper outlines newly added datasets for AbHAR by the researchers with added complexities for method validations.

1. Introduction

Abnormal Human Action Recognition (AbHAR) through visual data and sensor data has appealed, tremendous attention in the computer vision community. It is because of its gigantic potential in wide spectrum of areas i.e. Ambient Assistive Living (AAL) (Ismail et al., 2015; Dragone et al., 2015; Andò et al., 2015; Rafferty et al., 2017; Zolfaghari and Keyvanpour, 2016), and healthcare of elderly people (Zhao et al., 2014; Cardile et al., 2010), Intelligent Video surveillance systems (Hsieh et al., 2015; Li et al., 2014a; Coşar et al., 2017; Ben and Zagrouba, 2018), human–computer interfaces (HCI) (Mosquera et al., 2017; Zhu et al., 2017; Feng et al., 2016; Roudposhti et al., 2017; Liu et al., 2014), sports (Zhu et al., 2009; Shih, 2017; Stein et al., 2016), event analysis, robotics (Chien et al., 2005), intrusion detection system (Yu et al., 2016; Aburomman and Reaz, 2016), content-based video analysis (Song and Fan, 2006; Kulkarni et al., 2015), multimedia semantic annotation and indexing (Chou et al., 2017) etc. Abnormal Human Activity Recognition and Anomaly detection are used interchangeably which refers to non-conforming patterns in a given data. Abnormalities are the rarely occurring activities that are different from others. The accuracy and efficiency of the abnormal HAR system depend on the features representing the actions. A robust and highly discriminative feature representation (Vishwakarma et al., 2015b, a, 2016a; Aggarwal and Vishwakarma, 2016) of the action can lead to better recognition results under various challenges i.e. lighting conditions, cluttered background, partial or complete occlusion, crowded scenes, different viewpoint of the camera, size, shape, appearance and complexity of human actions. Such challenges have always pushed forth the researchers to explore new dimensions of the solution from vision-based to sensors based Surveillance Systems integrating multiple features, over the years. Various algorithms (Singh et al., 2016; Chathuramali et al., 2014; Nguyen et al., 2016; Panahi and Ghods, 2018) have been developed keeping different challenging scenarios in mind by the researchers, but one thing remains common that action is defined by two attributes: motion magnitude and motion orientation.

Anomaly detection has been surveyed and reviewed in number of articles, as well as books highlighting various techniques/methods (classification, clustering, nearest neighbor, statistical (Hodge and Austin, 2004), information theoretic and spectral-based techniques) to perform the detection more effectively under different constraints and also various applications such as cyber intrusion detection (Synder, 2001),

* Corresponding author.
E-mail addresses: chhavi1990delhi@gmail.com (C. Dhiman), dvishwakarma@gmail.com (D.K. Vishwakarma).

https://doi.org/10.1016/j.engappai.2018.08.014
Received 6 July 2017; Received in revised form 12 June 2018; Accepted 27 August 2018
Available online 27 September 2018

fraud detection (Chen and Gangopadhyay, 2016), medical anomaly detection, industrial damage detection, image processing, textual anomaly detection (Gorai et al., 2016), sensors networks. This survey focuses on the latest methodologies marking their impacts on the performance of a robust Abnormal HAR system based on two-dimensional (intensity) and three-dimensional (depth/skeleton) input by covering a wide range of applications from a single person based application-AAL, fall detection to multi-persons based application-crowd analysis.

Motivation: The European Statistical Office reported that by 2060, the ratio of young and old person will be 1:1 in the EU (EC, 2012). In addition to it, World Health Organization (World Health Organization, 2008) mentioned that injuries due to fall will shoot up by 100% by 2030. Henceforth, the smart home is drawing a lot of considerations by many researchers (Zolfaghari and Keyvanpour, 2016; Li et al., 2015a; Rashidi and Mihailidis, 2013) to support peoples' health. Technological inventions and development play a significant role in extending support to them through assistive and autonomous care facilities, preferably based on a low-cost setup that can be set up at homes or care centers. Nowadays, depth images have taken over intensity images. Firstly, because they are illumination invariant and can provide a 3D model of the scene. Secondly, with depth images segmentation, background subtraction, and motion estimation, silhouette extraction has been significantly simplified and robust by advancement in depth sensor hardware i.e. time-of-flight (TOF) cameras, low-cost structures light sensor (Microsoft Kinect). Availability of three-dimensional data (depth or skeleton) has opened a different direction for abnormal Human action recognition. Therefore, various real-time depth and skeleton based fall detection systems (Ma et al., 2014; Akagunduz et al., 2016; Zhang et al., 2012; Diraco et al., 2010) are evolving considering the affordable range of the common user and various challenges. In addition to this, deep architectures (Hammerla and Plotz, 2015) have also foot-stepped in the field of computer vision and are used for automatic assessment of Parkinson's disease, AAL applications and many more. The work includes both two-dimensional and three-dimensional AbHAR systems, in which, research is progressing rapidly. The significant contributions of the papers are as follows:

- A structured overview of Abnormal Human Activity Recognition (AbHAR) is provided through the comprehensive survey. Broadly, it covers abnormal human activity recognition i.e. handcrafted-2D and 3D AbHAR based on RGB, depth and skeleton evidence and deep single person and multiple person based methodologies as illustrated in the taxonomy, (Fig. 3).
- A 2D AbHAR state-of-arts for a single person and a multiple persons AbHAR systems have been presented, which target smart home surveillance and also any unusual behavior at a public place (crowd).
- A 3D AbHAR system state-of-arts have been reviewed and presented the depth based, and skeleton based abnormal activity recognition system designs have significant applications in elderly healthcare and AALs, ADL, and fall detection.
- The literature is updated with the application of recent advances of deep neural networks in the field of abnormal human action recognition, Section 3.3 by analyzing the challenges present.
- A systematic arrangement of discussed 2D and 3D features extraction methods have been discussed and well analyzed in Tables 1 and 2 and 4–7 by highlighting their key contributions and limitations.

Organization of our survey is as follows. Section 2, provides a panoramic summary of the related state-of-the-art survey works in the area of abnormal human action recognition followed by paper count analysis per year. It will help the reader to get an overview of key contributions of previous surveys done. Section 3, discusses the outlined taxonomy for categorizing anomaly detection techniques evolved from 2006 to 2018. Every aspect for a robust and real-time abnormal human action recognition system are closely discussed and analyzed in Section 4. Section 5, outlines recently introduced publicly available datasets used for abnormal activity recognition in the mentioned works. Finally, in Section 6, peculiar observations and possible directions are highlighted, that need to further explore for research in the field of AbHAR.

2. Earlier state-of-the-arts surveys

A Reliable and Intelligent Abnormal Behavior Surveillance Systems are the need of the hour to mitigate the effects of abnormal/ undesired activities in public place, health-care organizations and smart homes and other security threats. Therefore, tracking and identifying objects and human motion in surveillance videos, trailed by automatic encapsulation of the content has turn out to be a hot topic of research (Gowsikhaa et al., 2014). Fig. 1 illustrates statistical compilation of some papers on abnormal activity recognition, reviewed in our survey from each category that has evolved over the period 2006 to present. It is clearly evident from the literature that research has increased rapidly in the field of AbHAR (RGB, depth, and skeleton based) in last few years. A Few surveys (Rashidi and Mihailidis, 2013; Chaaraoui et al., 2012) are identified, discussing about computer vision solutions for elderly health care, home surveillance and AAL. Rashidi and Mihailidis (2013) talked about Ambient-Assisted Living tools and techniques for older adults from types of wearable and ambient sensors to vision sensors. Chaaraoui et al. (2012) highlighted the challenges of sensor technologies, limited assistive robot technologies and social security and privacy issues of AAL systems to make it widely acceptable among the users. Whereas, in Li et al. (2015a) the focus is brought to IoT, wearable devices, cloud computing, advanced robotics, sensor networks based assistive living products to discern the wider frontiers of AAL for healthcare, rehabilitation and assistive living. Some surveys (Chen and Wei, 2013; Presti and Cascia, 2016; Edwards et al., 2016; Synnott et al., 2015) conferred about depth imagery based human motion analysis, 3D some skeleton based human action classification and introduced new datasets for handling complex interactions and smart home activities respectively. A broad survey (Popoola and Wan, 2012) of video-based anomaly detection has brought forward diversified work much more in-depth. It defined the characteristics of an anomaly and context-based anomaly which may not be an anomaly in another frame of reference and discussed various scene behavior modeling methods, considering behavior abstraction. Probability of occlusion is relatively less in uncrowded scenes, therefore tracking based approach (Piciarelli and Foresti, 2011) is suitable for action pattern mining in uncrowded videos. However, in crowded scenes pixel level or frame level anomaly detection should be preferred to handle occlusion.

Over the year's behavior modeling focus has shifted from rule-based methods to probability-based statistical methods, being superior in robustness and scalability. But the real challenge still lies in making the system free from human interventions and minimizing false alarm (positive/negative) rate. For which global and local features have proved to be highly descriptive and discriminative in nature to encode the behavior. In addition to it, trajectory, speed, and direction, optical flow, object-based abstraction methods are also used for complete abstraction of the scene behavior. Whereas, for high dimensional behavior modeling, learning with Generative Topic Model framework (Popoola and Wan, 2012) generate more robust sparse spatial–temporal interest points than and can determine eloquent activities from co-occurrence of visual words automatically. Such complex scene models are validated with UMN dataset (Detection of unusual crowd activity dataset, 2006), web dataset (Web Dataset, 2009), UCSD (UCSD Anomaly Detection Dataset, 2013), CAVIAR (CAVIAR Test Case, 2005), PETS dataset (Patino et al., 2016), which are grouped based on the scene density i.e. single person, sparse, relatively dense. From the literature (Turaga et al., 2008), it is notified that graphical models (i.e. dynamic Bayes net, propagation nets, Petri nets), syntactic (i.e. context-free grammars, attribute grammar and stochastic CFG), and knowledge-based approaches (i.e. ontologies, logic rules, constraint satisfaction) are

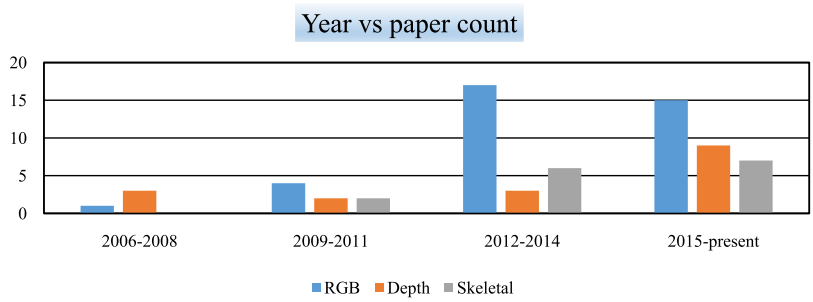

Fig. 1. The graph demonstrates the number of papers reported in this survey on AbHAR from 2006–present.

preferred to model complex scenes, the inherent structure and semantics of complex activities and provide higher level representation for activity recognition. Whereas, Chandola et al. (2009) covered wide application areas of anomaly detection i.e. anomaly detection in scripts, sensor network, image processing, medical and communal health anomaly detection, fraud detection, network intrusion detection.

Fall (Aslan et al., 2017) is one of the major unusual events occurring at any public or private place that needs to be identified well in time and help should be extended according to the severity of fall (child/older person). A comprehensive survey (Mubashir et al., 2013) is laid out describing various existing approaches for fall detection dividing them into three categories such as wearable device, ambient device, and vision-based approaches. Wearable and ambient device based methodologies are scenario dependent whereas vision-based approaches are generalized, which induce no intrusion to the person under surveillance.

In vision-based approaches (Nguyen et al., 2016; Panahi and Ghods, 2018; Rougier et al., 2011b; Jansen and Deklerck, 2006) spatiotemporal details, shape deformation features, posture information and inactivity information after fall, and 3D head position have been used enormously over the years to design an efficient real-time fall detection system with standard computing platforms, and low-cost cameras but a wide area is still untouched that deals with the robustness of the system. Recently in Chen and Wei (2013), an overview of various available depth sensors and their benefits over conventional cameras are stated. It is noticed that the growing research area is addressing human action recognition as normal or abnormal and focusing on depth based body part detection and pose estimation, body pose modeling, and space–time evidences. Synnott et al. (2015) discussed the need for simulated sensor data generated in the field of behavior analysis algorithm/models/data-driven learning approaches/classification mechanisms. All the pose oriented datasets introduced between 2009 and 2015 are reviewed in the survey (Edwards et al., 2016). A new dataset CONVERSE representing Complex Conversational Interactions between two individuals via 3D poses in the survey has opened more possibilities for Abnormal Human activity Recognition (AbHAR). This dataset caters real-world challenging scenarios incorporating frequent primitive actions, interactions, and motion over a period of time. It is clearly evident that in this decade, it is the vantage point for posed based 3D abnormal human activity recognition in research where enormous information is available Presti and Cascia (2016) discussed different aspects of data pre-processing, publically available benchmarks, and commonly used accuracy measurements along with feature representation and 3D Skeleton based action classification at length. In the year 2013 three kinds of literature (Thida et al., 2013; Jo et al., 2013; Kok et al., 2016) were presented which laid out three different perspectives of crowd analysis. Thida et al. (2013) emphasized macroscopic and microscopic modeling with crowd event detection. Jo et al. (2013) highlighted physic based approaches for the crowd and group analysis in computer vision whereas, Kok et al. (2016) discussed state-of-the-art methods from the physics and biologically inspired perspective for crowd activity analysis as shown in Fig. 2.

Li et al. (2015b) have drawn a general structure of crowd scene analysis. According to which, crowd activity can be recognized via holistic approach or object-based approach. Initially, authors (Zitouni et al., 2016) talked about various motion representation methods in crowd scenes- flow-based features, spatiotemporal features and trajectory based features and motion pattern segmentation techniques — Flow Field Model-based segmentation, correspondence-based clustering, and probability-based model clustering. The recent work by Tripathi et al. (2018) defined four aspects of crowd analysis with recent advances of CNN architectures, as Crowd counting and density estimation, Crowd motion detection, Crowd tracking, Crowd behavior understanding and their benefits over traditional methods (Loy et al., 2013; Hassner et al., 2012). The work addresses the fact that three steps must be followed to analyze crowd behavior with real-time challenges i.e. crowd modeling, crowd monitoring and crowd management it deals with all the strategic decision-making to control the crowd by maintaining social i to optimize operational efficiency.

3. Abnormal human activity recognition methods

Abnormal Human Activity Recognition (AbHAR) methods can be broadly categorized into Two-Dimensional AbHAR systems and Three-Dimensional AbHAR systems on the basic type of input fed to the system. The detailed classification of AbHAR system is shown in Fig. 3. A two-dimensional AbHAR system is fed with 2D silhouettes for single person AbHAR in the majority of mentioned approaches. Whereas, three-dimensional AbHAR systems are fed with depth silhouettes and skeleton structures of the person for AbHAR. Both kinds of approaches have their own set of applications and benefits. A detailed description of feature representation of the abnormal human activity is mentioned in Sections 3.1 and 3.2 with an analytical discussion in Section 3.3.

3.1. Two-Dimensional AbHAR

The RGB intensity and object detection (spatial information) are some elementary observations of an image. A significant amount of research has been done by researchers using 2D RGB/intensity/spatial information of an image for normal and abnormal HAR systems. In this section, 2D-AbHAR approaches are broadly categorized as single person AbHAR and Multiple Persons AbHAR. The former one is further grouped as silhouette based, Spatio-temporal based approaches and some miscellaneous techniques on the bases of their significant contributions in understanding the single person action recognition. Multiple Persons AbHAR methods are summarized as Spatio-temporal approaches, sparse representations based approaches and some miscellaneous ones. In each category different types of feature representations of action are discussed to provide a brief understanding of existing methods.

3.1.1. Single person AbHAR

One of the major application of Single Person AbHAR is in Ambient Assistance Living (AAL), home monitoring where a single person is under surveillance for abnormal activity recognition. Therefore, in this section, we have included single person AbHAR methodologies supporting elderly health care issues and AAL (Li et al., 2015a). Table 1 briefs

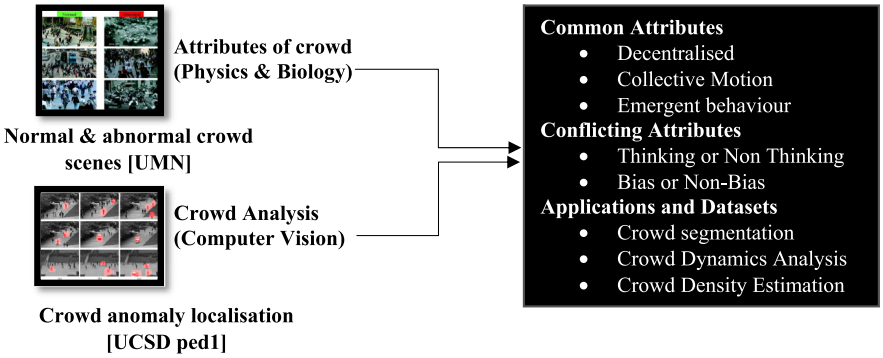

Fig. 2. Crowd Analysis (Kok et al., 2016).

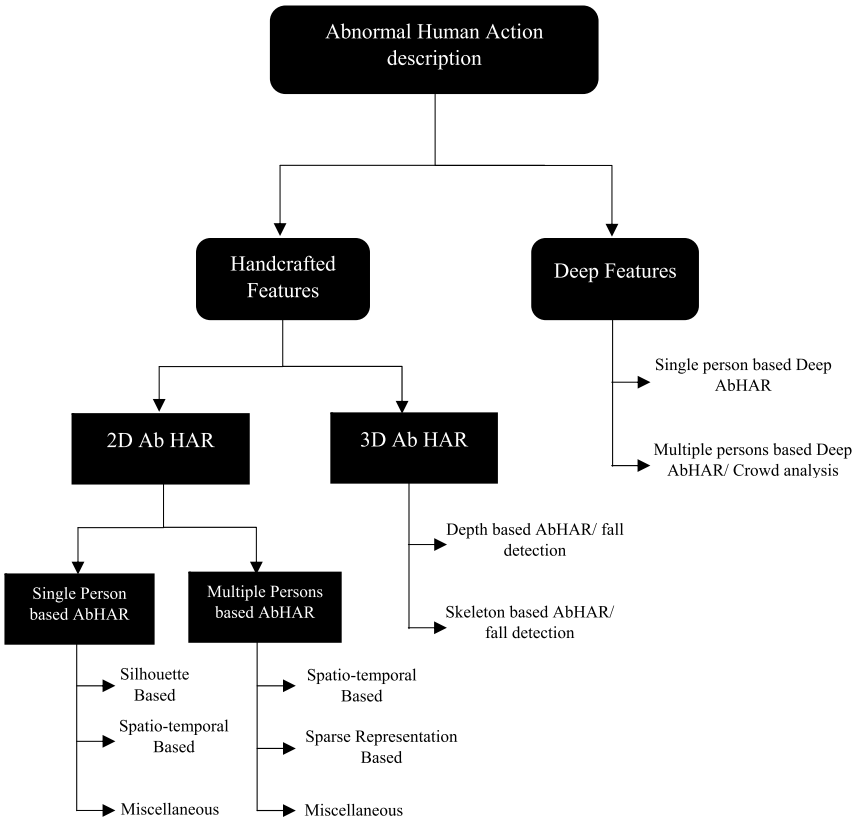

Fig. 3. Taxonomy of Abnormal Human Activity Recognition (AbHAR).

about various approaches based on two-dimensional fundamentals for single person activity recognition. It highlights the scene representation and feature extraction techniques with individual's key elements and limitations for a better understanding of the method.

It can be observed from Table 1that LOTAR framework (Riboni et al., 2016) offers a stronger feature representation platform for AAL application which analyzes both short term and long term anomalies by collecting data from multiple sensors i.e. temperature, pressure & RFID sensors along with vision sensors. For experimentation, the framework is employed in the real patient home, which needs to be extended to multiple individuals for realistic results. And in work (Huang et al., 2014), the habit of the person is studied and analyzed for the first time by fusing ISUS (Intelligent space for understanding and service) and multi-camera positioning algorithm. Single Person, AbHAR approaches are discussed in detail by categorizing them as silhouette based, spatiotemporal based and some miscellaneous approaches, below.

3.1.1.1. Silhouette-based approaches. Robustness of any AbHAR method depends on how efficiently action evidence in an image are extracted as a feature. R-Transform has been used frequently (Khan and Sohn, 2013, 2011) to target elderly health care issues and providing a highly practical solution to it. Khan and Sohn (2011) took into account six possible unusual activities (i) faint (ii) backward fall (iii) forward fall (iv) vomit (v) chest pain and (vi) a headache (see Fig. 5). It integrated R-Transform with Kernel Discriminant Analysis (KDA) to minimize interclass similarity of different activity postures as binary silhouette maintaining individual's privacy.

Khan and Sohn (2013) designed a two-level hierarchical anomalous human activity recognition system to increase the recognition rate for intra-class activities, particularly for falling forward-vomiting postures and falling backward-fainting postures. The first level performs KDA on the R-Transform of binary silhouettes. In the second level, a k-mean clustering algorithm is applied to make groups of similar postures and four state HMM classifier in training and recognition of the activities

Table 1
Scene representation & feature abstraction techniques for two-dimensional single person abnormal human activity recognition (AbHAR) approaches.

Scene representation & feature abstraction techniques	Year	Ref.	Key contributions	Limitations
HMM+SVM	2010	Hung et al. (2010)	Inexpensive sensor network	Collect data from RFID-based sensor network
Separable linear filter (spatiotemporal feature) +Topic HMM	2011	Zhu and Liu (2011a)	Efficient extraction of variable length action sequences and construct highly descriptive features.	THMM is sensitive to the noise of Spatio-temporal words.
R-Transform and Kernel Discriminant Analysis	2013, 2011	Khan and Sohn (2013) and Khan and Sohn (2011)	Scale and translational invariant and reduced intraclass false recognition rate (FRR)	Binary silhouettes used create ambiguities in discerning the certain postures in case of self-occlusion (front-view)
Jerk based Inactivity Magnitude (JIM) feature	2014	Candás et al. (2014)	An autonomous framework that makes no assumptions about the activity data distribution	The experiment needs to be performed with more users under different conditions.
Key points' duration histogram	2014	Huang et al. (2014)	multi-camera positioning algorithm advances the positioning accuracy	The method can be applied in the intelligent space where the elderly lives regularly
Shape Model based on OMEGA equation	2016	Al-Nawash et al. (2016)	automatic real-time video-based surveillance system with simultaneous tracking, semantic scene learning, and abnormality detection ability in an academic environment	The complexity of the algorithm is not discussed
LOTAR framework	2016	Riboni et al. (2016)	Detects both Short term and Long term anomalies	Heavily depends on indicator models developed by cognitive neuroscience experts.

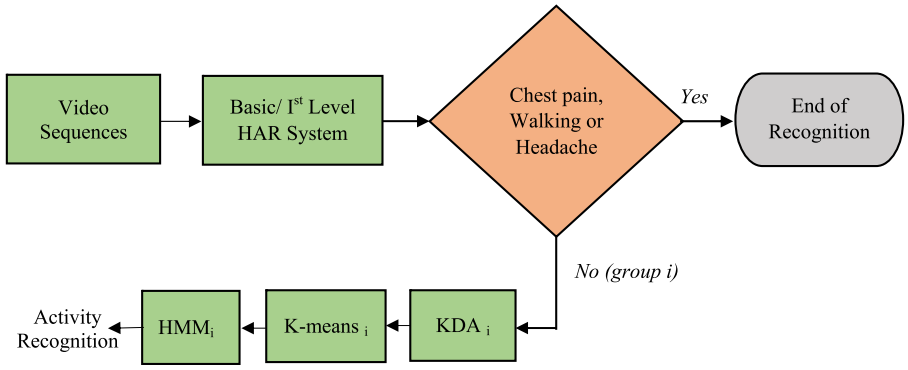

Fig. 4. Block diagram of hierarchical HAR system.

(see in Fig. 4). This hierarchical approach has increased the average recognition rate of similar activities to 97.1% in a confined environment which needs to be extended with real-time abnormal activity dataset augmented with noise. Binary silhouette fails to describe the posture in case of self-occlusion, this limitation can be removed using depth silhouettes . Early detection of short-term or long-term abnormal behavior of the person under AAL (Ambient Assistive Living) is emerging as a promising field of abnormal activity recognition. With the help of medical models, provided by cognitive neuroscience researchers a FABER hybrid technique is presented (Riboni et al., 2015) which focuses on abnormal activity routines to identify symptoms of mild cognitive impairment (MCI) at a fine-grained level with the assimilation of supervised learning and symbolic reasoning i.e. problems with memory, language, thinking, and judgment which is a significant information for early detection of MCI.

Generally, the camera or sensor-based monitoring systems (Crispim Junior et al., 2012) in the closed environment have the limitation of individual privacy. The approaches which monitor low-level behavioral indicators such as walking speed steps taken, they fail to acquire fine-grained depiction of the action during execution of Instrumental Activities of Daily Living (IADL). Some approaches (Sacco et al., 2012) require high implementation cost and cannot be applied for a long-term basis. For long-term abnormal behavior recognition, Riboni et al. (2016) developed a LOTAR framework (see Fig. 6) which

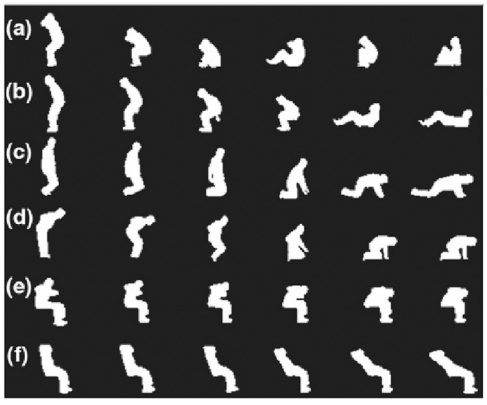

Fig. 5. Selected postures for the six abnormal activities: (a) fainting, (b) falling backward, (c) falling forward, (d) vomiting, (e) chest pain, and (f) a headache (Khan and Sohn, 2013).

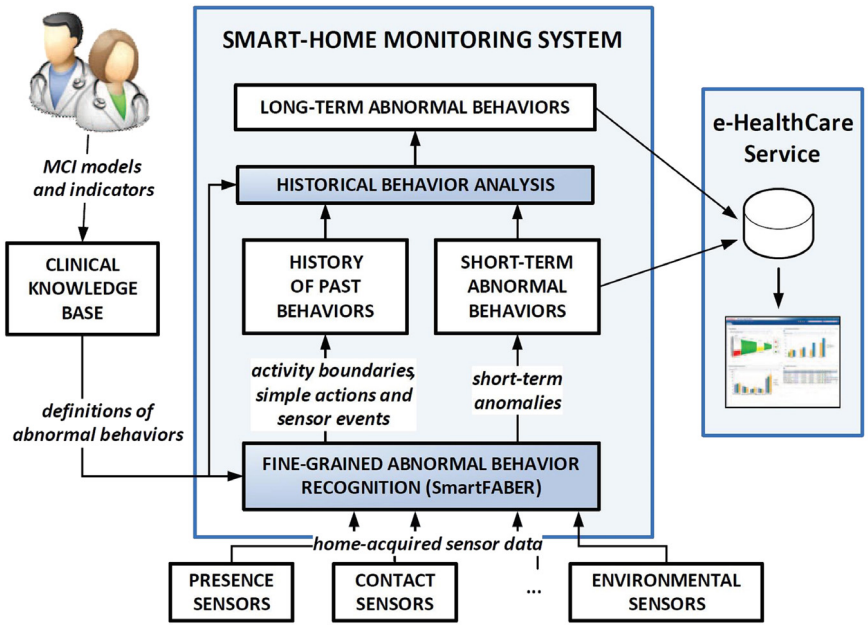

Fig. 6. The LOTAR framework to recognize long-term abnormal behaviors (Riboni et al., 2016).

is a hybrid behavior analysis system integrated with knowledge-based and data mining based machine learning techniques. It acquires data from non-intrusive sensors set up in the patient's home preserving the individual's privacy and fine-grained process statistics vital for long-term abnormal behavior recognition.

3.1.1.2. Spatio-temporal based approaches. It has been seen that spatiotemporal representation (Vishwakarma and Singh, 2016; Diba and Gool, 2016; Guo et al., 2008; Li et al., 2015c) is used abundantly to describe the actions in a video sequence for normal/abnormal human action recognition. Li et al. (2015c), utilised spatiotemporal contextual details for anomaly detection and localization from captured video volumes.

An unsupervised statistical learning structure is developed which excavates global activity patterns and local salient behavior patterns using clustering and sparse coding respectively along with multiple scale analysis to detect and localize the anomalous region efficiently. However, during this process, codebook and dictionary learning algorithm must update them for new input videos to cater a practical solution in real-time applications. Phase Spectrum of Quaternion Fourier Transform (PQFT) (Guo et al., 2008) is identified as a potential solution for the spatiotemporal Salience Detection in videos frames. It has been demonstrated experimentally that PQFT works with less computational complexity, being robust to noise and faster than existing methods i.e. spectral Residue (SR), Phase Spectrum of Fourier Transform (PFT), Neuromorphic Vision Toolbox (NVT), and Saliency Toolbox (SBT). Addition of motion cue with PQFT further strengthens the representation of spatiotemporal saliency in videos in the presence of white-colored noise. However, if the noise is similar to salient object, the approach fails to detect the saliency. PQFT cannot detect the closure patterns of the image, which humans can identify very quickly in an image. Vishwakarma et al. (2016b) defined an activity recognition framework using shape and motion orientation of the object as baseline features. In which spatial edge distribution of gradients and texture based segmentation technique are used to extract binary silhouette. However, the work needs to be evaluated with more complex and real-life environment to provide the solution for real-life applications.

3.1.1.3. Miscellaneous. Nowadays the existing methods (classification based, clustering based and statistical methods) for abnormal human behavior detection are facing the challenge of making the system independent of human arbitrations. Candás et al. (2014) made an attempt technique to detect human behavior abnormality autonomously, without making any assumption using Jerk based Inactivity Magnitude (JIM) feature. However, the experimental work needs to be extended for more users and under different conditions. Huang et al. (2014) utilised intelligent space for understanding and user service (ISUS) for abnormal habit recognition problem, for the first time, with multi-camera positioning algorithm to locate the person using the concept of distribution of time spent at different duration histogram of key points in the intelligent system under surveillance for the entire day that makes the habit recognition task simpler. However, the application of this model is limited to ISUS environment to monitor the habit of the elder people, which may not be under affordable range of an individual. There is a need to develop low-cost fall detector (Miguel et al., 2017) for smart homes based on artificial vision algorithms that can be available in affordable range to the common man.

3.1.2. Multiple persons AbHAR

Multiple Persons AbHAR is a generalized case of Crowd Activity analysis, which are used interchangeably in this review. It is a novel area of interest in computer vision which will potentially advance to new application domain such as automatic detection of riots or chaotic acts in crowd and localization of the abnormal areas in scenes for high resolution investigation. Occlusions and ambiguities in crowded scenes with complex activities and scene semantics are some of the challenges needs to be addressed for crowd activity analysis. It is systematically studied in the field of transportation and communal security. But the level of complexity is very high to define the complete crowd activity with the help of a feature. There are two approaches to analyze the crowd activity (1) identify group activity patterns — abnormal activity localization (Antic and Ommer, 2015; Fagette et al., 2014) (2) represent an individual's behavior in the crowd i.e. trajectory-based approach (Chathuramali et al., 2014), and shape dynamics based approach (Vaswani et al., 2005). Table 2outlines various approaches for multiple persons AbHAR specifying feature abstraction methods with their key elements of the approaches and their limitations.

It is observed from Table 2 that the optical flow based motion descriptions — SOF (Chathuramali et al., 2014), MHOF (Cong et al., 2013), HNF (HOG+HOF) (Zhao et al., 2015), Histogram of Maximal Optical Flow Projection (HMOFP) (Li et al., 2016), extract meaningful statistics

Table 2
Scene representation & feature abstraction techniques for two-dimensional multiple persons abnormal human activity recognition (AbHAR) approaches.

Scene representation & feature abstraction techniques	Year	Ref.	Key elements	Limitations
Trajectory Sparse Reconstruction Analysis	2012	Li et al. (2012)	Number of control points and weight factors, makes the method flexible to adapt to the complicated shape of curves.	Performance is highly sensitive to control point parameters of the trajectory.
Spatio-temporal volume construction	2013	Roshtkhari and Levine (2013)	Robust to spatial and temporal scale variations under varying brightness conditions.	Fails to detect an anomaly in sudden variations in dynamic backgrounds
Silhouettes and Optic Flow-based Features (SOF) +Dense Trajectory-based Features (DTF)	2014	Chathuramali et al. (2014)	Adaptive to dynamic background	The complexity of DTF descriptor generation is high
Laplacian Eigen-map (LE)	2013	Eng et al. (2013)	Local probabilistic model allows to detect abnormalities in both local and global contexts	pixel-level abnormality localization is responsible for high computation cost
Dynamic Patch Grouping(DPG) +Multilayer Histogram of Optical Flow (MHOF) Edge Orientation Histogram (EOH)	2013	Cong et al. (2013)	DPG adaptively cluster similar patches and process only foreground region which makes it computationally efficient	DPG performance reduces with increase in the size of patch and noise in the frame.
A mixture of dynamic textures (MDT) models	2014	Li et al. (2014b)	Spatial and temporal scores at multiple scales ensure global uniformity of anomaly decision	Occlusion is not discussed
Dominant sets	2014	Alvar et al. (2014)	sparse representation	A sudden change in direction and speed are responsible for FP error
Laplacian Sparse Representation (LSR) of spatiotemporal HNF (HOG+HOF) feature	2015	Zhao et al. (2015)	LSR is responsible for minimum feature quantization error	Computational time involved in feature generation is not discussed
Histogram of Maximal Optical Flow Projection (HMOFP) descriptor	2016	Li et al. (2016)	K-SVD is used for dictionary optimization	User-defined threshold τ controls the abnormal event detection
ICA-based linear representation model-(off-line long-term sparse representation (LTSR) and on-line short-term sparse representation (STSR))	2016	Wang et al. (2017b)	High computational efficiency due to selective visual attention analysis	Needs better saliency detection method.
Region Association Graph (RAG)	2017	Dogra et al. (2015)	Movement patterns-trajectories	Manual classification of trajectories is required for supervised training (HMM)

to describe densely crowded scenes. The trajectory-based methods — Region Association Graph (RAG) (Dogra et al., 2015), Dense Trajectory-based Features (DTF) (Chathuramali et al., 2014; Li et al., 2012) can easily handle the complexity of shapes but involves high computational cost. Recently Saini et al. (2017) extended the concept of RAG by defining a high-level non-overlapping block based scene representation using raw object trajectories for video scene segmentation. As far as real-time applicability of the method is concerned, ICA-based linear representation model (Wang et al., 2017b) is implemented in a real-time scenario with high computational efficiency due to selective visual attention analysis using both long term and short term sparse representation of crowded scenes. Multiple persons AbHAR based approaches are discussed in detail by categorizing them as Spatio-temporal based approaches, sparse representation based approaches, and miscellaneous ones.

3.1.2.1. Spatio-temporal based approaches. Effective encryption of complex details of the dynamic background and multiple variations of foreground in crowded or public place demands both spatial and temporal information (Zhan et al., 2016) of the scenes. In the work (Mehran et al., 2009) dynamics of a crowd is represented using social force model without tracking the objects**.** Where moving particles are considered as individuals whose new positions are measured with space–time average of optical flow. And Force Flow is observed for each pixel of every frame and arbitrarily selected spatiotemporal volumes of Force Flow are used to design the ordinary activity/ behavior of the crowd. In this work, bag of words methodology is used for classification of frames as normal/abnormal. Escape panic scenarios of UMN dataset (Detection of unusual crowd activity dataset, 0000), a challenging and interesting crowd videos datasets from the web (Web Dataset, 2009) are used for validation of this scheme. To overcome unstable foreground excavation, crowd distribution is defined with the concept of particle entropy (Gu et al., 2014) integrated with GMM. Eng et al. (2013) used the concept of Laplacian Eigen-map (LE) to detect and localize the abnormality in the video scene robustly. It represents the local regions and edges in frame as nodes of a graph. And the nodes are assigned with weights by considering the relation of local regions on the basis of joint similarity in feature, space and time domain. Since the dynamics of the crowd changes instantaneously, therefore, less computation time to perform the task in real life crowded scenes is one of the key issues, which is difficult to achieve in graph-based approaches. In one of the studies (Wang et al., 2017b) the author aims to detect and localize the abnormal act in crowd using the concept of visual attention analysis. The presented approach used two sparse coding strategies: Off-line long-term sparse representation and online short-term sparse representation, for spatiotemporal feature extraction. Dense Trajectory-based features (DTF) (Chathuramali et al., 2014) can handle dynamics of the background with near to perfect accuracy however, DTF fail when both normal and abnormal actions occur in the same frame and require high performance hardware. DTF–SOF descriptors are used to train SVM lending small time complexity. A recent work (Saini et al., 2017) highlighted the importance of complete scene understanding along with motion detection by defining block based scene representation. It extracted high-level features from object trajectories using non-overlapping block-based representation of surveillance scene. And Hidden Markov Model (HMM) learned the parameters that describes the dynamics of a given surveillance scene. Anomaly detection can also be formulated either as an outlier detection

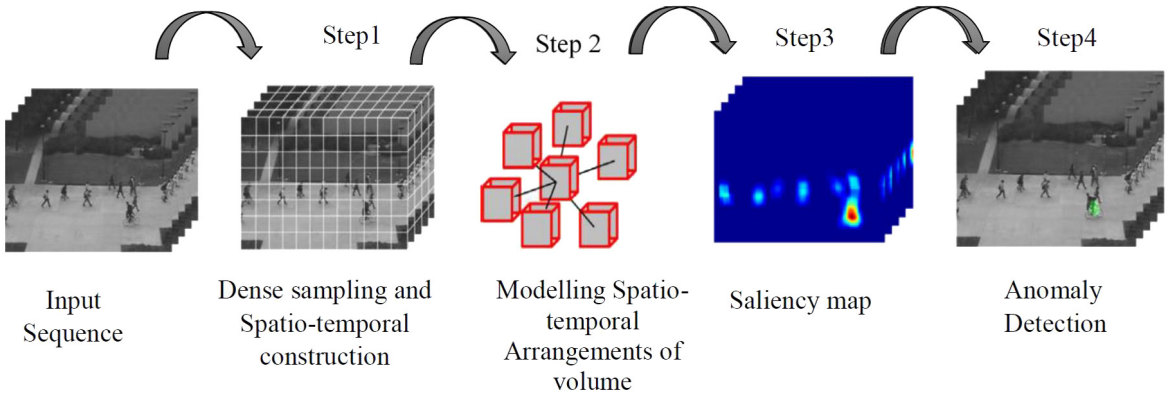

Fig. 7. An on-line, real-time learning method for detecting anomalies in videos using Spatio-temporal compositions (Roshtkhari and Levine, 2013).

process or as a reconstruction process (Roshtkhari and Levine, 2013). At different spatial and temporal scales, spatiotemporal compositions of the video volume are created and analogous spatiotemporal volumes are clustered. It helped to detect anomaly without background subtraction and tracking, as shown in Fig. 7. This framework performs well even in the concurrent occurrence of normal and doubtful events in the video. Zhu and Liu (2011b) used spatiotemporal interest point to give a compact representation to human actions and similar human actions patterns are grouped using Topic Hidden Markov Model (HMM: based on HMM and LDA) clustering technique. The presented THMM clustering approach offers smaller computational complexity. However, spatial and temporal arrangements of local features can provide fine details to develop compact video representation. Another work (Roshtkhari and Levine, 2013), segregated the video into three components: background, dominant activities and rare activities (bicyclist in walking pedestrians) and learn the video events for every pixel without supervision using closely built hierarchical codebook of spatiotemporal video volume (STVs), Fig. 8. The major advantage of this method is that it is adaptive to variable background conditions and illuminations and requires less number of initialization frames (200 frames compared to 6400 frames). But it fails to learn long time behaviors. Li et al. (2014b) utilised crowd dynamics as well as crowd appearance in the video jointly using mixture of dynamic textures (MDT) models. These texture models account for designing center-surrounded discriminant saliency detector and normal behavior model which produces spatial saliency score and temporal saliency score respectively at different scales. Here, multiscale scores strongly support capabilities of a conditional random field.

(CRF) that assures reliability of the abnormality decision globally. In addition to this, authors presented an anomaly detection dataset to address complex scenes of pedestrian crowds i.e. stochastic motion, complex occlusions, and entity interactions. It offers both frame-level and pixel-level ground truth, and also a protocol for anomaly detection algorithms authentication. Antic and Ommer (2011) presented a probabilistic model to detect abnormality in a video which consists irregular objects and sequence of activities. The task becomes more challenging when sufficient abnormal activities-training data is not available. In this model, while video parsing, a hypothesis layout is obtained that jointly describes the scene (foreground) instead of detecting individual hypothesis. This work is extended by the authors in Antic and Ommer (2015) in which their goal of video parsing is to determine a set of necessary normal spatiotemporal object hypotheses that mutually describe foreground of a video supporting normal training samples. Further, MAP inference in graphical method is explored for abnormality localization efficiently using it as a convex optimization problem (UCSD ped1, ped2) dataset. Abnormal event detection is formulated as a matching problem being more robust than statistic model-based methods, particularly for small training dataset. Cong et al. (2013) outlined region-based descriptor called "*Motion Context*", by incorporating both motion — the Multilayer Histogram of Optical Flow (MHOF) and appearance — Edge Orientation Histogram (EOH), evidence of the spatiotemporal video segments. Han et al. (2016) looks for latent action patterns among crowd activities using the concept of visual attention mechanism with spatiotemporal saliency-based representation and divide them in significant groups using affinity matrix based N-cut.

3.1.2.2. Sparse Representation based approaches. Sparse models provide unique discrimination ability to the descriptors which has made sparse representation a popular approach for abnormal behavior detection. Li et al. (2016) defined Histogram of Maximal Optical Flow Projection (HMOFP) descriptor and used it for dictionary optimization and computing 'l1' norm of sparse reconstruction coefficient (SRC) for crowd anomaly detection in the test frame. The combination of spatiotemporal features and its Laplacian Sparse representation is reported in Zhao et al. (2015) to identify a normalized combinational vector HNF (HOG+HOF), as shown in Fig. 9.

It is made more descriptive with minimized feature quantization error, by Laplacian sparse representation and maximum pooling. It is noted that the presented approach detects both global and local abnormal activities with 93% accuracy for UMN dataset (Detection of unusual crowd activity dataset, 0000). Whereas, for trajectory sparse reconstruction analysis (SRA) (Li et al., 2012) control point features of cubic B-spline curves are extracted from sequence of normal actions trajectory to build a normal dictionary set. In testing phase of abnormal event detection, sparse linear reconstruction coefficients and residuals help to decide normal or abnormal event. However, one of the limitations of this approach is that its performance is highly sensitive to the control point parameters of the trajectory. The recent work (Liu et al., 2017) observed the crowd motion using a double sparse representation with a dynamic dictionary updating mechanism that increases the size of dictionary dynamically for a small set of training samples. Dominant Set (DS) has emerged as a powerful technique to detect any abnormal behavior in an unsupervised framework (Alvar et al., 2014) that provide a locally adaptive boundary to represent the unknown data points sparsely. Experimentally, it is identified that Dominant Sets produce smallest overall error rate in comparison to other clustering algorithms i.e. KNN, a mixture of Gaussians, fuzzy k-means.

3.1.2.3. Miscellaneous. In a group activity, the shape of the objects deforms continuously. Co-occurrence statistics and dynamic Bayesian networks are used for group activity modeling which fails when objects are large in an activity. This problem can be avoided by treating interacting objects as point objects or landmarks. Kendall's shape theory proved to be valuable to represent shape deformations of a discrete landmark in Vaswani et al. (2005). A continuous-state Hidden Markov Model is defined for each landmark shape dynamics in an activity. Where, for each landmark, scaled Euclidean motion parameters, and corresponding shape, together, are used to describe the hidden state vector of the HMM. The conceptualization of the use of shape and its dynamics for abnormal activity detection makes it translation, in-plane rotation or sensor zoom invariant.

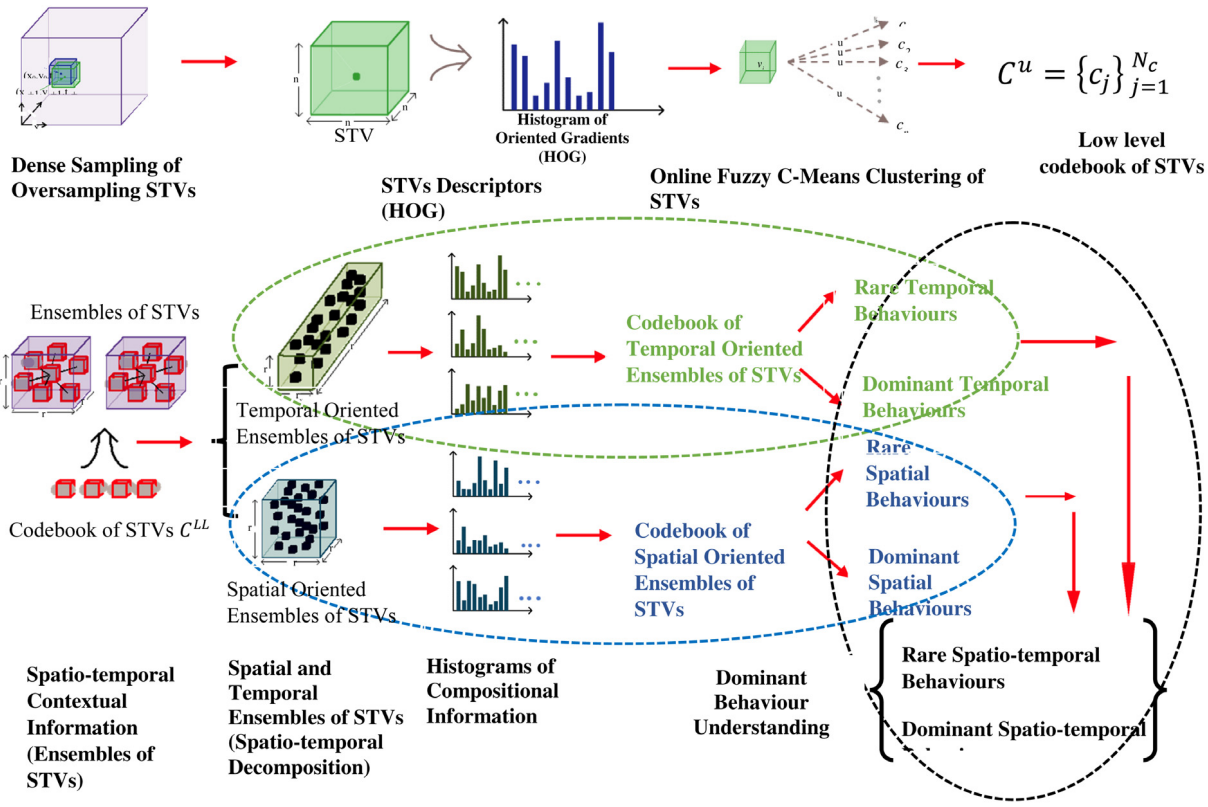

Fig. 8. Algorithm Overview: Online Dominant and Anomalous Behavior Detection in Videos (Roshtkhari and Levine, 2013).

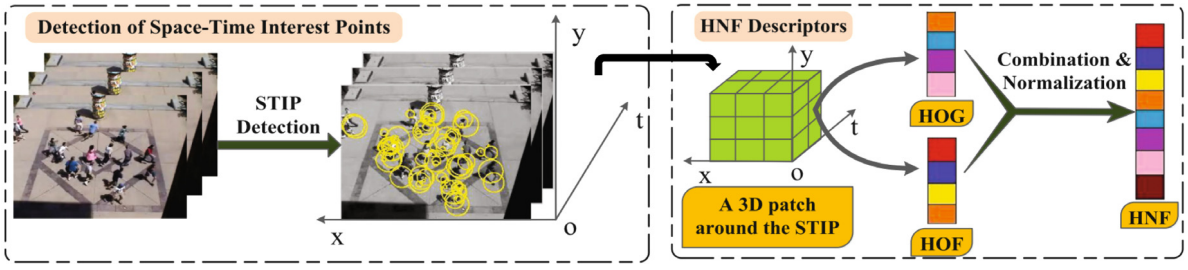

Fig. 9. Extraction of discriminative spatiotemporal features (Zhao et al., 2015).

In Hu et al. (2013) an unsupervised semiparametric statistical method is used to examine the (crowded) video with windows of adjustable shape and size for abnormal activity detection. The time complexity involved in the window scanning stage over the video is resolved with a fast two round scanning algorithm to advance the efficiency of the algorithm. People behavior can be defined as set of individual activities and group interactive behavior. And native motion details i.e. speed and direction, are sufficient to represent individual activity. However, for group interactive behavior interactive motion between neighbors information is required. In Cho and Kang (2014), a hybrid agent system is designed which comprises static and dynamic agents for efficient inspection of the corresponding individual and interactive behaviors in a crowded scene represented by a bag of words. Another unsupervised context-aware approach (Leach et al., 2014) targeted elusive anomalies in a behaviorally heterogeneous surveillance scene. Social context and scene context are used in conjunction for thorough activity analysis. Social contextual evidence supports anomaly propagation in a social network assuring better anomaly detection rate. Detection and localization of dense crowd in a scene is accomplished in Fagette et al. (2014) without any context-awareness in an unsupervised manner by multi-scale texture analysis. For which, firstly, three types of features are mined at different scales of the observation to design a high-dimensional multiscale feature vector for every pixel of the image. While binary classification (crowd or background) diffusion distance is preferred over traditional Euclidean distance to know the length and the density of the path between the data points. To reduce time and volume of computation complexity concept of coarse-graining is investigated using a quad-tree.

In Guo et al. (2016), authors follow pixel-wise approach, Weighted Quaternion Discrete Cosine Transform (QDCT) to find anomaly saliency map. The limitation of this approach is that QDCT cannot exploit information for long-term events. In Stephens and Bros (2016) specific human activities in a video are identified and distinguished by medium-term movement flow using streaklines model. Where each streakline is constructed by multiple optical flow vectors that keep a track of local movements in the scene. In Zhang et al. (2016), authors defined Locality Sensitive Hashing Filters (LSHF) which filter out abnormal activities by hashing regular activities into multiple feature buckets. To enhance the efficiency of anomaly detection in the video Particle Swarm Optimization (PSO) method is employed. Further, dynamic scene variations in the video, are updated to this framework by an online updating procedure.

Table 3
Evaluation of wearable devices and vision system (Rougier et al., 2011b) performance for fall detection.

	Wearable device	2D vision system	3D vision system
Lack of Significate movement	++	++	++
Lying position	+	+	++
Lying on the ground	–	+	++
Vertical speed	++	+	++
Impact shock	++	–	+
Body shape change	–	+	++

++: excellent performance.
+: satisfactory performance.

3.2. Three-dimensional AbHAR

Multi-dimensional data processing always produce more accurate and precise results in comparison to one-dimensional data processing. This fact has always attracted researchers towards 3D, 4D data. In recent years, due to the gigantic popularity of Microsoft depth sensors, the research work on abnormal Human Activity using depth information has proliferated. A number of new datasets (Fothergill et al., 2012; Escalera et al., 2014, 2013) have provided researchers enormous opportunities to create novel designs and test them on a bigger set of sequences. We can generate depth details of the image and also skeletal view with each joint position in three coordinate system. Some interesting applications are developed using depth and skeleton representation of an image.

3.2.1. Depth based AbHAR

Depth images not only simplifies and fasten up the low-level image processing but also delivers better processing outcomes in terms of background subtraction, object motion detection, and localization. In the following series various popular depth based approaches are discussed and summarized in Table 4 highlighting the key contributions and limitations of these methodologies with year of publication. From where it is identified that depth-silhouette based statistics such as height to width ratio, centroid, and silhouette shape deformations, are used very commonly to extract features of the person under motion. Human motion and shape variation features (Rougier et al., 2011b; Nguyen et al., 2016) can handle realistic challenges such as occlusion, different viewpoints etc.

Vision-based AAL solutions (Chaaraoui et al., 2012) are the best examples of depth based anomaly detection which assists the elderly while performing their daily living activities using depth information. A significant amount of work (Tran et al., 2014; Rashidi and Mihailidis, 2013) has been done to produce depth based AAL solutions. Jyothilakshmi et al. (2016) extended the Ambient concept Assistance of the patient through the application of Smart Ward System (SWS) using Kinect sensor camera. In existing patient assistance systems in hospitals, the patient is provided with a remote control to fine-tune the bed and to call a nurse for assistance. Many hospitals are not equipped with this automation system due to the expenses involved, and conventional gesture recognition systems are wearable sensors' dependent. The concept of Smart Ward System empowers the hospital workforce to emphasize majorly on patients with the facility of collection of data at the bedside and remove the chances of duplication and double-handling, i.e. security issues over data storage in the cloud. In Uddina et al. (2011); Triantafyllou et al. (2016), HMM model is used with depth features to develop a real-time solution which generates a warning alarm for the person walking in a room which can be extended to various home activities in the drawing room, living room, and kitchen. ICA helps to extract local features of human gait depth silhouette rather than PCA. Whereas in Uddin et al. (2014) Local Directional Patterns (LDPs) are used to extract local features from depth silhouette producing superior gait recognition rate using HMM model. The edge responses extracted via LDP are calculated in eight directions using Kirsch edge masks. To handle the problem of dehydration, RGB-D data is used (Tham et al., 2014) to monitor drinking activities at home with the help of Dynamic Time Warping (DTW) concept.

Fall (Mohamed et al., 2014) is one of the most frequently identified unusual events. Various methodologies have been presented for fall detection, highlighting this major health concern. Few studies presented that falls are the foremost reason for elder persons', disabled people or overweight/obese peoples' hospitalization. An extensive survey on fall detection is made in Yu (2008) which focused on identifying principles and approaches to existing work and literature. It classified the fall detection technologies into three categories: *wearable device, ambient device, and vision-based approaches,* as shown in Fig. 10. It is notified most wearable devices are cost-effective and easy to use, but they may not be very appealing to the wearers, in addition to this wearable device based approach is prone to the high rate of false alarm. Ambient device-based approaches commonly use pressure, vibration, acoustic sensors, and stabile-meters which need to be positioned in a permanent space to gather facts about the person in their operative range. Thus these sensors have limited range of usage and are noise sensitive due to environmental factors. The vision-based approach has an overhead over others that it is not intrusive to the person under monitoring in comparison to wearable devices and ambiance devices. They can be divided into three sub-categories: *Inactivity Analysis (IA), Shape Change Analysis (SCA), and Head Motion Analysis (HMA)*. Table 3gives an overall summary of wearable devices and 2D & 3D vision-based systems for fall detection. It also shows the circumstances where wearable and 2D & 3D vision-based system performances, are good and bad. The 3D vision system can detect fall with maximum accuracy between two with any combination of fall features such as lack of significant movement, lying position, lying on the ground, vertical speed, impact shock and body shape change.

3.2.1.1. Depth based fall detection. The concept of inactive period strengthens the severity of fall. Lying on the ground being inactive for longer duration is a consequence of strong fall. Fall detection cannot be made from one instance information, but discriminative features need to be analyzed for the entire duration of fall and also after the fall. The confirmation of inactivity is highly context dependent. The exact location of the person, time and duration of inactivity collectively gives wise decision such as staying in bed for long duration is not an alarming event but staying long on the floor after a fall will lead to an alarm. Hence contextual information helps to reduce the false alarming rate. Therefore, Jansen and Deklerck (2006) learnt about contextual details of the fall by quantifying the area of the body on the floor and 3D orientation of fall to understand the inactive duration of a person during and after fall. In daily activities, human body possesses various quick movements which may lead to false large motion identification. Therefore in Nguyen et al. (2016) center of mass of human 2D silhouette is quantified to observe overall motion (magnitude and orientation) of the object, with the help of image moments, along with human shape feature. It reduces the impact of sudden movements on fall detection. In addition to this, MHI (motion history image) provides exact location and trajectory of motion in the video sequence. However, 2D silhouette extraction approach needs to be improved to reduce false detection rate.

Human body posture is also used as a key element for fall detection by set of researchers (Ma et al., 2014; Akagunduz et al., 2016; Rougier et al., 2011b; Yu et al., 2013; Yao et al., 2017; Rougier et al., 2011a). Yu et al. (2013) represented human body posture with the centroid

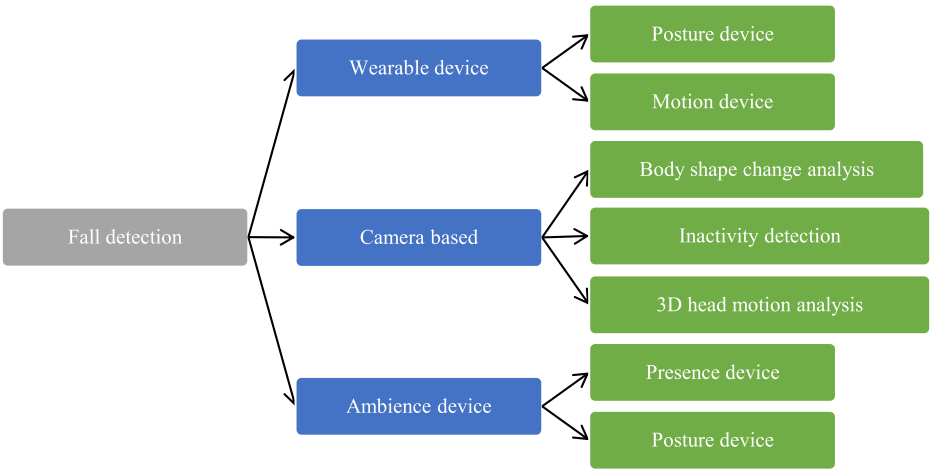

Fig. 10. An ordered structure of fall detection methods for elderly and patients (Yu, 2008).

Table 4
Feature extraction and representation techniques for three-dimensional-depth based fall detection techniques.

Method	Year	Technique involved	Performance	Dataset	Challenges	limitations
Yu et al.	2013	Online One-Class Support Vector Machine (OCSVM)	Accuracy-100%	The self-recorded dataset in a real home environment of elderly people	Handles occlusion well	Tested for a single dataset
Goudelis et al.	2015	History Triple features	Accuracy — 100% (both)	Le2i Fall Detection dataset, UR fall dataset	handles common visual distortions, camera placement variations and noise	Does not support real-time fall recognition
Nguyen et al.	2016	body width-to-height ratio, the angle between body and horizontal plane	Precision — 93.25%	Lie2 Fall detection dataset	Handles shadow, light reflection, complex, textured background, and view variations	suitable foreground segmentation technique required
Jansen and Deklerck	2006	context model defined after fall using area of the body and 3D orientation of fall	–	Synthetically generated fall sequences	–	Human interventions are required
Rougier et al.	2011b	full Procrustes distance and mean matching cost as shape deformation measures	Accuracy-99.6%	Self-recorded dataset (Fig. 11)	occlusion, use of the object, putting off clothes, different viewpoints	–
Yang et al.	2016	Position and orientation analysis	–	The self-recorded dataset in a real home environment of elderly people	Detects fall accurately when the fall orientation is aligned with the optical axis of the vision sensor	Require adaptive thresholds for fall detection
Panahi and Ghods	2018	The distance of the person's centroid to the floor	Sensitivity-100% Specificity –97.5%	The self-recorded dataset in the real home environment	Videos are recorded in different rooms and decorations, with different people, and in different lighting conditions.	the application of the Kinect v2 strongly depends on the distance to the target and the incidence angle of the sunlight
Yao et al.	2017	Person's torso angle and centroid height	97.5%	Self-recorded Kinect based dataset	High discriminating power between Fall- and fall-like events	Depends on threshold values for fall detection
Rougier et al.	2011a	human centroid height relative to the ground and 3D person velocity	98.7%	Self-recorded Kinect based dataset	Occlusion	suitable foreground segmentation technique required
Ma et al.	2014	Bag of Curvature Scale space (BoCSS)	86.83%, 97.2%	SDUFall Dataset, MultiCam dataset	–	Time cost is high
Akagunduz et al.	2016	Silhouette Orientation volume (SOV)	91.89%, 100%, 89.63%	SDUFall Dataset. Weizmann Dataset, six-class action recognition dataset	Scale invariant	–

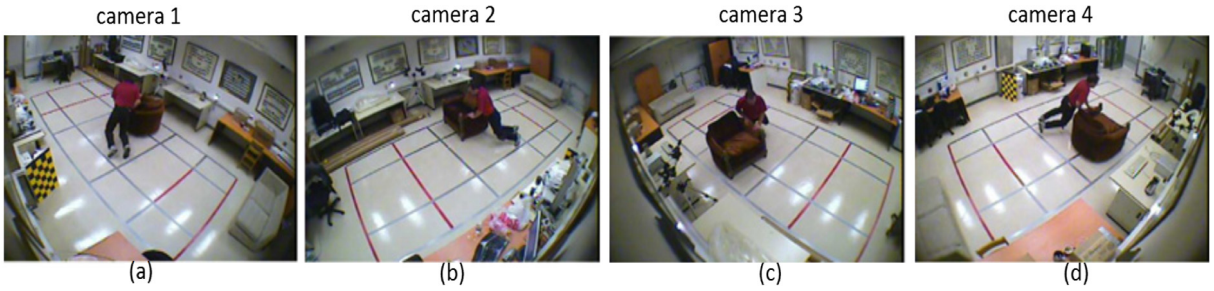

Fig. 11. A same person fall activity viewed from different viewpoints (Rougier et al., 2011b).

Fig. 12. Other examples considered in Rougier et al. (2011b) (a) Forward fall (b) sitting down (c) crouching down (d)–(e),(g) fall (loss of balance) (f) putting off a coat (h) fall from sofa (i) housekeeping (j) backward fall (k) occlusion (l) carrying a box.

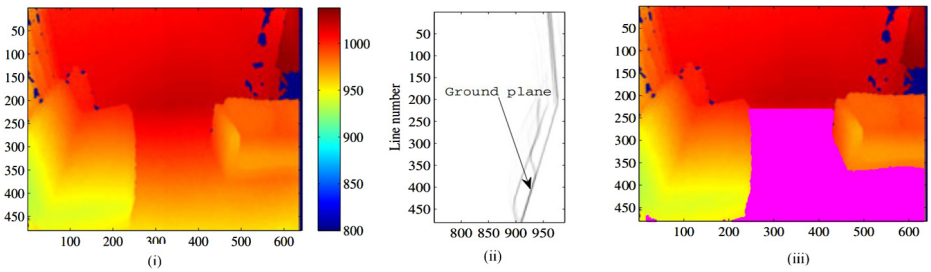

Fig. 13a. Ground Plane detection (i) depth image (ii) V-Disparity image (iii) ground plane segmentation (in magenta) (Rougier et al., 2011a). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

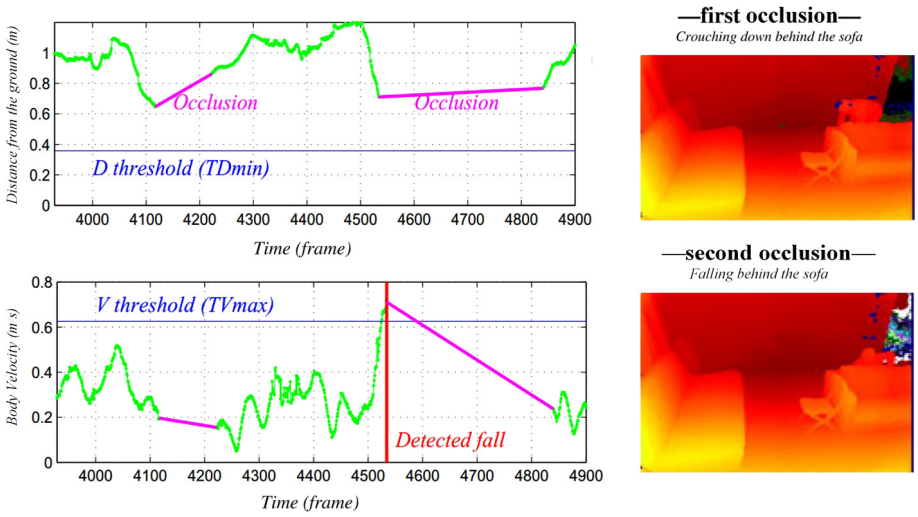

Fig. 13b. Distance from the ground and body velocity curves (left) obtained for occluded events (right). The fall is correctly detected with a high body velocity (Rougier et al., 2011a).

position of the body, ellipse feature and shape structure feature. They have defined two rules to reduce false alarm rate, which is as follows:- *First rule*: area ratio (AR) between the area of MEI frames and area of the current frame's foreground region compute the magnitude of motion. AR will be larger for fall activity than walking, sitting or bending. Fall will not be reported for a small value of AR, even though the system detects abnormal postures. *Second rule*: a fall is declared only if any abnormal posture occurs for a longer period than a threshold; it reduces false alarms (i.e. occasional bends, tying shoes) to an acceptable range. In spite of the excellent fall detection accuracy, the method demands human interventions for manual segmentation and video clip selection. Another approach (Rougier et al., 2011b) measured human shape deformation during fall using full Procrustes distance and mean matching cost as deformation measures. Realistic scenarios like occlusion, use of the object, putting off clothes, different viewpoints are addressed by authors in their dataset as shown in Figs. 11 and 12**.** Yao et al., (2017) introduced Human Torso Motion Model (HTMM) which can discriminate fall and fall-like activities such as bending and crouching down with 97.5% accuracy by observing changing rates of torso angle and the centroid height. Since the existing RGB-D action datasets i.e. CAD-60/120 do not provide fall sequences, ADL and fall sequences are recorded for experimentation. However, the method is dependent on threshold values which used with hit and trial approach to optimize the result, which needs to be identified every time for a new dataset. Rougier et al. (2011a) computed human centroid height relative to the ground and 3D person velocity. 3D person velocity helps the system to make fine discrimination between crouching down behind the sofa from fall behind the sofa — look alike cases as shown in Fig. 13b**.** Here, 3D velocity is preferred over 2D velocity of a person during fall because 2D velocity is generally very high near the camera for normal walking activity resulting misclassification between a fall and a walk. Detection of the ground plane for human centroid height computation relative to ground in depth images, accomplished using V Disparity histogram, is shown in Fig. 13a. It is observed that height (Gasparrini et al., 2014) and height velocity (Yang et al., 2015) based approaches fail to distinguish fall and fall-like actions and whereas bounding box width to height ratio based (Nguyen et al., 2016; Rougier et al., 2011a) and HTMM (Yao et al., 2017) based fall detection model has **the** higher discrimination power of fall and fall-like actions.

Ma et al. (2014) represented the actions by a bag of words model (BoCSS) using distinctive Curvature Scale Space (CSS) features of depth silhouette for fall detection, whereas Akagunduz et al. (2016) integrated orientation scale space (OSS) and morphological scale space of a curve to form robust Silhouette Orientation volume (SOV) global scale invariant descriptor to represent actions. However, these approaches have high computational cost.

Some researchers came up with fusion of sensor (accelerometer, floor sensor) and visual depth data to develop an improved human fall detection system which has improved the performance of the fall detection. Toreyin et al. (2006) integrated sound impact of falling person with height to width ratio of the bounding box on a person under falling condition to discriminate a fall from a normal sitting in the floor action. Zerrouki et al. (Zerrouki et al., 2016) used the concept of Univariate Statistical monitoring method Exponentially Weighted Moving Average (EWMA) control scheme to detect potential fall integrating accelerometric data and depth data with low computational cost. Though such fusions produce impressive outcome but this detection is dependent on sensor and its periphery which may not be in the comfort zone of the user. Therefore, the purpose of making the fall detection system non-intrusive to the user is defeated.

3.2.2. Skeleton based AbHAR

Skeleton representation of human body provides incisive details of the human posture in compact form which has definitely resolved the problem of need of effective segmentation technique to extract 2D silhouettes and simplifies the height centroid computation (Rougier et al., 2011a) from depth silhouettes. This has encouraged researchers to develop real-time applications using skeleton modality making the computation process faster, simpler and more effective. Nar et al. (2016) designed an effective real-time ATM intelligent monitoring system to recognize abnormal postures prevailing stronger security in the ATM i.e. fiddling with the camera, aggressive posture, and peeping. The work used angles between different bones as useful features to compute the optimum value of weights for computing probability of the current pose of the person under surveillance being abnormal. The computation of angle between joints (x, y, z) becomes quite simple and fast and more accurate with 3D skeleton coordinates. Hendryli and Fanany (2016) addressed the issue of automatic detection of abnormal activities of students in examination hall that generates warning to exam proctors if any doubtful activity is detected (Cheating activity). For this purpose, MCMCLDA (Multi-class Markov Chain Latent Dirichlet Allocation) framework is designed that access arm joints and head location as interest points directly from skeleton representation without considering irrelevant ones resulting better accuracy and higher computational speed than Harris3D detector. Nowadays, health professionals and researchers have also shifted their attention to Human gait analysis (Aggarwal and Vishwakarma, 2016). Because human gait not only helps to know different traits of a person but also variation in the human

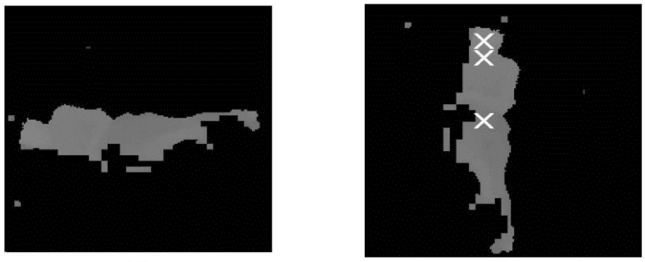

Fig. 14. An example of rotation. (a) Before rotation. (b) After rotation, the head, neck and hip center joints can be extracted correctly (Bian et al., 2012c).

gait from normal frame may be correlated with neurological diseases. Chaaraoui et al. (2015) developed generic machine learning framework (Bag- of-Key-Poses) using joint motion history feature i.e. 3D location of skeletal joints and motion cues. To handle complex behaviors both low and high-level multi-scale motion cues are extracted in Paiement et al. (2014). However, skeleton data acquired with a Kinect sensor likely to suffer from relatively large amount of noise, and also contain outliers, especially in case of partial occlusion. Therefore, the work incorporates diffusion maps to filter the outliers. Jalal et al. (2014), tried to develop a design continuous surveillance and daily activity recognition in indoor environments (i.e., smart homes, smart office and smart hospitals) turning the space into a smart living space. It also failed to handle complex activities and partial occlusion of the body while generating skeleton joints and resulting in noise.

3.2.2.1. Skeletal based Fall Detection. Research in abnormal activity recognition is receiving dynamically increasing attention, encouraged by the speedy growth of information (Khan and Hoey, 2017), intelligent video systems, and communication technologies. Medical consequences of a fall mainly depend on the response and rescue time. Hence, an accurate automatic fall detector (Bian et al., 2012a) is an essential element for an older adult living to expedite and improve the medical care provided to this group. Skeletal representation of a person is proving its strength by enhancing the performances of fall detection systems and many other abnormal activities in daily routine. Table 5 outlines various skeleton based AbHAR approaches. It is visible from here that trajectory of joints (Rougier et al., 2006; Bian et al., 2012b; Nizam et al., 2016) and Joint Motion History (JMH) (Chaaraoui et al., 2015) based action description is simple and very effective having high temporal efficiency and appreciable view and illumination invariance property for skeleton-based abnormal human action detection.

However, the distance between silhouette center and the floor (Rougier et al., 2011a) or shape deformation concept based fall detection work is not able to discriminate the initiative action and fall accident well i.e. fall in bed and fall in the floor without defining normal inactivity zones, a person is sleeping on the sofa or bed and falls down to the floor. 3D human skeleton joins distance from the floor and joins hitting velocity, joint position and its height from the ground (Bian et al., 2012b; Nizam et al., 2016) collectively elicit robust results by discriminating a fall from slowly lying down on the floor and other similar cases. While falling, human's body orientation changes dramatically which leads to poor tracking of joints. Therefore in Bian et al. (2012c) the author initially corrected the trunk orientation (from hip point to neck) of the person before applying a fast Randomized Decision Forest (RDF) algorithm for human skeleton extraction which has improved the accuracy of fall detection rate as shown in Fig. 14. The presented work is able to detect minor fall like falling from the sofa when our half body (legs) is still on the sofa by simultaneously tracking head, which silhouette center-based approach fails to identify. A view independent statistical method (Zhang et al., 2012) takes a decision based on the information how human behave in last few frames after falling. It defined a feature set $f = [f_1, f_2, f_3, f_4, f_5]$, where f_1 is duration of fall, f_2 is total head drop, f_3 is the maximum speed of fall, f_4 is the smallest head height, f_5 is the fraction of frames where head has a smaller height than in the previous frame. The five features are combined using the Bayesian network. But few false alarms are received for a person lying on the floor and a person jumping on the bed. Diraco et al. (2010) used distance of 3D centroid from floor plane as threshold to confirm high performance in terms of consistency and competence on a large real dataset which uses Bayesian segmentation to detect moving regions. It also introduced 3D Geodesic Distance Map-Based Posture Recognition (3dGDMPR) system prototype, Fig. 15. Occluded leg is used as a test case and overlapping region postures are detected and interpolated as mentioned in Fig. 16.

3.3. Deep features based action description

It is observed that over the passage of time the concept of feature engineering is evolving from 2D features to 3D features in order to improve the action representation. However, complexity involved in designing handcrafted features is very high which is one of the key reasons to shift the features from shallow region to deep that has brought the practical applicability of the action recognition algorithms at another level of excellence by empowering the knowledge of deep learning to the recognition systems. Though the concept of deep learning and architectures (Koohzadi and Charkari, 2017) was existing since 1980s but could not perform up to the mark due to lack of sufficient datasets and computational resources. In 1998, LeNet (Lecun et al., 1998) came up as the first real-world successful application of CNN for handwritten digit recognition. However, in successive years, various deeper architectures have been reported in Russakovsky et al. (2014) and are being used in different application areas such as computer vision (Herath et al., 2017; Paul et al., 2013; Taylor et al., 2010), speech recognition (Vesperini et al., 2018), brain–computer interaction (Cecotti and Graser, 2010) and natural language processing (Yin et al., 2017) with the availability of large datasets and hardware resources. Deep models learn a hierarchy of features by constructing high-level features from low-level ones. CNN is a type of deep model which is made up of neurons and learnable weights and biases which were initially applied on 2D images for visual object segmentation (Iannizzotto et al., 2005) and recognition (Wang et al., 2018; Uddin et al., 2017) tasks. And later, many researchers experimented CNN with videos by considering video frames as still images to recognize action in each frame. However, this approach was limited to learn only spatial information. Some authors (Ji et al., 2013) tried to incorporate temporal information by expanding the 2D CNN to 3D CNN. 3D CNN applies 3D convolution in CNN convolution layers by using 3D kernel to multiple contiguous frames stacked together to encode the motion information with spatial one. In subsequent years, the work is further extended by bringing the concept multi-stream CNNs (Simonyan and Zisserman, 2014) based action recognition, which has fortified feature description of an action to a higher level by making the deep recognition system to analyze not just raw images at a time but multiple set of inputs such as: RGB image, optical flow (Simonyan and Zisserman, 2014), dynamic images (Jing et al., 2017), and depth images (Liua et al., 2016). On the other hand, long short term memory (LSTM) (Liu et al., 2018) has emerged as one of the most popular unsupervised models that learns temporal arrangements of frames and predict time series data.

3.3.1. Single person based abnormal human action recognition

As far as application of deep architectures in the area of abnormal human action recognition for AAL and smart homes, is concerned, a limited work has been carried out yet. It is the need of an hour to apply deeper architectures to develop stronger algorithms (Zhang et al., 2017) for single person specific abnormal actions recognition applications i.e. AAL, smart homes. The generalized deep frameworks for human action recognition are summarized in Table 3 and single person based abnormal human action recognition work is discussed below.

Table 5
Feature extraction and representation techniques for three-dimensional-skeleton based abnormal human activity recognition (AbHAR).

Feature representation for depth based anomaly detection	Key Contribution	Limitations	Year	Ref.
Joint Motion History Feature (JMH) + Bag Of Key Poses	high temporal efficiency	Multi-scale feature extraction technique is required to handle complex human activities	2015	Chaaraoui et al. (2015)
The angle between different bones, the 3D trajectory of head motion, Human joints height, joint positions and falling velocity	The velocity of joints a stronger classifier than distance variations.	Adaptive thresholds need to be used to handle individual differences	2014 2012 2016 2006	Tran et al. (2014) Nar et al. (2016) and Rougier et al. (2006) Bian et al. (2012b) and Nizam et al. (2016)
Dynamic Time Warping	Illumination invariant	Manual segmentation of drinking video needs to be replaced by automatic segmentation procedure	2014	Tham et al. (2014)
Multi-class Markov Chain LDA Based on MODEC (Multimodal Decomposable Models) Features	MODEC interest point detection is faster than Harris3D	Incapable of distinguishing elbow and wrist when the object wears clothes with indistinguishable colors	2016	Hendryli and Fanany (2016)
Trunk orientation correction and Randomized Decision Forest (RDF)	Trunk orientation correction improves the accuracy of fall detection.	Time involved in trunk correction per frame not discussed	2012	Bian et al. (2012c)
Depth comparison feature (DCF) a+ RDF	DCF is extracted pixel wise	Computation cost is high for large datasets	2018	Abobakr et al. (2017)
Bayesian Network with Statistical parameters	Viewpoint invariant and user independent		2012 2010	Zhang et al. (2012) and Diraco et al. (2010)
PCA (20 skeleton joints) +HMM	human posture sequence modeling	Non-availability of the large-sized dataset	2013	Dai et al. (2013)
3D trajectory of head joints	Illumination invariant of the lights and perform well in dark room	Head occlusion case not discussed	2015	Bian et al. (2015)

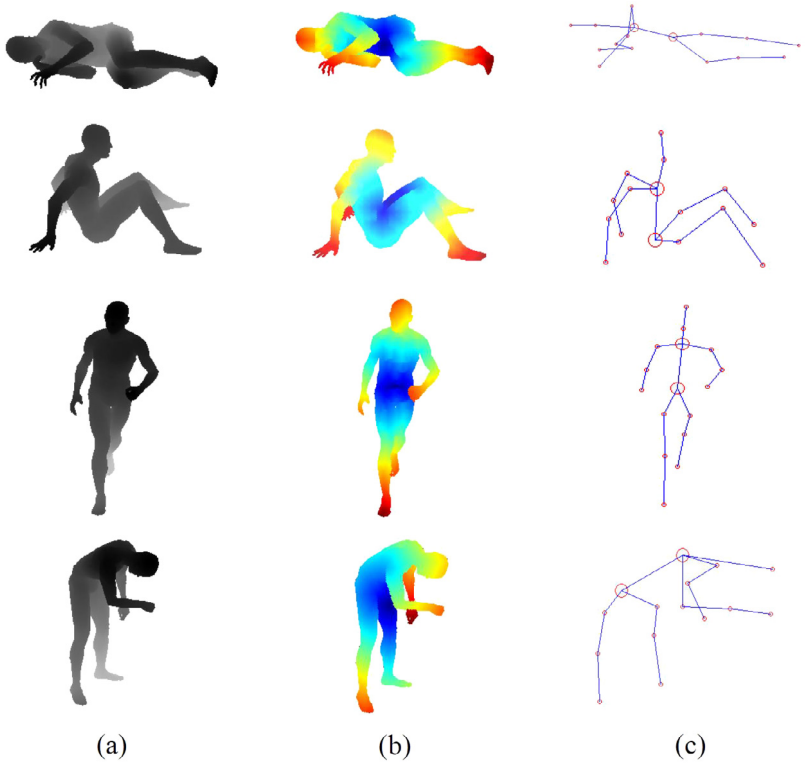

Fig. 15. Four main postures examined in Diraco et al. (2010): lie, sit, stand and bend. (a) Depth map (b) Geodesic distance map and (c) skeleton. Upper and lower nodes of the skeleton Reeb Graph are encircled with red circle.

The common signs of dementia (Alzheimer, Parkinson's disease) can be acknowledged by behavioral variations such as disturbance in sleep, difficulty in walking and fail to complete tasks. Such variations can be identified in early stages by monitoring the elderly person in smart homes which may prove to be more beneficial than medical diagnosis. However, availability of the real world data of dementia patients is a big challenge for non-medical diagnosis of the patient. Therefore, Arifoglu and Bouchachia (2017) introduced an approach to generate synthetic data to represent behavior of dementia patients and addressed the problem of abnormal behavior detection for elderly

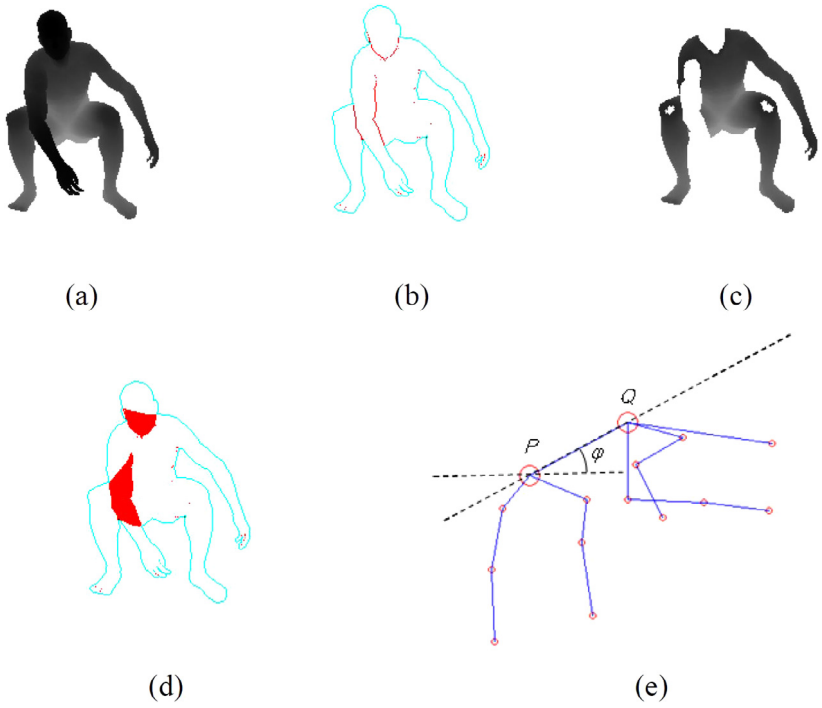

Fig. 16. Detection and interpolation of overlapped regions (a) the arm occludes a leg (b) Laplacian operator detect occluded bounds (c) and level sets of the height function (depth map) detect inbound region (d) interpolation is performed to obtain a connected mesh (e) body spine orientation is approximated by the orientation φ of the PQ segment in the body skeleton (Diraco et al., 2010).

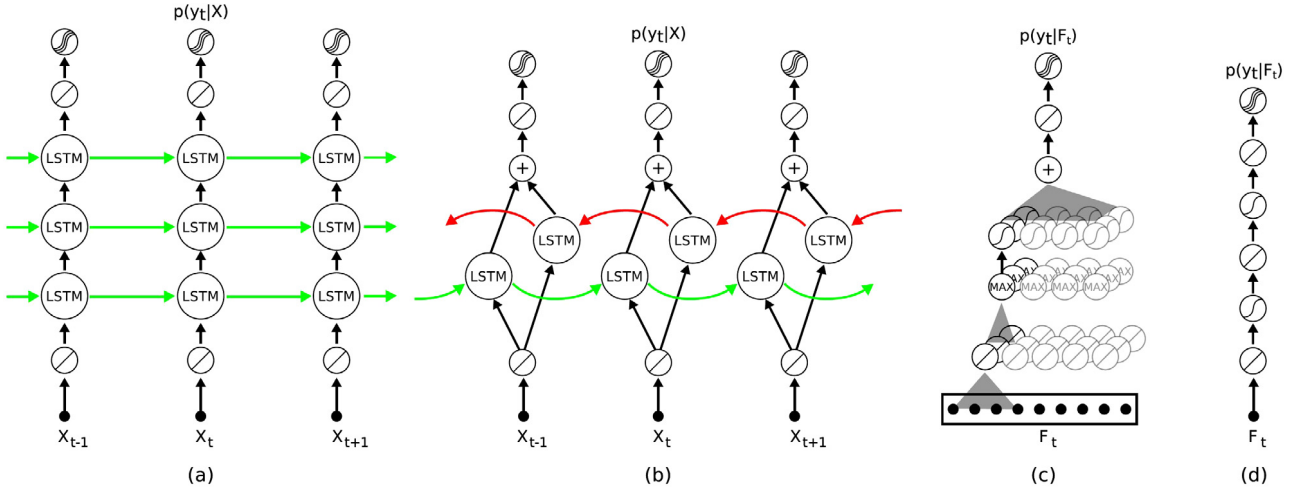

Fig. 17. (a) LSTM network hidden layers containing LSTM cells and a final Softmax layer at the top. (b) bi-directional LSTM network with two parallel tracks in both future direction (green) and to the past (red). (c) Convolutional networks that contain layers of convolutions and max-pooling, followed by fully-connected layers and a Softmax group. (d) Fully connected feed-forward network with hidden (ReLU) layers (Hammerla et al., 2016). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

people with dementia using three variants of Recurrent Neural Networks (RNNs)- Vanilla RNNs (VRNN), Long Short-Term RNNs (LSTM) and Gated Recurrent Unit RNNs (GRU), see Fig. 17. Recently, Park et al. (2018) analyzed the sequence of actions recorded in ambient intelligent environment i.e. smart home and city with multiple sensors by introducing the concept of Residual-Recurrent Neural Network (Residual-RNN). The extensive experiments assert that the proposed model (in Fig. 18) performs better than LSTM and Gated Recurrent Units (GRU) in respect of recognition accuracy. However, GRU performs slightly better in terms of computational speed. In another work by Hammerla et al. (2016), LSTM, bi-directional LSTM, and convolutional networks are used to detect abnormal actions of the patient under surveillance using movement data acquired from wearable sensors.

Considering the fact that deep models are outperforming existing state-of-arts (Zhang et al., 2018a, b; Hou et al., 2018) in almost all the fields being used such as face recognition, handwriting detection (Tolosana et al., 2018), audio signal processing, a variety of plans are discovered that use pre-trained, fully-connected networks for automatic assessment of Parkinson's Disease (Hammerla and Plotz, 2015), replace emission models i.e. GMM, random forest (RF) in Hidden Markov Models (HMMs) with Deep Neural Networks (DNN) (Zhang et al., 2015; Alsheikh et al., 2016) ; and design a mobile audio sensing

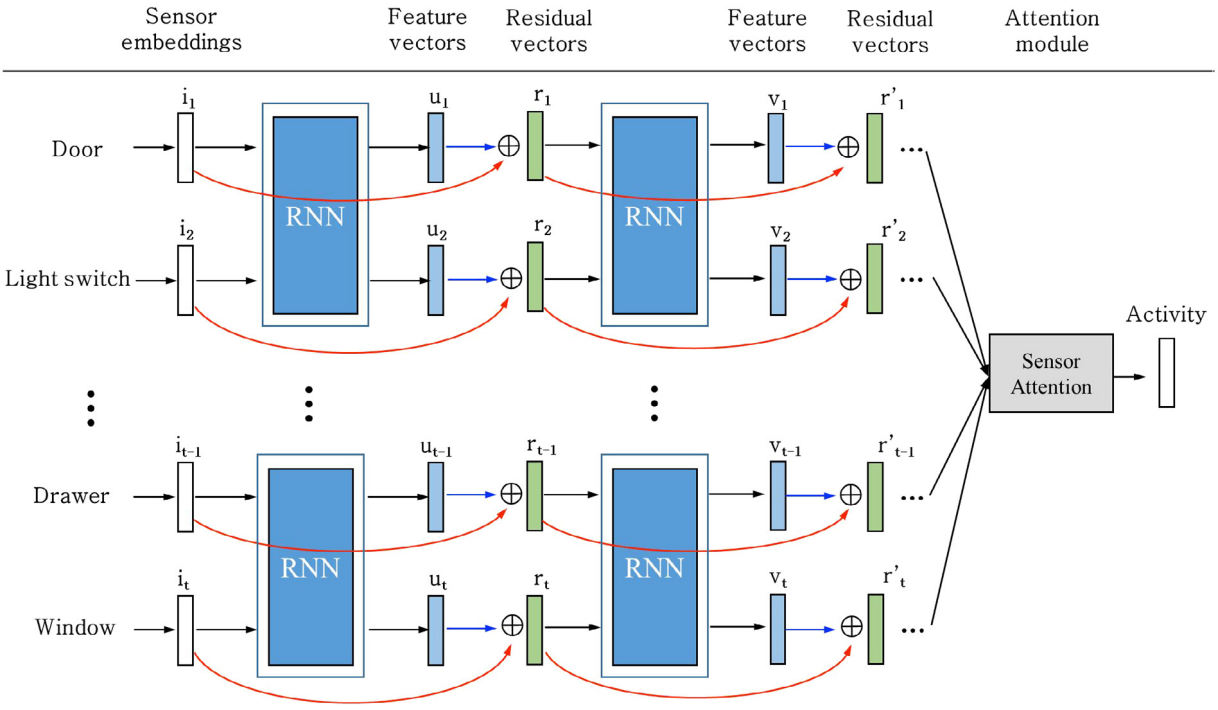

Fig. 18. Residual-RNN structure for activity recognition in smart homes (Park et al., 2018).

framework using DNNs (Lane et al., 2015). These works (Hammerla and Plotz, 2015; Zhang et al., 2015; Alsheikh et al., 2016) are restricted to the accelerometer and wearable sensor data, that is very scarcely available and labeled. It limits the application of the methods in real-world scenarios where vision sensors are used profoundly.

3.3.2. *Multiple persons based abnormal event detection*

Evolution of abnormal event detection approaches in crowded scenes from shallow to deep is reported in a recent survey (Tripathi et al., 2018) which has emphasized on four attributes of the crowd to be considered for crowd analysis:-crowd counting, crowd motion detection, crowd tracking, and crowd behavior understanding. Crowd analysis is one of the demanding issues in public domain. And deep learning models i.e. variants of CNN, LSTM, autoencoders, RNN are providing improved solutions (Sindagi and Patel, 2018) compared with hand-crafted features for real-life challenges in crowded scenes like occlusion, cluttered background etc.

In the work (Vignesh et al., 2017) the property of LSTM network to learn the long-term dependence between the frames is fused with CNN features to recognize the unusual event in videos. Here, the LSTM network possess two layers — 1×1024 LSTM layer followed by 1×512 bi-directional layer. This set of arrangement of layers overcomes the limitation of standard recurrent neural networks for inflexible size of input data and also that the forthcoming input cannot be obtained from the current state. Whereas, bi-directional LSTM has forward and a backward LSTMs operational in reverse time that combine the significant features of past and future. Though, the approach produces impressive results for ARENA dataset but the experimental work deals with only one type of activity. In Medel and Savakis (2016), author introduced the concept of Conv-LSTM to model and predict video sequences. Conv-LSTM layers are memory insensitive which makes training testing phase slow. In real-world scenario, the probability of abnormal event occurrence is lesser than a normal event. 3D ConvNet fails in the absence of abnormal events and loosely supervised datasets, in spite of being able to discriminate normal and abnormal events. Chong and Tay (2017) proposed a spatiotemporal end-to-end model for anomaly detection in videos or crowded scene. Which performs well even in the absence of abnormal events in video samples. It has a CNN based spatial feature extractor and a temporal encoder–decoder which collectively learns the temporal patterns of the input volume. To understand the crowd scenes Shao et al. (2016) developed a spatiotemporal CNN named as Slicing SCC (S-CNN) which decompose the 3D volume (x, y, t) input video into 2D spatial (x, y) and 3D temporal slices (x, y), (x, t), and (y, t) by exploiting selectiveness of spatial filters. The semantic feature maps claims that such kind of slicing helps to discard background clutter of scene and select only the people under action. In addition to this, in a recent work (Li and Chuah, 2018) defined a robust and efficient human activity recognition framework (ReHAR) that combined optical flow and CNN based image feature extractor with global average pooling layer which is fed to stack of two LSTM networks to predict the group activities of Basketball dataset and UCF sports action dataset. Global average pooling layer and stack of two LSTM structures helps to perform the task faster supporting real time action recognition.

The development of generalized deep learning architectures for human action recognition is summarized in Table 6. Which indicates that various modalities-RGB, depth, skeleton, optical flow, MSDI, and dynamic images are integrated to enhance the performance of the system by the researchers. Each set of representation has its benefit over another such as dynamic images (Jing et al., 2017) and optical flow (Hana et al., 2018) both provide motion representation, however, from the aspect of time of computation and storage capacity dynamic images offer a more compact representation of motion in a video sequence than optical flow. Set of researchers (Liu et al., 2018; Amir et al., 2016; Song et al., 2016; Ding et al., 2016) have exploited view-invariant and succinct property of skeletons for deep action recognition models and observed impressive performance. However, the real-time applicability of such algorithms is still an important issue from the point of single person surveillance. As listed in Table 7, it can be seen that only a little work is carried out on the problem of single person abnormal human action recognition using the benefits of deep architectures, which needs to be focused seriously, considering the need of better healthcare services, smart homes, and AAL. The work which is listed is also based on wearable sensor data. Sometimes wearing the sensors is not a pleasant experience and if the person forgets to

Table 6
Generalized deep approaches for human action recognition.

Method	Year	Technique		Dataset (Accuracy)	Modality
Simonyan and Zisserman	2014	Two stream CNN	Spatial and temporal streams	UCF-101(88%), HMDB51 (59.4%)	RGB + optical flow
Diba and Gool	2016		Appearance and motion streams	UCF101 (90.2%)	RGB
Hana et al.	2018		Spatial and temporal streams	UCF101 (95.1%), KTH (93.1%)	RGB + optical flow
Jing et al.	2017	Three stream CNN	Appearance, short-term temporal and long-term temporal streams	UCF-101(88.6%), HMDB51 (57.9%)	RGB + optical flow + Dynamic images
Wang et al.	2017a		spatial, local temporal and global temporal streams	UCF-101(93.4%), HMDB51 (68.3%), Hollywood2 (74.6%), YouTube (83.8%)	single frame + optical flow + Motion Stacked Difference Image (MSDI)
Liua et al.	2016	$3D^2CNN$		MSR Action-3D (98.14%) and UTKinect-Action3D (95.5%)	Depth image + skeleton
Ji et al.	2013	3D CNN		KTH (90.2%)	RGB
Ijjina and Chalavadi	2017	MEI and MHI based Temporal Templates for RGBD 2D ConvNet features		MIVIA action (93.37%), NATOPS gesture (86.52%), SBU Kinect interaction (90.98%), and Weizmann dataset (100%)	RGB + Depth
Bilen et al.	2016	Dynamic Image Network (2D CNN)		HMDB51 (65.9%) and UCF101 (91.5%)	RGB
Wang et al.	2015a	Trajectory-Pooled Deep-Convolution Descriptor (TDD) +iDT		HMDB51 (65.9%) and UCF101 (91.5%)	RGB
Shao et al.	2016	S-CNN	2D CNN + 1D Temporal pooling	WWW crowd video dataset (62.55%)	RGB
Amir et al.	2016	Spatiotemporal LSTM (Tree Traversal) with Trust Gates		NTU RGB+D Dataset (77.7%) SBU Interaction Dataset (93.3%) UT-Kinect Dataset (97.0%) Berkeley MHAD (100%)	skeleton
Song et al.	2016	Deep LSTM with Spatio-Temporal Attention Model		NTU RGB+D Dataset (81.2%) SBU Interaction Dataset (91.51%)	skeleton
Ding et al.	2016	Profile HMM		MSRAction3D (86.4%), UTKinect-Action (91.7%), and UCF Kinect (97.6%) Dataset	skeleton
Liu et al.	2018	Global context aware attention LSTM network		UT-Kinect Action (99%), NTU RGB+D (84%) , SYSU-3D (79.1%), SBU-Kinect (94.6%) dataset	skeleton

wear the sensors, the actions of the person cannot be observed through the recognition system. However, it should be acknowledged that deep models have enhanced the performance of crowd analysis systems by handling the real-life challenges which are reducing the barrier of scene complexity for an excellent analysis of crowd activities.

Development and establishment of deeper architectures do involve some set of challenges such as a small set of training data, computational resources etc. Deep learning based neural networks cannot produce a realistic performance with small training datasets. Therefore, rigorous data augmentation techniques (Tran et al., 2017; Hanab et al., 2018) (i.e. cropping, rotating, and flipping input images) are utilised to analyze to small dataset samples for human action recognition. The work (Aquino et al., 2017) introduced 'Only augmented' and 'Balanced augmented,' two augmentation strategies and their effect on CNN architectures which is pointing the upgraded performance of CNNs. Transfer learning (Cook et al., 2013; Sargano et al., 2017) is another good approach to handle limited data samples for a specific application without involving considerable training time from scratch and large dataset by transferring the learned knowledge in one set of use to another.

The work (Hammerla and Plotz, 2015; Zhang et al., 2015; Alsheikh et al., 2016) performed in the direction to improve performance of smart homes, elderly people is majorly sensor based which is limited to a small set of data samples. If, we apply transfer learning techniques on visual datasets for abnormal action recognition and use LSTMs, RNNs to analyze temporal details involved, it can improve the image of single person abnormal human recognition.

Table 7
Single and multiple person based deeper abnormal human action recognition approaches.

Method	Year	Varient of CNN		Dataset (Accuracy)	Short-comings
A. **Single Person based Deep Abnormal HAR methods**					
Hammerla et al.	2016	Bi-directional-LSTM-S		Ubicom — Daphnet Gait dataset (DG) (F1-score-0.868)	The analysis is based on Wearable sensors' movement data
Arifoglu and Bouchachia	2017	Variants of RNN-Vanilla RNNs (VRNN), Long Short-Term RNNs (LSTM) and Gated Recurrent Unit RNNs (GRU)		Artificially generated dementia specific abnormal samples (96.7%)	A small set of sensor-based dataset
Park et al.	2018	Residual-RNN		MIT dataset (90.85%)	The absence of vision sensors in the smart and small dataset
B. **Multiple Persons based Deep Abnormal HAR methods**					
Karpathy et al.	2014	Two stream CNN	Context stream (low resolution image) and Fovea Stream (high resolution image)	Sports-1M (80.2%) and UCF101 (65.4%)	Weekly annotated data samples
Chong and Tay	2017	Autoencoders		Avenue (80.3%), UCSD Ped1 (89.1%)and Ped2 (87.4%), Subway entrance(84.7%) and exit (94%)datasets-AUC	Performance is sensitive to view variations.
Medel and Savakis	2016	Conv-LSTM		UCSD Ped1 (85.1%), Ped2 (92.3%), Avenue (85.1%), Subway Enter (81.57%) and Exit (65.9%)Datasets-Precision	–
Vignesh et al.	2017	CNN-LSTM-SVM-TA		ARENA Dataset (96.4%)	Model is evaluated for only one type of abnormal event
Hinami et al.	2017	Multi-task Fast R-CNN		Avenue (89.8%) and UCSD Ped2 (92.2%)- AUC	Fast R-CNN must be trained with richly annotated image datasets
Ravanbakhsh et al.	2017	Generative Adversarial Nets		UCSD Ped1 (97.4%, Ped2 (93.5%) and UMN dataset (99%)	–
Ravanbakhsh et al.	2016	Plug- and Play CNN		UCSD Ped1 (95.7%), Ped2 (88.4%) UMN (98%) dataset	–
Li and Chuah	2018	CNN-LSTM single frame		NCAA Basketball Dataset (94%) and the UCF Sports Action Dataset (92.8%).	Optical flow needs to be extracted manually

4. Discussion

For vision-based human action recognition systems as normal or abnormal in a different set of applications-AAL, smart homes, crowd analysis etc., a compact representation of the action videos needs to be defined considering the discrimination ability of the description along with time and computational complexity. The paper has covered three types of approaches to representing the actions for abnormal human recognition application specifically-2D AbHAR, 3D AbHAR and deep recognition systems for abnormal activities.

During the survey it is observed that every approach has certain limitations. Therefore the selection of the action description method must be application specific always. For RGB images/video sequences abnormal action detection and recognition can be categorized as a single person based and multiple person based approaches. It is noticed that for single person based AbHAR, silhouette and spatiotemporal based approaches became quite popular. In Khan and Sohn (2011) R-Transform is applied on silhouettes and KDA is applied to improve the discrimination ability of the descriptor. However, the system is evaluated for the hypothetical dataset. And (Sacco et al., 2012) demands high implementation cost and cannot perform the long-term real-time analysis. Whereas real-time dataset. However, LOTAR framework (Riboni et al., 2016) is developed for real-time abnormal human behavior detection at an early stage. However, this approach depends on multiple sensors other than vision sensor-camera. The integration of sensors and simultaneous acquisition and analysis of different sensors data become another challenge in real time. For abnormal behavior recognition with multiple persons spatiotemporal details (Roshtkhari and Levine, 2013; Roshtkhari and Levine, 2013) provide significant aspects of the abnormal crowd behavior. Sparseness (Chathuramali et al., 2014) of extracted features also play a major role in the accurate detection of the abnormal crowd behavior such as stampede.

For indoor activities where depth cameras can be fixed, skeleton and depth generation of a person is possible. It is witnessed during the survey that in closed environment single person based abnormal actions are analyzed keeping in mind the smart homes, elderly health care, and fall detection applications. From the performance point of view full Procrustes distance (Rougier et al., 2011b), OCSVM (Yu et al., 2013), History Triple Factor (HTF) (Goudelis et al., 2015), BoCSS (Ma et al., 2014) methods recognize the fall efficiently while handling occlusion, view variations quite well. However, (Ma et al., 2014; Goudelis et al., 2015) involve high computation cost not supporting real-time fall detection and (Yu et al., 2013) is evaluated only for one type of fall dataset.

The developed single person deep abnormal behavior recognition methods (Hammerla et al., 2016; Arifoglu and Bouchachia, 2017; Park

et al., 2018) utilised variants of RNN-Vanilla RNNs (VRNN), Long Short-Term RNNs (LSTM) and Gated Recurrent Unit RNNs (GRU) and Residual-RNN architectures to understand long-term details. However, no vision based single person deep abnormal behavior recognition model has been developed yet due to non-availability of sufficiently large abnormal datasets for deep networks driven by millions of data. Whereas for multiple person based AbHAR, a significant number of datasets are now available which is one of the major reasons that ample experiments have been performed using the combinations of CNNs and LSTMs (Vignesh et al., 2017; Medel and Savakis, 2016; Li and Chuah, 2018; Hinami et al., 2017; Ravanbakhsh et al., 2016), Generative Adversarial Nets (Ravanbakhsh et al., 2017) to define the deep model. However, some of the frameworks are trained with loosely annotated data, and some are not able to achieve real-time computational efficiency and accuracy.

5. Datasets

With the evolution of new technologies, there has been a tremendous rise in the number and also variations in publicly available datasets for experimentation. Recently introduced 3D AbHAR and AAL Datasets are summarized in Table 8 specifying the methodologies used in the introducing paper, accuracy, precision and recall obtained and set of challenges addressed by these datasets. Commonly used datasets for crowd activity analysis are UMN dataset (Detection of unusual crowd activity dataset, 2006) UCSD-ped1, ped2 dataset (UCSD Anomaly Detection Dataset, 2013),PETS dataset (Patino et al., 2016), web videos (Web Dataset, 2009), subway surveillance dataset, the boat-sea, boat-river, traffic-Belleview and airport-wrong-direction datasets (Anomalous Behavior Data Set, 2018). Recently a new dataset HIT-JUT (Han et al., 2016) is introduced for crowd activity detection which includes fight, shoot, and escape scenarios. It is captured in crowded indoor scenes with 15 people shot from two different views using two cameras. Saini et al. (2017) defined anomalies as Type-I and Type-II and their mathematical definitions. According to Type-I anomaly target resides within a region for a long duration of time. And in Type-II anomaly a target switches between two or more regions for a sustained duration. The above mentioned datasets (Akagunduz et al., 2016; Web Dataset, 2009; Diraco et al., 2010; Patino et al., 2016; Han et al., 2016; Anomalous Behavior Data Set, 2018) only exhibit Type-I anomaly. The In-House dataset introduced in the work (Saini et al., 2017) includes both Type-I and Type-II anomalies and provides wider set of anomaly situations. The dataset is recorded for 70 min using a static camera (25 FPS) hosted on top of a building, to capture public movement on a busy working day. The camera supports both HD as well as SD quality video recording. The video frames are prepossessed to 640 × 480. Whereas for single person Abnormal Human Action Recognition from 2011 to 2014 very less number of datasets were publicly available. In Khan and Sohn (2013, 2011), 2D-silhouette based abnormal dataset is introduced that considers six abnormal activities: forward fall, backward fall, chest pain, faint, vomit, and headache, stored in audio video interleave (AVI) format. The activities are executed by six persons (4 males, 2 females) by repeating ten sequences for each activity. There are 150 silhouettes for each activity and 900 silhouettes for each person performing the activities. Later, in the concern of ADL and elderly health care, good number of AbHAR 3D datasets are reported i.e. CAD60, CAD-120, TST fall detection dataset, UR fall detection, SisFall dataset, Le2i datasets.

CAD-60 (Cippitelli et al., 2016), CAD-120 are two popular datasets designed for home monitoring. CAD-120 includes 120 RGBD video sequences with 10-high level activities such as *making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, Picking objects, cleaning objects, taking food, arranging objects, having a meal and* 12 object affordance labels: *reachable, movable, pourable, pour to, containable, drinkable, openable, placeable, closable, scrubbable, scrubber, stationary*. To provide Activity of daily living (ADL) and fall solutions TST Fall detection dataset v2 (Gasparrini et al., 2015) offers different variations of fall : forward fall, backward fall, side fall, fall ending up sitting and daily activities such as walk back and forth, sit on a chair, walk and grasp object from the floor, lying down . The dataset has been collected using Microsoft Kinect v2 and IMU (Inertial Measurement Unit) manufactured by Shimmer Research and simulated by 11 volunteers. The database contains 264 different actions for a total no. of 46 k skeleton samples and 230 k acceleration values. These actions are stored as depth frames, skeleton joints in-depth, time information helpful for synchronization and two raw acceleration streams, provided by SHIMMER devices constrained to the waist and right wrist of the volunteer. UR fall detection (Kwolek, 2014) dataset contains 70 (30 falls + 40 activities of daily living) sequences. Fall events are chronicled with two Microsoft Kinect cameras and corresponding accelerometric data. ADL events are recorded with only one device (camera 0) and accelerometer. Sensor data was collected using PS Move (60 Hz) and x-IMU (256 Hz) devices. On the basis of a survey (Sucerquia et al., 2017), exact pre and post falling situation are identified with the help of three questions (i) which activity the person was performing when the fall happened? (ii) Reason for fall? A sliding, a faint, a trip, other? (iii) In which orientation did the fall happen and what part of the body received the impact? SisFall dataset, selected Lateral fall/Fall forward/fall backward while walking caused by a slip/trip or falling asleep or while getting up/sitting down, vertical fall while fainting after surveying. In total it includes 15 types of falls and 19 ADL activities. Le2i robust fall detection dataset (Nguyen et al., 2016) has 221 videos (320 × 240 pixels at 25 fps) of different living environments with various daily activities. These video sequences includes real-world challenges such as shadow and light reflection, complex textured background, actors wear different colored clothes.

6. Conclusion and future directions

The literature has covered rapid development of feature designing strategies, from hand-crafted features to deep features, for video sequences to support a real-time, robust and computationally efficient abnormal human activity recognition framework depending on the context of application such as fall detection, Ambient Assistive Living (AAL), homeland security, surveillance or crowd analysis. The feature designing approach also change with dimension of input in the application i.e. RGB, depth and skeleton. In past few years, due to the diffusion of infrared sensor (Microsoft Kinect), abnormal human activity recognition using depth and skeleton sequences has proliferated. But the use of Kinect sensor is limited to confined region, therefore experiments are performed using depth sensors only for fall detection, AAL and smart home applications for abnormal human action recognition. Depth and skeleton representation provide view and illumination invariance, which are observed as a very serious issue in crowd or public scenes where the background changes dynamically with illumination variations in open areas. For multiple persons based abnormal human activities RGB images are used. Deep features based abnormal human action description which has surpassed the performance of hand-crafted features due to the ability of learning the video scenes dynamically but demands high computational resources.

The paper also incorporated through study of the publicly available dataset in the field of fall detection, AAL and abnormal human action recognition to outline the scope of experiments. As far as AbHAR datasets are concerned, many Kinect-based 3D datasets — pose based human activity datasets : *MoCap* (Subtle Walking From CMU Mocap Dataset, 2018) , MHAD (Teleimmersion Lab, 2018) , MSRAction3D dataset (MSR ACtion 3D Dataset, 0000) ; human interaction datasets :*G3Di* (Bloom et al., 2014) , K3Hi (K3HI Kinect-based 3D Human Interaction Dataset, 2018) , CONVERSE (Edwards et al., 2016), In-HOUSE Dataset (Saini et al., 2017), Fu Kinect Fall Dataset (Aslan et al., 2017) are being generated but the type of dataset stipulating the abnormal actions specifically, identified very few (Nguyen et al., 2016; Khan and Sohn, 2013, 2011; Han et al., 2016; Gasparrini et al., 2015; Sucerquia et al., 2017).

Table 8
3D AbHAR and AAL datasets summary.

Dataset	Method	Year	Accuracy	Precision	Recall	resolution	challenges
CAD-60	Key pose extraction and association using clustering algorithm (Cippitelli et al., 2016)	2016	–	93.9%	93.5%	640 × 480	Illumination Invariant
	Neural integration pose motion features (Parisi et al., 2015)	2015		91.9%	90.2%		
	Pose Kinetic energy (Shan and Akella, 2014)	2014		93.8%	94.5%		
CAD-120	Conditional Random Field (CRF) (Koppula and Saxena, 2013)	2013	93.5%	95.0%	93.3%	640 × 480	
	Latent graphical structure (Hu et al., 2014)	2014	87.0%	89.2%	83.1%		
TST fall detection v2	Fusion of Kinect sensor and joint acceleration matrix (Gasparrini et al., 2015)	2015	99%	–	–	512 × 424	–
UR Fall detection	Hough Transform on V-disparity values of images	2014	99.67%	–	–	640 × 480	View-invariant
Le2i dataset	Feature Set — MHI based motion magnitude and orientation measure with human shape variations	2016	–	93%	100%	320 × 240 (25 fps)	• shadow • light reflection, • complex textured background, • actors wear different colored clothes
In-HOUSE Dataset	Non-overlapping block-based surveillance scene segmentation (Saini et al., 2017)	2015	83%	100%	83%	640 × 480	Demonstrates both Type-I and Type II anomaly
Fu Kinect Fall Dataset	coded regions on nested circles (Aslan et al., 2017)	2017	97.92% (max)	–	–	–	View invariance

In future, abnormal human action recognition systems need to address specific issues to bring abnormal action recognition methodologies to a real-time platform within the affordable range of the common users that will foster the security level to a higher level in the day-to-day routine of the common user.

- During the survey, it was witnessed that a few handcrafted (Al-Nawash et al., 2016; Roshtkhari and Levine, 2013; Wang et al., 2017b; Uddina et al., 2011; Yang et al., 2016; Triantafyllou et al., 2016; Yu et al., 2013) and deep features (Li and Chuah, 2018) based recognition systems' performances cannot be maintained near to real-time performances due to non-availability of large-sized real-time dataset samples for validation. That limits not only the generalizability of the systems but also robustness. Human skeletons provide view invariance to the action, but the availability of abnormal actions from a different view is still a big challenge. The work (Diraco et al., 2010) also depend on artificially generated data which does not reflect real-life challenges. Therefore, some meaningful datasets must be developed to represent abnormal actions in different scenarios-office, home, the coffee shop.
- Researchers are working on developing deep architectures from primary CNN to RCNN, RNN, auto-encoders. However, availability of deep architectures (Wang et al., 2015b) based efficient and deeper abnormal action recognition system with required computational resources to the common man is a serious the future concern.
- Though three dimensional data have enhanced the performance of recognition systems yet computational complexity is always high while using three-dimensional data. Therefore, first challenge lies in transferring the knowledge of three-dimensional data as depth or skeleton based feature descriptor to a real-time AbHAR systems with improved true detection rate and less computational complexity.

Acknowledgments

The authors would like to acknowledge the support of the Biometric Research Laboratory, Department of Information Technology, Delhi Technological University, New Delhi, India.

References

Abobakr, A., Hossny, M., Nahavandi, S., 2017. A skeleton-free fall detection system from depth images using random Decision Forest. IEEE Syst. J. 1–12.
Aburomman, A.A., Reaz, M.B.I., 2016. Ensemble of binary SVM classifiers based on PCA and LDA feature extraction for intrusion detection. In: Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC). IEEE, Xi'an, China.
Aggarwal, H., Vishwakarma, D.K., 2016. Covariate conscious approach for Gait recognition based upon Zernike moment invariants. IEEE Trans. Cognitive Develop. Syst. PP (99), 1-1.
Akagunduz, E., Aslan, M., Sengur, A., Wang, H., Ince, M., 2016. Silhouette orientation volumes for efficient fall detection in depth videos. IEEE J. Biomed. Health Inform. PP (99), 2168–2194.
Al-Nawash, M., Al-Hazaimeh, O.M., Mohamad, S., 2016. A novel framework for intelligent surveillance system based on abnormal human activity detection in academic environments. Neural Comput. Appl. 28 (1), 565–572.
Alsheikh, M.A., Seleim, A.A.S., Niyato, D., Doyle, L., Lin, S., Tan, H.P., Deep activity recognition models with triaxial accelerometers, CoRR, arXiv:abs/1511.04664, 2016.
Alvar, M., Torsello, A., Miralles, A.S., Armingol, J.M., 2014. Abnormal behavior detection using dominant sets. Mach. Vis. Appl. 25 (5), 1351–1368.
Amir, J.L., Xu, S.D., Wang, G., 2016. Spatio-temporal LSTM with trust gates for 3D human action recognition. In: European Conference on Computer Vision (ECCV), Amsterdam, The Netherland.
Andò, B., Baglio, S., Lombardo, C.O., Marletta, V., 2015. An event polarized paradigm for ADL detection in AAL context. IEEE Trans. Instrum. Meas. 64 (7), 1814–1825.
Anomalous Behavior Data Set, [Online]. Available: http://vision.eecs.yorku.ca/research/anomalous-behaviour-data/. (Accessed 16 May 2018).
Antic, B., Ommer, B., 2011. Video parsing for abnormality detection. In: 13th International Conference on Computer Vision, Barcelona.
Antic, B., Ommer, B., Spatio-temporal video parsing for abnormality detection, arXiv, arXiv:abs/1502.06235, 2015, pp. 1–15.
Aquino, N.M.R., Gutosk, M., Hattori, L.T., Lopes, H.S., 2017. The effect of data augmentation on the performance of convolutional neural networks. In: Brazilian Society of Computational Intelligence, Niterói, Rio de Janeiro.
Arifoglu, D., Bouchachia, A., 2017. Activity recognition and abnormal behaviour detection with recurrent neural networks. In: International Conference on Mobile Systems and Pervasive Computing, Leuven, Belgium.

Aslan, M., Akbulut, Y., Sengor, A., CevdetInce, M., 2017. Skeleton based efficient fall detection. J. Faculty Eng. Architecture Gazi Univ. 32 (4), 1025–1034.
Ben, A.M., Zagrouba, E., 2018. Abnormal behavior recognition for intelligent video surveillance. Expert Syst. Appl. 91, 480–491.
Bian, Z., Chau, L.P., Thalmann, N.M., 2012a. Fall detection based on skeleton extraction. In: International Conference on Virtual-Reality Continuum and its Applications in Industry, Singapore.
Bian, Z.P., Chau, L.P., Thalmann, N.M., 2012b. A depth video approach for fall detection based on human joins height and falling velocity. In: Proceedings of International Conference on Computer Animation and Social Agents, Singapore.
Bian, Z.P., Chau, L.P., Thalmann, N.M., 2012. Fall detection based on skeleton extraction. In: 11th International Conference on Virtual-Reality Continuum and its Applications in Industry, Singapore.
Bian, Z.P., Hou, J., Chau, L.P., Thalmann, N.M., 2015. Fall detection based on body part trackingusing a depth camera. IEEE J. Biomed. Health Inform. 19 (2), 430–439.
Bilen, H., Fernando, B., Gavves, E., Vedaldi, A., Gould, S., 2016. Dynamic image networks for action recognition. In: CVPR, Las Vegas.
Bloom, V., Argyriou, V., Makris, D., 2014. G3Di: A gaming interaction dataset with a real time detection and evaluation framework. In: Workshop at the European Conference on Computer Vision. Cham.
Candás, J.L.C., Peláez, V., López, G., Fernández, M.Á., Álvarez, E., Díaz, G., 2014. An automatic data mining method to detect abnormal humanbehaviour using physical activity measurements. Pervasive Mob. Comput. 15, 228–241.
Cardile, F., Iannizzotto, G., Rosa, F.L., 2010. A vision-based system for elderly patients monitoring. In: 3rd International Conference on Human System Interaction, Rzeszow.
CAVIAR test case scenarios, 2005. http://homepages.inf.ed.ac.uk/rbf/CAVIAR/.
Cecotti, H., Graser, A., 2010. Convolutional neural networks for P300 detection with application to brain-computer interfaces. IEEE Trans. Pattern Anal. Mach. Intell. 33 (3), 433–445.
Chaaraoui, A.A., Padilla-López, J.R., Flórez-Revuelta, F., 2015. Abnormal gait detection with RGB-D devices. In: 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia.
Chaaraoui, A.A., Pérez, P.C., Revuelta, F.F., 2012. A review on vision techniques applied to human behaviour analysis for ambient-assisted living. Expert Syst. Appl. 39 (12), 10873–10888.
Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection : a survey. ACM Comput. Surv. 41 (3).
Chathuramali, K.G.M., Ramasinghe, S., Rodrigo, R., 2014. Abnormal activity recognition using spatio-temporal features. In: 7th International Conference of Information and Automation of Sustainablilty, Colombo.
Chen, S., Gangopadhyay, A., 2016. Health care fraud detection with community detection algorithms. In: IEEE International Conference on Smart Computing (SMARTCOMP), St. Louis, MO.
Chen, L., Wei, J.F., H., 2013. A survey on human motion analysis using depth imagery. Pattern Recognit. Lett. 34, 1995–2006.
Chien, T.L., Su, K.L., Guo, J.H., 2005. The multiple interface security robot - WFSR-II. In: IEEE International Safety, Security and Rescue Rototics, Workshop, Kobe, Japan.
Cho, S.H., Kang, H.B., 2014. Abnormal behavior detection using hybrid agents in crowded scenes. Pattern Recognit. Lett. 54, 64–70.
Chong, Y.S., Tay, Y.H., 2017. Abnormal event detection in videos using spatiotemporal autoencoder. In: International Symposium on Neural Networks, Japan.
Chou, C.L., Chen, H.T., Lee, S.Y., 2017. Multimodal video-to-near-scene annotation. IEEE Trans. Multimed. 19 (2), 354–366.
Cippitelli, E., Gasparrini, S., Gambi, E., 2016. Ahuman activity recognition system using skeleton data from RGBD sensors. Comput. Intell. Neurosci. 2016, 1–14.
Coşar, S., Donatiello, G., Bogo, V., Garate, C., Alvares, L.O., Brémond, F., 2017. Toward abnormal trajectory and event detection in video surveillance. IEEE Trans. Circuit Syst. Video Technol. 27 (3), 683–695.
Cong, Y., Yuan, J., Tang, Y., 2013. Video anomaly search in crowded scenes via spatio-temporal motion context. IEEE Trans. Inform. Forensics Secur. 8 (10), 1590–1599.
Cook, D.J., Feuz, K.D., Krishnan, N.C., 2013. Transfer learning for activity recognition: a survey. Knowledge Inform. Syst. 36 (3), 537–556.
Crispim Junior, C.F., Joumier, V., Hsu, Y.L., Chung, P.C., Dechamps, A., Pai, M.C., Robert, P., Bremond, F., 2012. Alzheimer's patient activity assessment using different sensors. Gerontechnology 11 (2), 266–267.
Dai, X., Wu, M., Davidson, B., Mahoor, M., Zhang, J., 2013. Image-based fall detection with human posture sequence modelling. In: IEEE International Conference on Healthcare Informatics, Philadelphia, USA.
Detection of unusual crowd activity dataset, 2006. http://mha.cs.umn.edu/proj_events.shtml#crowd.
Diba, A., Gool, L.V., 2016. Efficient two-stream motion and appearance 3D CNNs for video classification. In: European Conference on Computer Vision, ECCV, Amsterdam, The Netherlands.
Ding, W., Liu, K., Fu, X., Cheng, F., 2016. Profile HMMs for skeleton-based human action recogniti. Signal Process., Image Commun. 42, 109–119.
Diraco, G., Leone, A., Siciliano, P., 2010. An active vision system for fall detection and posture recognition in elderly healthcare. In: Design, Automation & Test in Europe Conference & Exhibition, Dresden.
Dogra, D.P., Reddy, R., Subramanyam, K., Ahmed, A., Bhaskar, H., 2015. Scene representation and anomalous activity detection using weighted region associated graph. In: 10th International Conference on Computer Vision Theory and Applications, Berlin, Germany.
Dragone, M., Amato, G., Bacciu, D., Chessa, S., Coleman, S., Rocco, M.D., Gallicchio, C., Gennaro, C., Lozano, H., Maguire, L., McGinnity, M., Micheli, A., O'Hare, M.P., G., Renteria, A., Saffiotti, A., Vairo, C., Vance, P., 2015. A cognitive robotic ecology approach to self-configuring and evolving AAL systems. Eng. Appl. Artif. Intell. 45, 269–280.
EC, 2012. Active ageing special eurobarometer 378, tech. rep. DG COMM "Research and Speechwriting" Unit, European Comission. In: Conducted by TNS Opinion & Social at the request of Directorate-General for Employment, Social Affairs and Inclusion.
Edwards, M., Deng, J., Xie, X., 2016. From Pose to Activity : Surveying datasets and introducing CONVERSE. Comput. Vis. Image Underst. 144, 73–105.
Eng, H.L., Thida, M., Remagnino, P., 2013. Laplacian eigenmap with temporal constraints for local abnormality detection in crowded scenes. IEEE Trans. Cybern. 43 (6), 2147–2156.
Escalera, S., Baro, X., Gonzalez, J., Bautista, A.M., Madadi, M., Reyes, M., Ponce-Lopez, V., Escalante, J.H., Shotton, J., Guyon, I., 2014. Chalearn looking at people challenge 2014: Dataset. In: Computer Vision - ECCV 2014 Workshop, Zurich, Switzerland.
Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., Athitsos, V., Escalante, H.J., 2013. Multi-modal gesture recognition challenge 2013: Dataset. In: International Conference on Multimodal Interaction, Sydney, Australia.
Fagette, A., Courty, N., Racoceanu, D., Dufour, J.Y., 2014. Unsupervised dense crowd detection by multiscale texture analysis. Pattern Recognit. Lett. 44, 126–133.
Feng, B., He, F., Wang, X., Wu, Y., Wang, H., Yi, S., Liu, W., 2016. Depth-projection-map-based bag of contour fragments for robust hand Gesture Recognition. IEEE Trans. Hum.-Mach. Syst. PP (99), 1–13.
Fothergill, S., Mentis, H., Kohli, P., Nowozin, S., 2012. Instructing people for training gestural interactive systems. In: Conference on Human Factors in Computing Systems, Austin, Texas.
Gasparrini, S., Cippitelli, E., Gambi, E., Spinsan, S., Wahslen, J., Orhan, I., Lindh, T., 2015. Proposal and experimental evaluation of fall detection solution based on wearable and depth data fusion. In: ICT Innovations, Advances in Intelligent Systems and Computing. Cham.
Gasparrini, S., Cippitelli, .E., Spinsante, S., Gambi, E., 2014. A depth-based fall detection system using a Kinect sensor. Sensors 14 (2), 2756–2775.
Gorai, A., Pal, R., Gupta, P., 2016. Document fraud detection by ink analysis using texture features and histogram matching. In: International Joint Conference on Neural Networks (IJCNN), Vancouver, BC.
Goudelis, G., Tsatiris, G., Karpouzis, K., Kollias, S., 2015. Fall detection using History Triple Features. In: Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Corfu, Greece.
Gowsikhaa, D., Abirami, S., Baskaran, R., 2014. Automated human behavior analysis from surveillance. Artif. Intell. Rev. 42 (5), 747–765.
Gu, X., Cui, J., Zhu, Q., 2014. Abnormal crowd behavior detection by using the particle entropy. Int. J. Light Electron Opt. 125 (14), 3428–3433.
Guo, C., Ma, Q., Zhang, L., 2008. Spatio-temporal saliency detection using phase spectrum of quaternion Fourier transform. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, Alaska.
Guo, H., Wu, X., Cai, S., Li, N., Cheng, J., Chen, Y.L., 2016. Quaternion discrete cosine transformation signature analysis in crowd scenes for abnormal event detection. Neurocomputing 204, 106–115.
Hammerla, N., Halloran, S., Pltz, T., 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. In: International Joint Conference on Artificial Intelligence, New York.
Hammerla, N.Y., Plotz, T., 2015. Let's (not) stick together: pairwise similarity biases cross-validation in activity recognition. In: ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), Umeda, Osaka, Japan.
Han, T., Yao, H., Sun, X., Zhao, S., Zhang, Y., 2016. Unsupervised discovery of crowd activities by saliency-based clustering. Neurocomputing 171, 347–361.
Hana, Y., Zhanga, P., Zhuob, T., Huangc, W., Zhang, Y., 2018. Going deeper with two-stream ConvNets for action recognition in video surveillance. Pattern Recognit. Lett. 107, 83–90.
Hanab, D., Liua, Q., Fan, W., 2018. A new image classification method using CNN transfer learning and web data augmentation. Expert Syst. Appl. 95 (1), 43–56.
Hassner, T., Itcher, Y., Gross, O.K., 2012. Violent flows: Real-time detection of violent crowd behavior. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI.
Hendryli, J., Fanany, M.I., 2016. Classifying abnormal activities in exam using multi-class Markov chain LDA based on MODEC features. In: Fourth International Conference on Information and Communication Technologies, Bandung, Indonesia.
Herath, S., Harandi, M., Porikli, F., 2017. Going deeper into action recognition: a survey. Image Vis. Comput. 60, 4–21.
Hinami, R., Mei, T., Satoh, S., 2017. Joint detection and recounting of abnormal events by learning deep generic knowledge. In: ICCV, Venice Italy.
Hodge, V.J., Austin, J., 2004. A survey of outlier detection methodologies. Artif. Intell. Rev. 22 (2), 85–126.
Hou, J.C., Wang, S.S., Lai, Y.H., Tsao, Y., Chang, H.W., Wang, H.M., 2018. Audio-visual speech enhancement using multimodal deep convolutional neural networks. IEEE Trans. Emerg. Top. Comput. Intell. 2 (2), 117–128.

Hsieh, J.W., Chuang, C.H., Alghyali, S., Chiang, H.F., Chiang, C.H., 2015. Abnormal scene change detection from a moving camera using bag of patches and spider web map. IEEE Sens. J. 15 (5), 2866–2881.

Hu, N., Englebienne, G., Lou, Z., Kröse, B., 2014. Learning latent structure for activity recognition. In: IEEE International Conference on Robotics and Automation, Hong Kong.

Hu, Y., Zhang, Y., Davis, L., 2013. Unsupervised abnormal crowd activity detection using semiparametric scan statistic. In: Computer Vision and Pattern Recognition Workshop, Portland, Oregon.

Huang, B., Tian, G., Wu, H., Zhou, F., 2014. A method of abnormal habits recognition in intelligent space. Eng. Appl. Artif. Intell. 29, 125–133.

Hung, Y.X., Chiang, C.Y., Hsu, S.J., Chan, C.T., 2010. Abnormality detection for improving elder's daily life independent. In: International Conference on Smart Homes and Health Telematics, Korea.

Iannizzotto, G., Lanzafame, P., Rosa, F.L., 2005. A CNN-based framework for 2D still-image segmentation. In: International Workshop on Computer Architecture for Machine Perception, Palermo, Italy.

Ijjina, E.P., Chalavadi, K.M., 2017. Human action recognition in RGB-D videos using motion sequence information and deep learning. Pattern Recognit. 72, 504–516.

Ismail, S.J., Rahman, M.A.A., Mazlan, S.A., Zamzuri, H., 2015. Human gesture recognition using a low cost stereo vision in rehab activities. In: IEEE International Symposium on Robotics and Intelligent Sensors (IRIS), Langkawi.

Jalal, A., Kamal, S., Kim, D., 2014. A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environment. J. Sensors 14 (7), 11735–11759.

Jansen, B., Deklerck, R., 2006. Context aware inactivity recognition for visual fall detection. In: Pervasive Health Conference and Workshops. IEEE, Innsbruck.

Ji, S., Xu, W., Yang, M., Yu, K., 2013. 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. 35 (1), 221–231.

Jing, L., Ye, Y., Yang, X., Tian, Y., 2017. 3D convolutional neural network with multi-model framework for action recognition. In: International Conference on Image Processing (ICIP), Beijing, China.

Jo, H., Chug, K., Sethi, R.J., 2013. A review of physics-based methods for group and crowd analysis in computer vision. J. Postdr. Res. Postdr Aff. 1 (1), 4–7.

Jyothilakshmi, P., Rekha, K.R., Nataraj, K.R., 2016. Patient assistance system in a super speciality hospital using a kinect sensor camera. In: International Conference on Electrical, Electronics, and Optimization Techniques, Chennai.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthanka, R., Fei-Fei, L., 2014. Large-scale video classification with convolutional neural networks In: Computer Vision and Pattern Recognition, Columbus, Ohio.

Khan, S.S., Hoey, J., 2017. Review of fall detection techniques: a data availability perspective. Med. Eng. Phys. 39, 12–22.

Khan, Z.A., Sohn, W., 2011. Abnormal human activity recognition system based on R-transform and kernel discriminant technique for elederly home care. IEEE Trans. Consum. Electron. 57 (4), 1843–1850.

Khan, Z.A., Sohn, W., 2013. A hierarchical abnormal human activity recognition system based on R-transform and kernel discriminant analysis for elderly health care. Computing 95 (2), 109–127.

Kok, V.J., Li, M.K., Chan, C.S., 2016. Crowd behavior analysis: a review where physics meets biology. Neurocomputing 177, 342–362.

Koohzadi, M., Charkari, N.M., 2017. Survey on deep learning methods in human action recognition. IET Comput. Vis. 11 (8), 623–632.

Koppula, H.S., Saxena, A., 2013. Learning spatio-temporal structure from RGB-D videos for human activity detection and anticipation. In: 30 th International Conference on Machine Learning, Atlanta, USA.

Kulkarni, P., Patil, B., Joglekar, B., 2015. An effective content based video analysis and retrieval using pattern indexing techniques. In: International Conference on Industrial Instrumentation and Control (ICIC), Pune.

Kwolek, M.K.B., 2014. Human fall detection on embedded platform using depth maps and wireless accelerometer. Comput. Methods Programs Biomed. 117 (3), 489–501.

K3HI: Kinect-based 3D Human Interaction Dataset, [Online] Available: http://www.lmars.whu.edu.cn/prof_web/zhuxinyan/DataSetPublish/dataset.html. (Accessed 16 May 2018).

Lane, N.D., Georgiev, P., Qendro, L., 2015. Deepear: robust smartphone audio sensing in unconstrained acoustic environments using deep learning. In: ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp), Umeda, Osaka, Japan.

Leach, M.J., Sparks, E., Robertson, N.M., 2014. Contextual anomaly detection in crowded surveillance scenes. Pattern Recognit. Lett. 44, 71–79.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE.

Li, X., Chuah, M.C., 2018. ReHAR: Robust and Efficient Human Activity Recognition, in IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe.

Li, C., Han, Z., Ye, Q., Jiao, J., 2012. Visual abnormal behavior detection based on trajectory sparse reconstruction analysis. Neurocomputing 119, 94–100.

Li, Y., Li, X., Jia, L., 2014a. Abnormal crowd behavior detection based on optical flow and dynamic threshold. In: Proceeding of the 11th World Congress on Intelligent Control and Automation, Shenyang, China.

Li, R., Lua, B., Maier, K.D.M., 2015a. Cognitive assisted living ambient system: a survey. Digit. Commun. Netw. 1 (4), 229–252.

Li, W., Mahadevan, V., Vasconcelos, N., 2014b. Anomaly detection and localization in crowded scenes. IEEE Trans. Pattern Anal. Mach. Intell. 36 (1), 18–32.

Li, A., Miao, Z., Cen, Y., Liang, Q., 2016. Abnormal event detection based on sparse reconstruction in crowded scenes. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai.

Li, H.C.T., Wang, M., Ni, B., Hong, R., Yan, S., 2015b. Crowd scene analysis : a survey. IEEE Trans. Circuits Syst. Video Technol. 25 (3), 1–20.

Li, N., Wu, X., Xu, D., Guo, u., Feng, W., 2015c. Spatio-temporal context analysis within video volumes for anomalous-event detection and localization. Neurocomputing 155, 309–319.

Liu, W., Fan, Y., Lei, T., Zhang, Z., 2014. Human gesture recognition using orientation segmentation feature on random rorest. In: IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), Xi'an.

Liu, P., Tao, Y., Zhao, W., Tang, X., 2017. Surveillance scene segmentation based on trajectory classification using supervised learning. Neurocomputing 269, 3–12.

Liu, J., Wang, G., Duan, L.Y., Abdiyeva, K., Kot, A.C., 2018. Skeleton-based human action recognition with global context-aware attention LSTM networks. IEEE Trans. Image Process. 27 (4), 1586–1599.

Liua, Z., Zhangb, C., Tian, Y., 2016. 3D-based deep convolutional neural network for action recognition with depth sequences. Image Vis. Comput. 55 (2), 93–100.

Loy, C.C., Chen, K., Gong, S., Tao, X., 2013. Crowd counting and profiling: methodology and evaluation. In: Modeling, Simulation and Visual Analysis of Crowds. In: The International Series in Video Computing, vol. 11, Springer, New York, NY, pp. 347–382.

Zolfaghari, S., Keyvanpour, M.R., 2016. SARF: Smart activity recognition framework in Ambient Assisted Living. In: Federated Conference on Computer Science and Information Systems (FedCSIS), Gdansk.

Ma, X., Wang, H., Xue, B., Zhou, M., Ji, B., Li, Y., 2014. Depth-based human fall detection via shape features and improved extreme learning machine. IEEE J. Biomed. Health Inform. 18 (6), 1915–1922.

Medel, J., Savakis, A., 2016. Anomaly detection using predictive convolutional long short-term memory units, arXiv:1612.00390v2 [cs.CV].

Mehran, R., Oyama, A., Shah, M., 2009. Abnormal crowd behavior detection using social force model. In: IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL.

Miguel, K.d., Brunete, A., Hernan, M., 2017. Home camera-based fall detection system for the elderly. Sensors (Basel) 17 (12).

Mohamed, O., Choi, H.J., Iraqi, Y., 2014. Fall detection systems for elderly care: A survey. In: International Conference on New Technologies, Mobility and Security (NTMS), Dubai, United Arab Emirates.

Mosquera, J.H., Loaiz, H., Nope, S.E., Restrepo, A.D., 2017. Identifying facial gestures to emulate a mouse: navigation application on Facebook. IEEE Latin Amer. Trans. 15 (1), 121–128.

MSR action 3D dataset, [Online]. Available: http://www.uow.edu.au/~wanqing/#Datasets.

Mubashir, M., Shao, L., Seed, L., 2013. A survey on fall detection : principles and approaches. Neurocomputing 100, 144–152.

Nar, R., Singal, A., Kumar, P., 2016. Abnormal activity detection for bank ATM surveillance. In: International Conference on Advances in Computing, Communications and Informatics, Jaipur, India.

Nguyen, V.A., Le, T.H., Nguyen, T.T., 2016. Single camera based fall detection using motion and human shape features. In: 7th International Symposium on Information and Communication Technology, Hochiminh city, Vietnam.

Nizam, Y., Mohd, M.N.H., Mohd, H., Jamil, M.M.A., 2016. Development of human fall detection system using joint height, joint velocity and joint position from depth maps. J. Telecommun. Electron. Comput. Eng. 8 (6), 125–131.

Paiement, A., Tao, L., Camplani, M., Hannuna, S., Damen, D., Mirmehdi, M., 2014. Online quality assessment of human motion from skeleton data. In: Proceedings of the British Machine Vision Conference, Nottingham.

Panahi, L., Ghods, V., 2018. Human fall detection using machine vision techniques on RGB–D images. Biomed. Signal Process. Control 44, 146–153.

Parisi, G.I., Weber, C., Werm, S., 2015. Self-organizing neural integration of pose-motion features for human action recognition. Front. Neurorobot. 9 (3), 1–14.

Park, J., Jang, K., Yang, S.B., 2018. Deep neural networks for activity recognition with multi-sensor data in a smart home. In: IEEE World Forum on Internet of Things (WF-IoT), Singapore.

Patino, L., Cane, T., Vallee, A., Ferryman, J., 2016. PETS 2016: Dataset and challenge. In: CVPR, Las Vegas, NV, USA.

Paul, M., Haque, S., Chakraborty, S., 2013. Human detection in surveillance videos and its applications - a review. EURASIP J. Adv. Signal Process. 1.

Piciarelli, C., Foresti, G.L., 2011. Surveillance-oriented event detection in video streams. IEEE Intell. Syst. 26 (3), 32–41.

Popoola, O.P., Wan, K., 2012. Video-based abnormal human behavior recognition—a review. IEEE Trans. Syst. Man Cybern. 42 (6), 865–878.

Presti, L.L., Cascia, M.L., 2016. 3D skeleton based human action classification : A survey. Pattern Recognit. 53, 130–147.

Rafferty, J., Nugent, C.D., Liu, J., Chen, L., 2017. From activity recognition to intention recognition for assisted living within smart homes. IEEE Trans. Hum.-Mach. Syst. PP (99), 1–12.

Rashidi, P., Mihailidis, A., 2013. A survey on ambient-assisted living tools for older adults. IEEE J. Biomed. Health Inform. 17 (3), 579–590.
Ravanbakhsh, M., Nabi, M., Mousavi, H., Sangineto, E., Sebe, N., 2016. Plug-and-play Cnn for crowd motion, arXiv:1610.00307.
Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., Sebe, N., 2017. Abnormal event detection in videos using generative adversarial nets. In: IEEE International Conference on Image Processing (ICIP), Beijing, China.
Riboni, D., Bettini, C., Civitarese, G., Janjua, Z.H., Helaoui, R., 2015. Fine-grained recognition of abnormal behaviors for early detection of mild cognitive impairment. In: IEEE International Conference on Pervasive Computing and Communications, St. Louis, USA.
Riboni, D., Civitarese, G., Bettini, C., 2016. Analysis of long-term abnormal behaviors for early detection of cognitive decline. In: IEEE International Workshop on PervAsive Technologies and care systems for sustainable Aging-in-place, Sydney.
Roshtkhari, M.J., Levine, M.D., 2013. Online dominant and anomalous behaviour detection in videos, In: Conference on Computer Vision and Pattern Recognition (CVPR), Portland, Oregon.
Roshtkhari, M.J., Levine, M.D., 2013. An on-line real-time learning method for detecting anomalies in videos using spatio-temporal compositions. Comput. Vis. Image Underst. 117 (10), 1436–1452.
Roudposhti, K.K., Dias, J., Peixoto, P., Metsis, V., Nunes, U., 2017. A multilevel body motion-based human activity analysis methodology. IEEE Trans. Cogn. Develop. Syst. 9 (1), 16–29.
Rougier, C., Auvinet, E., Rousseau, J., Mignotte, M., Meunier, J., 2011a. Fall detection from depth map video sequences. In: International Conference on Smart Homes and Health Telematics, Montreal.
Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J., 2006. Monocular 3D head tracking to detect falls of elderly people. In: IEEE International Conference on Engineering in Medicine and Biology Society, New York.
Rougier, C., Meunier, J., St-Arnaud, A., Rousseau, J., 2011b. Robust video surveillance for fall detection based on human shape deformation. IEEE Trans. Circuit. Syst. Video Technol. 21 (5), 611–622.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei, F.L., 2014. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.
Sacco, G., Joumier, V., Darmon, N., Dechamps, A., Derreumaux, A., Lee, J.H., Piano, J., Bordone, N., Konig, A., Teboul, B., Davi, R., Guerin, O., Bremond, F., Robert, P., 2012. Detection of activities of daily living impairment in Alzheimer's disease and mild cognitive impairment using information and communication technology. Clin. Interv. Aging 7, 539–547.
Saini, R., Sk, A.A., Dogra, D.P., Roy, P.P., 2017. Surveillance scene segmentation based on trajectory classification using supervised learning. In: Proceedings of International Conference on Computer Vision and Image Processing, Roorke.
Sargano, A.B., Wang, X., Angelov, P., Habib, Z., 2017. Human action recognition using transfer learning with deep representations. In International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA.
Shan, J., Akella, S., 2014. 3D human action segmentation and recognition using pose kinetic energy. In: IEEE Workshop on Advanced Robotics and its Social Impacts, Evanston, Illinois, USA.
Shao, J., Loy, C.C., Kang, K., Wang, X., 2016. Slicing convolutional neural network for crowd video understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA.
Shih, H.C., 2017. A survey on content-aware video analysis for sports. IEEE Trans. Circuits Syst. Video Technol. PP (99), 1-1.
Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems (NIPS), Montreal; Canada.
Sindagi, V.A., Patel, V.M., 2018. A survey of recent advances in CNN-based single image crowd counting and density estimation. Pattern Recognit. Lett. 107, 3–16.
Singh, K., Vishwakarma, D.K., Walia, G.S., Kapoor, R., 2016. Contrast enhancement via texture region based histogram equalization. J. Modern Opt. 63 (15), 1444–1450.
Song, X., Fan, G., 2006. Joint key-frame extraction and object segmentation for content-based video analysis. IEEE Trans. Circuits Syst. Video Technol. 16 (7), 904–914.
Song, S., Lan, C., Xing, J., Zen, W., Liu, J., 2016. An end-to-end spatio-temporal attention model for human action recognition from skeleton data, CoRR, arXiv:1611.06067.
Stein, M., Janetzko, H., Lamprecht, A., Seebacher, D., Schreck, T., Keim, D., Grossniklaus, M., 2016. From game events to team tactics: Visual analysis of dangerous situations in multi-match data. In: International Conference on Technology and Innovation in Sports, Health and Wellbeing (TISHW), Vila Real.
Stephens, K., Bros, A.G., Grouping multi-vector streaklines for human activity identification. In: IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop, Bordeaux, 2016.
Subtle Walking From CMU Mocap Dataset, [Online]. Available: http://users.eecs.northwestern.edu/~jwa368/my_data.html. (Accessed 16 May 2018).
Sucerquia, A., López, J.D., Vargas-Boni, J.F., 2017. SisFall: a fall and movement dataset. Sensors 17 (198), 1–14.
Synder, D., 2001. Online Intrusion Detection Using Sequences of System Calls. (MS thesis), Department of Computer Science, Florida State University.
Synnott, J., Nugent, C., Jeffers, P., 2015. Simulation of smart home activity datasets. Sensors 15 (6), 14162–14179.
Taylor, G.W., Fergus, R., LeCun, Y., Breg, C., 2010. Convolutional Learning of Spatio-temporal Features, In: ECCV, Greece.
Teleimmersion Lab, [Online]. Available: http://tele-immersion.citris-uc.org/berkeley_mhad/. (Accessed 16 May 2018).
Tham, J.S., Chang, Y.C., Fauzi, M.F.A., 2014. Automatic identification of drinking activities at home using depth. In: International Conference on Control, Automation and Information Sciences, Gwangju, Korea.
Thida, M., Yong, Y.L., Pérez, P.C., Eng, H.l., Remagnino, P., 2013. A literature review on video analytics of crowded scenes. In: Intelligent Multimedia Surveillance: Current trends and research. Springer, Berlin Heidelberg, pp. 17–36.
Tolosana, R., Rodriguez, R.V., Fierrez, J., Garcia, J.O., 2018. Exploring recurrent neural networks for on-line handwritten Signature Biometrics. IEEE Access 6, 5128–5138.
Toreyin, B.U., Dedeoglu, Y., Cetin, A.E., 2006. HMM based falling person detection using both audio and video. In: Signal Processing and Communications Applications, Antalya.
Tran, T.T.H., Le, T.L., Morel, J., 2014. An analysis on human fall detection using skeleton from Microsoft kinect. In: International Conference on Communication and Electronics (ICCE), Danang, Vietnam.
Tran, T., Pham, T., Carneiro, G., Palm, L., Reid, I., 2017. A Bayesian data augmentation approach for learning deep models. In: Neural Information Processing Systems (NIPS), Long Beach California.
Triantafyllou, D., Krinidis, S., Ioannidis, D., Metaxa, I.N., Ziazios, C., Tzovaras, D., 2016. A real-time fall detection system for maintenance activities in indoor environments. IFAC-PapersOnLine 49 (28), 286–290.
Tripathi, G., Singh, K., Vishwakarma, D.K., 2018. Convolutional neural networks for crowd behaviour analysis: a survey. Vis. Comput. PP, 1–24.
Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O., 2008. Machine recognition of human activities: a survey. IEEE Trans. Circuits Syst. Video Technol. 18 (11), 1473–1488.
UCSD Anomaly Detection Dataset, 2013. http://svcl.ucsd.edu/projects/anomaly/dataset.htm.
Uddin, M.Z., haksar, W., Torresen, J., 2017. Facial expression recognition using salient features and convolutional neural network. IEEE Access 5.
Uddin, M.Z., Kim, J.T., Kim, T.S., 2014. Depth video-based gait recognition for smart home using local directional pattern features and hidden markov model. Indoor Built Environ. 23 (1), 133–140.
Uddina, M.Z., Kim, T.S., Kim, J.T., 2011. Video-based indoor human gait recognition using depth imaging and hidden markov model : a smart system for smart home. Indoor Built Environ. 20 (1), 120–128.
Vaswani, N., Chowdhury, A.K.R., Chellappa, R., 2005. Shape activity: a continuous-state HMM for moving/deforming shapes with application to abnormal activity detection. IEEE Trans. Image Process. 14 (10), 1603–1616.
Vesperini, F., Vecchiotti, P., Principi, E., Squartini, S., Piazza, F., 2018. Localizing speakers in multiple rooms by using deep neural networks. Comput. Speech Lang. 49, 83–106.
Vignesh, K., Yadav, G., Sethi, A., 2017. Abnormal event detection on BMTT-PETS 2017 surveillance challenge. In: CVPR, Honolulu, Hawaii.
Vishwakarma, D.K., Dhiman, A., Maheshwari, R., 2015a. Human motion analysis by fusion of silhouette orientation and shape features. In: 3rd International Conference on Recent Trends in Computing (ICRTC), Ghaziabad, India.
Vishwakarma, D.K., Kapoor, R., Dhiman, A., 2016a. A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics. Robot. Auton. Syst. 77, 25–38.
Vishwakarma, D.K., Kapoor, R., Dhiman, A., 2016b. Unified framework for human activity recognition: An approach using spatial edge distribution and R-transform. AEU-Int. J. Electron. Commun. 70 (5), 341–353.
Vishwakarma, D.K., Rawat, P., Kapoor, R., 2015b. Human activity recognition using gabor wavelet transform and ridgelet transform. In 3rd International Conference on Recent Trends in Computing (ICRTC), Ghaziabad, India.
Vishwakarma, D.K., Singh, K., 2016. Human activity recognition based on spatial distribution of gradients at sub-levels of average energy silhouette images. IEEE Trans. Cogn. Develop. Syst. 9 (4), 316–327.
Wang, L., Ge, L., Li, R., Fang, Y., 2017a. Three-stream CNNs for action recognition. Pattern Recognit. Lett. 92 (1), 33–40.
Wang, L., Qiao, Y., Tang, X., 2015. Action recognition with trajectory-pooled deep-convolutional descriptors. In: CVPR, Boston, MA, USA.
Wang, L., Xiong, Y., Wang, Z., Qiao, Y., 2015. Towards good practices for very deep two-stream ConvNets, CoRR, arXiv:abs/1507.02159.
Wang, L., Xu, Y., Cheng, J., Xia, H., Yin, J., 2018. Human action recognition by learning spatio-temporal features with deep neural network. IEEE Access 6, 17913–17922.
Wang, C., Yao, H., Sun, X., 2017b. Anomaly detection based on spatio-temporal sparse representation and visual attention analysis. Multimedia Tools Appl. 76 (5), 6263–6279.
Web Dataset: Abnormal/Normal Crowds, 2009. Available from http://crcv.ucf.edu/projects/Abnormal_Crowd/Normal_Abnormal_Crowd.zip.
World Health Organization (WHO), 2008. Global report on falls prevention in older age, Geneva.
Yang, L., Ren, Y., Hu, H., Tian, B., 2015. New fast fall detection method based on spatio-temporal context tracking of head by using depth images. Sensors 15, 23004–23019.
Yang, L., Ren, Y., Zhang, W., 2016. 3D depth image analysis for indoor fall detection of elderly people. Digit. Commun. Netw. 2, 24–34.

Yao, L., Min and. K. Lu, W., 2017. A new approach to fall detection based on the human torso motion model. Appl. Sci. 7.
Yin, W., Kann, K., Yu, M., Schütz, H., Comparative Study of CNN and RNN for Natural Language Processing, 2017. arXiv:1702.01923.
Yu, X., 2008. Approaches and principles of fall detection for elderly and patient. In: 10th International Conference on E-health Networking, Applications and Services, Singapore.
Yu, S.J., Koh, P., Kwon, H., Kim, D.S., Kim, H.K., 2016. Hurst parameter based anomaly detection for intrusion detection system. In: IEEE International Conference on Computer and Information Technology (CIT), Nadi.
Yu, M., Yu, Y., Rhuma, A., Mohsen, S., Naqvi, R., Wang, L., Chambers, J.A., 2013. An online one class support vector machine-based person specific fall detection system for monitoring an elderly individual in a room environment. IEEE J. Biomed. Health Inform. 17 (6), 1002–1014.
Zerrouki, N., Harrou, F., Sun, Y., Houacine, A., 2016. Accelerometer and camera-based strategy for improved human fall detection. J. Med. Syst. 40 (12), 1–6.
Zhan, Y., Lu, H., Zhang, L., Ruan, X., 2016. Combining motion and appearance cues for anomaly detection. Pattern Recognit. 51, 443–452.
Zhang, Z., Liu, W., Metsis, V., Athitsos, V., 2012. A viewpoint-independent statistical method for fall detection. In: 21st International Conference on Pattern Recognition, Tsukuba.
Zhang, Y., Lu, H., Zhang, L., Ruan, X., Sakai, S., 2016. Video anomaly detection based on locality sensitive hashing filters. Pattern Recognit. 59, 302–311.
Zhang, Z., Ma, X., Song, R., Ron, X., Tian, X., Tian, G., Li, Y., 2017. Deep learning based human action recognition: A survey, in Chinese Automation Congress (CAC), Jinan, China.
Zhang, Y., Shang, K., Wang, J., Li, N., Zhang, M.M., 2018a. Patch strategy for deep face recognition. IET Image Process. 12 (5), 819–825.
Zhang, L., Wu, X., Luo, D., 2015. Human activity recognition with HMM-DNN model. In: International Conference on Cognitive Informatics & Cognitive Computing, Beijing, China.
Zhang, X.Y., Yin, F., Zhang, Y.M., Liu, C.L., Bengio, Y., 2018b. Drawing and recognizing chinese characters with recurrent neural network. IEEE Trans. Pattern Anal. Mach. Intell. 40 (4), 849–862.
Zhao, X., Naguib, A.M., Lee, S., 2014. Kinect based calling gesture recognition for taking order service. In: 23rd IEEE International Symposium on Robot and Human Interactive Communication, Edinburgh, Scotland, UK.
Zhao, Y., Qiao, Y., Yang, J., Kasabov, N., 2015. Abnormal activity detection using spatio-temporal feature and laplacian sparse representation. In: International Conference on Neural Information Processing, Switzerland.
Zhu, X., Liu, Z., 2011a. Human behavior clustering for anomaly detection. Front. Comput. Sci. 5 (3), 279–289.
Zhu, X., Liu, Z., 2011b. Human behavior clustering for anomaly detection. Front. Comput. Sci. China 5 (3), 279–289.
Zhu, G., Xu, C., Huang, Q., Rui, Y., Jiang, S., Gao, W., Yao, H., 2009. Event tactic analysis based on broadcast sports video. IEEE Trans. Multimed. 11 (1), 49–67.
Zhu, G., Zhang, L., Shen, P., Song, J., 2017. Multimodal gesture recognition using 3D convolution and convolutional LSTM. IEEE Access PP (99), 1-1.
Zitouni, M.S., Bhaskar, H., Dias, J., Al-Mualla, M., 2016. Advances and trends in visual crowd analysis: a systematic survey. Neurocomputing 186, 139–159.

Ms. Chhavi Dhiman received B.Tech. from Indira Gandhi Technical University for Women, New Delhi, India in 2011, and M.Tech. from Delhi Technological University, New Delhi, India in year 2014. She is currently working towards the Ph.D. degree in the Bio-Metric Research Laboratory, Delhi Technological University, New Delhi. Her current research interest includes deep learning, pattern recognition, human abnormal action identification and classification.

Dinesh Kumar Vishwakarma received the B.Tech. degree from Dr. Ram Manohar Lohia Avadh University, Faizabad, India in 2002, the M.Tech. degree from the Motilal Nehru National Institute of Technology, Allahabad, India in 2005, and the Ph.D. degree from Delhi Technological University, New Delhi, India, in 2016. He is currently an Associate Professor with the Department of Information Technology, Delhi Technological University, New Delhi, India. His current research interests include Human–Computer Interaction, Machine Learning, Deep Learning, Pattern Recognition, Computer Vision, Human Action and Activity Recognition, Sentiment Analysis, Spam Analysis, and Fake News Analysis. He is also reviewer of various journals of IEEE/IET, Springer, and Elsevier.