Automated human behavior analysis from surveillance videos: a survey

D. Gowsikhaa · S. Abirami · R. Baskaran

Abstract   With increasing crime rates in today's world, there is a corresponding awareness for the necessity of detecting abnormal activity. Automation of abnormal Human behavior analysis can play a significant role in security by decreasing the time taken to thwart unwanted events and picking them up during the suspicion stage itself. With advances in technology, surveillance systems can become more automated than manual. Human Behavior Analysis although crucial, is highly challenging. Tracking and recognizing objects and human motion from surveillance videos, followed by automatic summarization of its content has become a hot topic of research. Many researchers have contributed to the field of automated video surveillance through detection, classification and tracking algorithms. Earlier research work is insufficient for comprehensive analysis of human behavior. With the introduction of semantics, the context of a surveillance domain may be established. Such semantics may extend surveillance systems to perform event-based behavior analysis relevant to the domain. This paper presents a survey on research on human behavior analysis with a scope of analyzing the capabilities of the state-of-art methodologies with special focus on semantically enhanced analysis.

Keywords   Activity recognition · Motion detection · Motion tracking · Object classification · Video surveillance

1 Introduction

In earlier days, the crime level and population was comparatively lesser than in the modern world. Surveillance required was concentrated only on smaller places, and security guards

D. Gowsikhaa · S. Abirami (✉) · R. Baskaran
Department of Information Science and Technology, College of Engineering, Anna University,
Chennai, India
e-mail: abirami_mr@yahoo.com

D. Gowsikhaa
e-mail: gowsikhaa@gmail.com

were stationed at these places. With rapid increase in crime rate,[1] observing all the activities in public places without surveillance cameras would need security guards almost everywhere. This led to additional security in the form of video surveillance. Video surveillance is done by installing CCTV or IP cameras, at places to be secured, without security guard presence at the place under surveillance. These videos can be manually monitored through video walls. Video surveillance acts as a security mechanism to monitor areas prone to issues like theft, drug trafficking, border trespassing, vandalism, fights, etc. It may also be used in home-care systems (Lao et al. 2010) for monitoring children and old people, or patients in hospitals (Chen 2010). The third kind of usage is for pattern analysis (Leykin and Tuceryan 2007) where, people behavior and shoppers' buying behavior are collected and patterns found. If an area under surveillance has many cameras, it is tedious to monitor all of them manually. It is said that manual supervisors tend to miss some activities when they continuously monitor video walls. This led to the transition of manual video surveillance to automated video surveillance.[2] Automation reduces man power wasted in manual monitoring and subsequent human errors, thus reducing the cost of employment, reducing the cost of storage and leading to a fool-proof monitoring.

Cameras are the eyes of a video surveillance system. The placement of cameras (see footnote 1) at perfect spots helps in viewing objects without occlusion. IP cameras are used for applications which require videos with higher quality. Video analytics (see footnote 1) is used for *optimizing storage* as well as *analyzing human behavior*. Since storing all the videos requires a lot of memory space, storage can be optimized by not recording static scenes. This is done by triggering the video record sequence only when there is motion in a scene, thereby reducing cost of storage.

Human behavior cannot be determined directly from an input video. A video must be analyzed in order to find out what is in it. It must then be processed to obtain desired results. The basic steps involved in processing a video are video segmentation, Motion detection (background modeling, foreground segmentation), object classification, motion tracking and finally, activity recognition as shown in Fig. 1.

A human can intuitively find out objects in motion just by watching a video. To make a machine do it, the video must first be split into a sequence of images. Motion detection helps a machine to find out which part of the video is moving. The Motion detection module includes background modeling, foreground modeling and foreground processing. Background modeling extracts a reference background image from a sequence of frames. Foreground modeling segments out a noisy image with moving objects in it. Foreground processing removes noise from the result of the previous stage and results in a blob representing the object's silhouette.

In object classification, the blob in the foreground is categorized into object types. In motion tracking, an object's movement is tracked from one frame to another. These phases, i.e. motion detection, object classification and motion tracking form the "building blocks" of human behavior analysis. With the results obtained from these, a behavior recognition methodology can be formulated using domain specific poses and semantics. A generalized approach to human behavior recognition can be designed for research purposes. Systems which are to be used commercially are preferred to be domain specific. For example, system at a railway station for detecting suspicious activities needs to detect activities like fighting, got hurt, stealing, running, etc.

---

[1] IP Video Market Information. http://ipvm.com

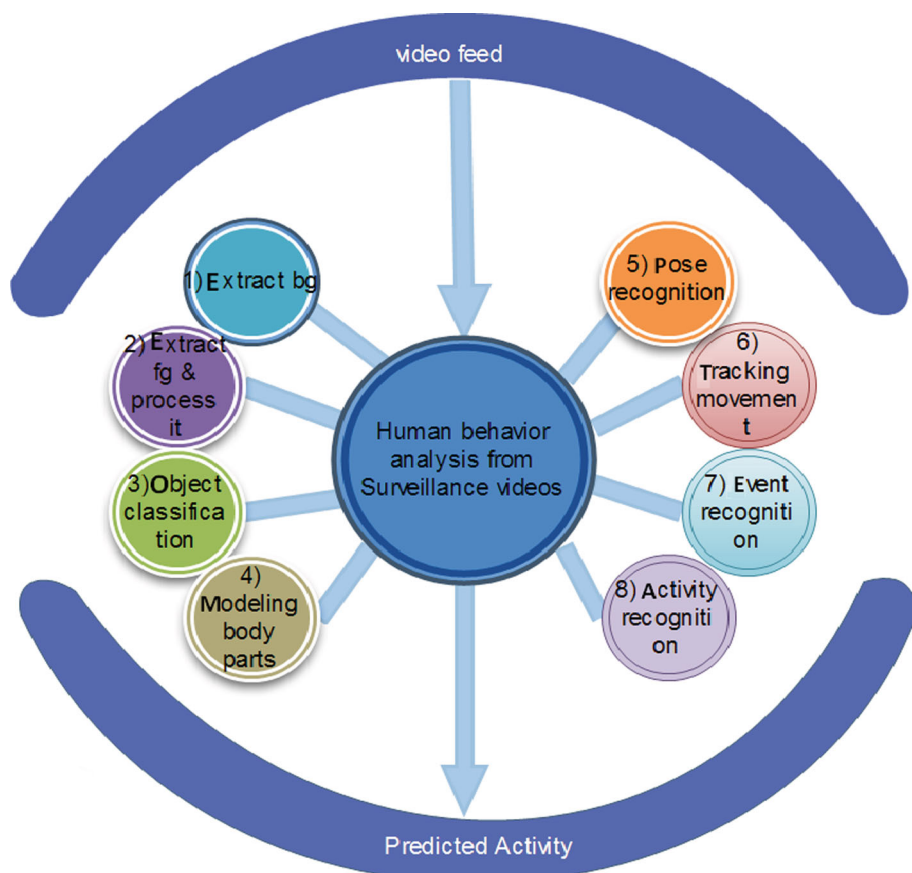[2] Web's Premier Article Directory. http://www.articledashboard.com

Fig. 1 A map representing the human activity prediction architecture

The types of behaviors that can be analyzed are represented in Fig. 2. They are single person activities, single person interacting with an object (vehicles, sports gears, household objects, etc), two person interactions and multiple people interactions.

Single person and multiple person activity analysis do not vary much with respect to the domain. Examples for this category are walking, standing, sitting, fighting, handshaking, etc. Group behavior analysis with varying number of group members is handled in Lin et al. (2010). Human—non human interactions and human—environment interactions relies on domain knowledge. For example, stealing, object exchange, trespassing and vandalism are more domains specific.

The detailed organization of this paper is presented in Fig. 3,

Section 2 presents a detailed survey of the low-level techniques. Section 3 presents a survey of the high-level processing methods. Section 4 presents the challenges in this area.

2 Literature survey: low-level processing techniques

In this survey, the techniques forming the basic building blocks of human behavior recognition process, which do not require any kind of semantics, are classified under low-level

Single Person - no / less context oriented

- Tracking the position of a single person through scenes
- Single person activity recognition
- Single person activity recognition with respect to context

Human - nonhuman interactions - more context oriented

- Based on Human – nonhuman distance
- Interaction with respect to context and semantics

Human - Facility interactions - more context oriented

- Location of human in a scene with respect to the context
- Interaction with respect to semantics

Multi-person - less context oriented

- Tracking the position of multiple person
- Identifying multiple person interactions
- Identifying multiple person interactions with respect to context

Group/crowd behavior - context oriented

- Simple tracking
- Tracking and behavior analysis with repect to context

Fig. 2 Categorization of human behaviors

| I Introduction | → | A. Mind map for system classification
B. Basic System Architecture |

| II Literature Review Low-Level Technologies | → | A. Motion Detection
B. Object Classification
C. Motion Tracking |

| III Literature Review High-Level Technologies | → | A. Pre-processing
B. Human Behavior Recognition and analysis
C. Comparison Tables |

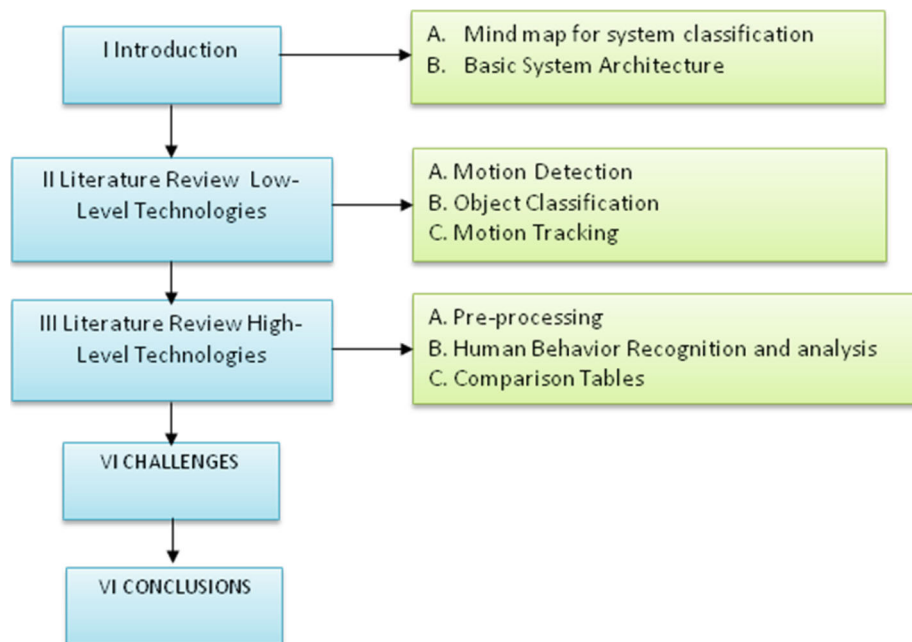| VI CHALLENGES |

| VI CONCLUSIONS |

Fig. 3 Paper organization

processing techniques. The low-level techniques consist of background modeling, foreground detection, object classification and simple motion tracking.

2.1 Motion detection methods

In motion detection, only the moving objects are segmented out from the original frame. To separate the foreground from the background, a reference frame is extracted from the video. From the current frame, the background/reference frame is subtracted to segment out moving objects. The required moving objects are the 'foreground' and the unwanted background information is the 'background'.

背景帧用于分割出移动物体

Background modelling can be done using medium complexity techniques such as approximate mean and approximate median. A low complexity technique for background modelling is implemented by simply treating the first frame as the background frame for the entire video. Another such simple method is to treat the previous frame of an image sequence as the background frame.

Common approaches in foreground modelling are background subtraction, temporal differencing and optical flow techniques. In background subtraction, the pixel by pixel difference is taken between the current frame and the reference frame. Background Subtraction techniques (Javed et al. 2002; McIvor et al. 1999; Li et al. 2010) are most frequently used. A simple Background subtraction method works only for videos with static backgrounds. Adaptive background subtraction method makes use of dynamic background updation. It is more robust to changing backgrounds and environmental conditions, but consumes more memory for storing intermediate values. This method can be enhanced further by adding a *scene classification* (Huang et al. 2011) module.

Temporal differencing (Mahadevan and Vasconcelos 2009) makes use of the pixel-wise difference between two to three consecutive frames in an image sequence to extract moving regions. The background extraction process is not used here. Differencing techniques generally are poor in extracting all relevant pixels, e.g., there may be cavities left inside moving entities. This happens when the object is moving slowly, or when it has uniform textures, in which case it might need additional *filtering algorithms*. In Xiaofeng et al. (2010) *Three frame differenced image* is used instead of using the current frame directly. Optical flow method (Shafie et al. 2009; Denman et al. 2009) depends upon the distribution of apparent velocities of movements of brightness patterns in an image and gives information about the spatial arrangement of the objects viewed. Optical flow methods are computationally complex and are used for complex dynamic image analysis.

2.2 Object classification methods

To track an object and process it further, it must be classified correctly. Motion detection phase results in foreground blobs. During classification, Yunus (2009), Roach et al. (2001), the foreground blob is processed, classified and labelled. The accuracy of the classification phase depends upon the features extracted during the foreground blob processing. In simple systems, there may be a binary classification where a blob can be classified either as a person blob or a non-person blob. Advanced methods (fuzzy classifiers) may classify a blob as a part of two or more classes.

Object classification is frequently done using advanced techniques like training models such as neural networks. Objects are classified based on features like shape of the object or the type of motion. At a simpler level, template matching technique can be used. In template

matching, the current foreground image of the object is compared with a set of templates and objects are classified based upon similarity measures. If the template and the object in foreground match or are similar, the appropriate object type is said to have been found. Object classification can be done either after or before background subtraction (Ogale 2006). When classifying directly from the video input, edge detection algorithms are used for finding out edges of the current frame and then separating the moving part from it through background subtraction.

The two important categorizations (Ko 2008) of object classification algorithms are Motion-based classification methods and Shape-based classification methods. Motion-based methods require parameter estimation for any motion based parameter. For example, a human and a non-human can be differentiated based on its movement patterns. In Yoshimitsu et al. (2010), an object that does not move for 'n' frames is considered to be non-human. Shape-based method uses the features that best represent the object's shape (Lao et al. 2010) for classification, but they are sensitive to noise. The commonly used features are colour, texture, pixel coordinates, etc. Object classification can be done using Support Vector machine (SVM), Hidden Markov Model (HMM) or a neural network.

Techniques for classifying a human and a non human (Rahimt et al. 2010) vary. A human and a non-human can be easily differentiated with the help of its contour shape. Contours can be obtained by two ways. In the first method, after motion detection, the foreground image is binarized (Xiaofeng et al. 2010) and then using neighbour masking algorithm (Liao et al. 2010), contour points are formed for the moving object. In the Second method (Tabb et al. 2004), the input frame is processed directly using an edge masking algorithm, and then the image with edges is used as the input for motion detection phase. The motion detection phase directly results in the contour of the foreground blob. When choosing shape control points from a contour, normalization can be done. Normalization scales up the distances as well as the number of contour points. So, this method can adapt itself for grownups as well as children. The choice of a pre-defined number of coordinates results in uniformity of the features. Rougier et al. (2011), uses a combination of the two methods mentioned above to get a better result. Here the foreground silhouette is combined with the canny edge image.

2.3 Motion tracking methods

After object classification, the movement of an object from one frame to another must be tracked in order to find out the temporal consistency of an object and to know whether the object under surveillance is new to the scene, or whether it has been previously seen. Tracking methods can be region based, feature based, contour based, part based, model based or combinations of these techniques. Tracking a particular person from a crowd or tracking a particular vehicle in a highly populated transit scene is difficult since the objects in background are similar to the object being tracked. Unique co-ordinates can be calculated from the contour, and if the number of coordinates drastically increases or decreases, *occlusion* is predicted.

A region-based technique (Brox et al. 2010) considers variation in the image regions corresponding to the moving objects. A contour-based technique (Techmer 2001; Lee et al. 2006) tracks only the contour of the object instead of the whole object. A feature-based technique (Park et al. 2005; Chen et al. 1996) extracts features from a person and uses them to track the body parts in the subsequent frames. Part based and shape control based techniques fall under the feature-based category. Part based technique (Chen 2010) detects salient points from interested objects and then decomposes the object into a combination of these salient points. Shape Control Point (Kim et al. 2011) is a technique that is highly efficient

in tracking objects when occluded and tracking when two objects cross each other. In the model-based approach (Lao et al. 2010), a human body model is constructed and tracked. The most common modeling technique used is the stick figure modeling, also known as the skeleton model. Apart from these five techniques, the Hybrid tracking method (Brox et al. 2010) uses a combination of two or more of these techniques.

A simple motion tracking algorithm tracks the displacement of a human. Trajectories can be calculated from the displacement of the object's centre of his/her foot (Yoshimitsu et al. 2010). Motion tracking in an *advanced level*, can be divided into three parts, finding out the current pose by body pose modelling, occlusion handling and estimating the trajectories. Applications which require recognition of simple poses, can just estimate the position of head, hands and legs to form a basic skeleton of the human (Lao et al. 2010). 2D modelling (Lao et al. 2010) is sufficient for this method. For complex applications, along with head, hands and legs, the joints (like ankles, knees) and neck can be identified. This requires 3D modelling (Brown and Capson 2011) for representation. For vehicles, the shape of vehicles can be used to handle occlusion. The shape of the vehicle can be used to regenerate (Feris et al. 2011) the missing parts of the vehicle.

3 High-level processing techniques

Motion detection methods (Kiryati et al. 2008) were used to optimize the storage space for later investigations. Video analytics is now an intelligent system and infers annotations from video sequences. Behaviour Analysis aims to extract information and process it to get a better understanding of the activities. High-level processing techniques include phases that make use of semantics instead of using the 'basic building blocks' as such.

3.1 Pre-processing

Avoiding false recognition of activities are as important as detecting activities correctly. In phases like object classification, pose recognition and activity recognition, semantics are used to increase accuracy and predict activities that may happen in future. Semantics can be included even at the tracking phase by including occlusion handling module and body part modelling relevant to the domain.

3.2 Human behaviour recognition and analysis

Actions may be poses or events. Different ways have been suggested to classify activities. The simplest classification is normal or abnormal (Kiryati et al. 2008). In Foroughi et al. (2008) activities are classified into normal, unusual and abnormal. Park and Aggarwal (2004), classifies the activities as positive, neutral and negative activities. In Lin et al. (2010), the group activities are classified into symmetric and asymmetric activities. Lao et al. (2010), recognizes a three person activity by analysing three combinations of two person activities.

Poses are the basic actions that can be directly identified. After modelling the body parts, poses can be estimated with lesser semantics through a classifier or by a simple rule based methodology. In Lao et al. (2010), the coordinates of the body parts or the generated skeleton is given as input to a classifier and then the poses are classified into sit, stand, lie, squat and relax. Another method is to draw a bounding rectangle around the object for which the pose must be identified. From the bounding box, the height and width of the object can be calculated and ratio between them is compared with a range of threshold values to find out
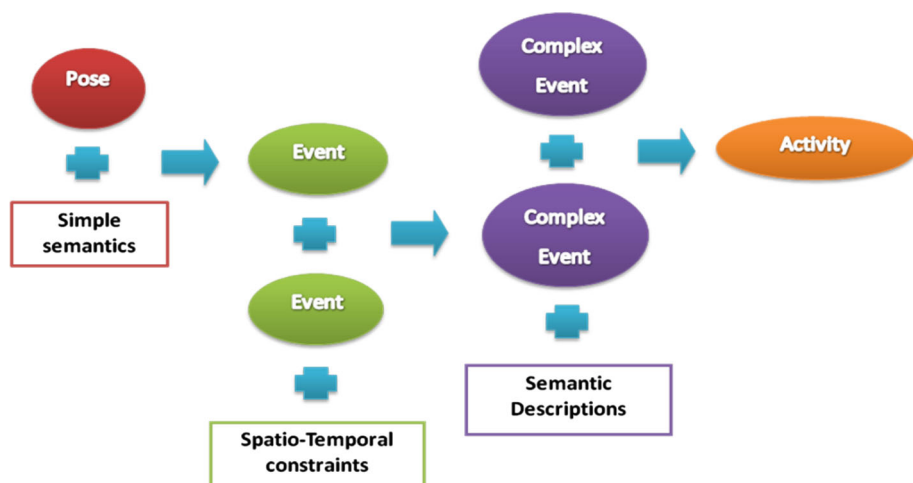
Fig. 4 Relationship between activities, events and poses

the exact pose (Lao et al. 2010). To get a precise pose, instead of using the bounding box to find the ratio, contour or convex hull of the object can be used. In Foroughi et al. (2008), instead of using the bounding box for predicting the action of a person, an approximated ellipse of the person is used. It is shown that, the box, bounds the human along with the objects possessed by him/her, so it may result in a false body ratio. In Park and Aggarwal (2004), body parts are first detected and then ellipses are constructed for each and every body part separately. Finally three convex hulls (head, body and legs) are formed for a single person.

Events that can be identified from the basic pose are stand, walk, and run. Speed is used as a trait to find out sub events from poses. When the person's speed is constant and slow and the current pose is stand, it can be concluded that the person is walking. Fig. 4 represents the relationship between the constraints, actions and activities.

Complex events are identified from poses using spatio-temporal constraints (Bremond et al. 2005). A Spatial constraint is the distance between objects and the location of the object. Temporal constraints deal with the duration of/between actions. If the distance between two objects decreases from frame to frame then it means that the objects are 'moving closer/approaching'.

Once the poses and events are found, activity prediction is done using a classifier with the help of semantic descriptions. Semantics give meaning to the actions. The preceding actions and the current action can be correlated to obtain a result. Different aspects of an activity (Huang 2011) can be used through an ADD/OR tree method to predict the activity. To predict the activity of a single person, Liao et al. (2010) uses information such as gaze direction, spatio-temporal activity of a person and then extracts his/her behaviour from the video.

Instead of using 2D silhouettes for activity recognition, Li et al. (2010) has used 3D points. The comparative results show that activity recognition with 3D points give better prediction results. A homography mapping technique (Lao et al. 2010) is used for mapping the locations of all the landmarks in the area under surveillance. Similarly, in Yoshimitsu et al. (2010) location maps are used for representing the floor, fence, wall, obstacle and gate. Such methods must be used for scenes with absolutely static backgrounds.

As stated in Table 1, different activity recognition schemes use different poses and spatio-temporal constraints according to the domain chosen. The actions such as walk, stand, sit, move closer to, move away from and stay at are common for most of the systems.

The descriptions for the input of an activity-classifier depend upon the domain chosen. The semantic descriptions play a vital role in activity recognition. Table 2 summarizes some of the semantic descriptions that may be verified with a domain expert.

Semantics play a vital role even for person-non person interactions. Table 3 explains the three types of basic actions in predicting the activities of a person with respect to an object. (i) Stealing an object from another person or stealing an object from the background. (ii) Abandoning an object into the background. The person who has abandoned an object can be found out by locating the person who has interacted with the object (Yoshimitsu et al. 2010). (iii) Exchanging objects (between two persons). If it is a direct exchange, both the persons involved in the action are aware of what is happening and they will contact each other at some point of time. If it is indirect, both the persons will be aware of what is happening and they will not contact each other. Person A will abandon the object and after a certain number of frames, person B steals/picks up the abandoned object. If a new object remains in the background for 'n' number of frames, then it is updated back to the background (Yoshimitsu et al. 2010).

Visual cues can be used for predicting the behavior of a human being. A system can learn visual cues related to emotions by recognizing certain regions of face or body parts which identify the emotions. Temporal segmentation is a sensitive process. The correct segmentation of each and every atomic action will decide the type of activity predicted. In Robertson security submission (2006), the magnitudes of data points are plotted along a time series, and the local minima are treated as the break point for segmenting atomic actions. This segmentation technique assumes that a human will come to a resting position after each and every atomic action. The resting position represents the local minima of the plot. Table 4 evaluates the performance of the systems by comparing the datasets used, accuracy rates and time complexities.

Table 5 summarizes some of the significant work in the area of human behavior analysis. Some of the works are more research oriented and complex, and thus, not suitable for commercial implementations. Most of the research oriented work concentrate on a single technique. Domain oriented work use a combination of these techniques. For constructing a complete system, a combination of methods can be used and modified according to the domain.

4 Challenges in human behavior analysis

Predicting activities from surveillance videos is easy for a human. But automatic activity prediction involves difficulties at almost every stage. Some of them are mentioned in this section.

- **Cavities:** Acquiring the foreground without any *cavities* (Huang 2011) is a challenging task during the foreground segmentation phase.
- **Human body modeling:** Similarly, *modeling human body parts* (Lao et al. 2010) help in tracking the human motion in successive frames. Incorrect identification of body part modeling will result in faulty pose prediction.

Table 1 Comparing Activities Recognized in Different Works

Ref. no.	Attributes used	Poses recognized	Spatio-temporal relationships	Activities recognized	Applications used	Classifier used
Lao et al. (2010)	Human	Pointing, lying, squatting, raising hands over head	After, meets, during, finishes, overlaps, equal, starts	Sitting in couch, playing in piano, robbery event	Home care monitoring system, bank robbery scenario	CHMM
Yoshimitsu et al. (2010)	Pedestrian, object	Standing, crawling, lying	Not specified	Illegal tailgating, loitering, suspicious pedestrian interaction, abandoned objects	Railway station	Rule-based approach
Foroughi et al. (2008)	Human	Lie, walk, sit, limp, fall, run, stumble, bend	Not specified	Not specified	Fall detection	SVM
Robertson security submission (2006)	Agent	Walking, running, stopped	Nearside, far side, left to right, away, towards	Combination of poses and relationships	Urban surveillance	HMM
Peursum et al. (2005)	Actor	Drink, read, type, walk, sit down, stand up	Not specified	Gets tea bag, open tap, sitting in chair, drink tea, load printer, etc	Office environment	HMM
Bremond et al. (2005)	Equipment, individual	Not specified	Far from, stays at, moves away, moves close, before, after	Blocking, fraud, fight, vandalism, overcrowding, jumping	Metro station, bank, lock chamber, apron monitoring system	AND/OR tree
Lin et al. (2010)	Human	Not specified	Not specified	Ingroup, walktogether, runtogether, ignore, approach, split, chase	Outdoor	AHMM
Park and Aggarwal (2004)	Human	Move forward, move backward, stretch, stay stationary, raise lower, withdraw	Not specified	Approaching each other, departing each other, pointing, shaking hands, hugging, standing hand-in-hand, punching, pushing, kicking	indoor	Dynamic Bayesian Network (DBN)

Table 2 Sample of semantic descriptions used in the state of art

Ref. No.	Activity	Descriptions
Yoshimitsu et al. (2010)	Normal	Pose of a pedestrian is crawling + he is located in bench
	Abnormal	Pose of pedestrian is crawling + he is located in floor
	Illegal tailgating	Pose of pedestrian is crawling/lying down + location is gate
	Suspicious behavior	If there is contact with other pedestrians before a particular time when a certain pedestrian is lying down on the floor and it crowds
Foroughi et al. (2008)	Walk	Walking naturally a few meters
	Run	Jogging or fast walking
	Stumble	Unable to keep balance or symmetric and synchrony of movement
	Limp	Suffering from gait abnormality
	Forward fall	Forward fall on knees, chest or arms
	Backward fall	Backward fall caused by slipping
	Sideway fall	Lateral fall to right or left on legs
	Bend down	Bending down, catching something on the floor and then rising up
	Sit down	Sitting down and then standing up
	Lie down	Lying down on the floor
	Inexplicable	Nearside pavement, walking Road, walking + Nearside pavement, walking
Robertson security submission (2006)		
Lin et al. (2010)	InGroup	The people are in a group and not moving very much
	WalkTogether	People walking together
	Fight	Two or more groups fighting
	RunTogether	The group is running together
	Ignore	Ignoring of one another
	Approach	Two people or groups with one (or both) approaching the other
	Split	Two or more people splitting from one another
	Chase	One group chasing another
Park and Aggarwal (2004)	Approach	To draw closer to
	Depart	To go away from
	Point	To indicate the position or direction of especially by extending a finger
		To clasp usually right hands by two people (as in greeting or farewell)
	Shake hands	Stand with hands clasped (as in intimacy or affection)
		To press tightly especially in the arms
	Standing hand-hand	To strike with a forward thrust especially of the fist
	Hug	To press against with force in order to drive or impel
	Punch	To strike out with the foot or feet
	Push	
	Kick	

- **Handling occlusions:** *Occlusion handling* (Brox et al. 2010), must be done to track through people even when they cross each other, or when a person disappears from frame and reappears after a few seconds.
- **Scene classification:** In the activity recognition phase for an environment with dynamic backgrounds, a better understanding of the *background scene* (Huang et al. 2011) will

Table 3 predicting the activities of a person with respect to an object

Activity	Attributes	Contact with person	Context Awareness
Stealing from background	Person 1, object	–	–
Stealing from a person	Person 1, person 2, object	Direct	Person 1/person 2
Abandoning	Person 1, object	–	–
Exchanging	Person 1, person 2, object	Direct indirect—has abandoning followed by stealing from background	Person 1 and person 2

be useful. Even if the background scene is static, automatic background classification (example, indoor/outdoor) will make the system generic.

- **Person identification:** Recognizing a previously seen person correctly when he/she re-enters the scene, instead of treating him/her as a new person is also important.
- **Techniques for activity perception:***Temporal segmentation*, *semantic descriptions* for multiple people interaction patterns is to be considered during the activity perception stage.
- **Cameras revisited:** Usage of a single camera is simple, but has drawbacks of occlusions due to the restricted view of a single camera. Multiple camera usage eliminates such a drawback, but has its own defects. Apart from the algorithms and techniques used for behavior analysis, the usage of appropriate type of cameras for applications and the number of cameras used also has an impact on the efficiency of the work.
- **Modeling scenes:** Most of the systems model the scenes in 3D. The 2D images captured from a single camera will be mapped to a 3D world for better visualization and for later analysis. Recently, multiple cameras (Chang et al. 2010; Gandhi and Trivedi 2006) are used for tracking the same object from different perspectives. For a method which uses multiple cameras, a reference point is needed in the scene, so that the different views can be correlated with respect to a common point.
- **Standardization:** There is no uniform standard for defining the human poses. The presence of such a standard would make the result comparisons easy and help in the uniformity of poses and semantic descriptions.
- **Domain specificity:** The background methodologies and the types of actions recognized make the works more domains specific. Domain specificity can be due to the following factors,

  - Using ground measurements (Lao et al. 2010)
  - Using landmarks for combining views of multiple cameras
  - Requirement of detecting domain specific activities

A background subtraction method which can handle all the issues like environmental changes in outdoor, indoor illumination changes will become more complex and can impact the performance of the system. A few combinations of these issues can be used by predicting the possibilities of environmental changes that could happen in the area under video surveillance. Thus domain specificity will lead to system accuracy for the required purpose.

Each stage is interdependent. Handling all these challenges result in added complexity. Some of these challenges that fall under the 'low-level technique' category have already researched well. There are still a lot of unhandled areas that involve semantics. Usage of standardized and fine-tuned semantics will lead to better performance and system understanding.

Table 4 Performance evaluation

Ref. No.	Datasets	Subject details	Accuracy in %	Field
Huang (2011)	–	–	Video sequence IR–78.64 (Precision) 80.25 (Recall) MR–86.65 (Precision) 92.46 (Recall) SC–70.44 (Precision) 71.13 (Recall) WS–86.84 (Precision) 86.73 (Recall)	Motion detection
Tabb et al. (2004)	Computer generated	CG human, horses, dogs moving in different directions (training)	Human–81 (average confidence value) Nonhuman–65 (average confidence value)	Human—nonhuman classification
BenAbdelkadert et al. (2002)	database of fronto-parallel sequences	Female–7 Male–38	Height & Stride–49 Stride–21	Person identification (object classification)
Bremond et al. (2005)	Barcelona and Brussels surveillance videos (both live and recorded)	–	85	Behavior recognition
Foroughi et al. (2008)	Own dataset	24 (Different heights/gender/age group–20 to 30)	89.49	Fall detection
Seki et al. (2000)	Own dataset	–	–	Motion detection with dynamic background
Park and Aggarwal (2004)	Own dataset	12 (in 6 pairs)	78	Recognition of positive/ negative/ neutral interactions between two persons
Lao et al. (2010)	–	Home-care scenario—6 persons (with different gender, height, age and clothes)	86.25	Robbery event recognition, home-care monitoring
Lin et al. (2010)	BEHAVE dataset	–	Miss–12.88 False alarm–2.3	Group event detection
Chang et al. (2010)	BEHAVE dataset	–	–	Fight detection in a gang

Table 4 continued

Ref. No.	Datasets	Subject details	Accuracy in %	Field
Kiryati et al. (2008)	Own dataset	–	–	Storage optimization and abnormal activity detection
Peursum et al. (2005)	Own dataset	–	Precision–70.18 (per frame)	Action recognition
			Recall–68.78 (per frame)	
Li et al. (2010)	Own dataset	–	86.8	Action recognition for tennis

Table 5 Comparison of Low-level and high-level techniques in human behavior analysis

Ref. No.	Publica-tion	Motion detection	Object classification	Object tracking	Interaction prediction	Testing environment
Liang et al. (2009)	2009	Subtracting the average distribution of image vectors with the observed image vectors, the changes in background are learnt. Assumptions: variations in an image occur only at the region of the moving object	–	–	–	Outdoor
Javed et al. (2002)	2002	Differentiating and removing a person's silhouette from simple noises, shadows and noises due to illumination by filtering the foreground image with the help of type of the boundary pixels using color and gradient information. Assumptions: noisy blob is assumed to have diffused highlights, smaller size and unclear edges	–	–	–	Indoor & Outdoor
Mahadevan and Vasconcelos (2009)	2009	Moving object segmentation with the help of temporal differencing, noise removal and cavity removal	–	–	–	Indoor & Outdoor

Table 5 continued

Ref. No.	Publica-tion	Motion detection	Object classification	Object tracking	Interaction prediction	Testing environment
Shafie et al. (2009)	2009	Background modeling sing rapid matching and accurate matching using moving average method	–	–	–	Indoor & Outdoor
Tabb et al. (2004)	2004	Active contour model (ACM) for detecting moving objects	Classifying the objects into human/non-human with the help of the shape of a scale invariant contour system using artificial neural networks	–	–	Outdoor
Kim et al. (2011)	2011	Background generation, updation and differencing using Block matching algorithm for motion detection	–	Motion tracking during occlusions over similar objects with the help of shape control points	–	Outdoor
Kiryati et al. (2008)	2008	Uses inter and intra frames for representing motion in video	–	Uses regional information by dividing frame to blocks and direction of motion	Recognizing abnormal activities from the video using the macroblock motion vectors	Outdoor
Park and Aggarwal (2004)	2004	–	–	Blob tracking using pixel color and position	Behavior analysis using body part modeling, operation triplets and verbal semantic descriptions	Indoor

Table 5 continued

Ref. No.	Publica-tion	Motion detection	Object classification	Object tracking	Interaction prediction	Testing environment
Robertson security submission (2006)	2006	–	–	Color based tracking	A probabilistic activity estimation by a rule-based reasoning approach using gaze direction, spatio temporal actions and behavior directly obtained from the video	Outdoor
Yoshimitsu et al. (2010)	2010	STMRF model	Objects are classified as non-humans if it is static in the video for 'n' frames	STMRF model	Analyzing behaviors of people near railway stations with respect to their interaction, location and type of action	Indoor
Lao et al. (2010)	2010	Not specified	–	Not specified	Analyzing behavior of people by modeling body parts by constructing skeletons and estimating pose by a classifier with the help of homography mapping	Indoor
BenAbdelkadert et al. (2002)	2002	Not specified	–	Tracking a person in consecutive frames by simple spatio temporal coherence Assumptions: Minimum height when legs are apart during walk and maximum height when legs are closer to each other during walk	Analyzing gait with the help of a person's smooth stride cycle with respect to his/her height	Outdoor

5 Conclusions

Even though much research has done in the area of Human Behavioural Analysis, issues and challenges still prevail. Recognizing the background scene as indoor and outdoor is not sufficient, it can be extended upon by further scene classification during pre-processing. Motion detection and tracking methodologies can be significantly enhanced using temporal segmentation and semantic descriptions. The classifier chosen for activity recognition has impact on the type of activity predicted. Most of the algorithms do not efficiently handle multiple object interaction recognition. There is active research in progress in this area, with the increasing need for methods to find out the interaction between groups of individuals.

Activity recognition is currently done for events which take place continuously; it can be extended to discrete event recognition methods. There can be two different scenarios with respect to discrete activities, with two persons far away from each other but involved in an activity (e.g. waving to each other) and the other one being, a single person performing the same activity repeatedly with pauses. Semantic descriptions play a vital role in finding the type of activity.

One of the areas which need further concentration is defining a standardized format for semantic descriptions of different activities.

The methodologies in this survey are mainly classified with respect to semantics involved, thus dividing the whole process into high level and low level. This survey reviewed a few algorithms in the domain of Human behaviour analysis including both the low-level and high-level techniques in order to get a better understanding of the state of art techniques.

References

BenAbdelkadert C, Cutler R, Davist L (2002) Person identification using automatic height and stride estimation. In: Proceedings of 16th international conference on pattern recognition, pp 377–380

Bremond F, Thonnat M, Zuniga M (2005) Video understanding framework for automatic behavior recognition. Behav Res Methods 38(3):416–426

Brown JA, Capson DW (2011) A framework for 3D model-based visual tracking using a GPU-accelerated particle filter. IEEE Trans Vis Comput Graph 80(1):60–80

Brox T, Rosenhahn B, Gall J, Cremers D (2010) Combined region and motion-based 3D tracking of rigid and articulated objects. IEEE Trans Pattern Anal Mach Intell 32(3):402–415

Chang M-C, Krahnstoever N, Lim S, Yu T (2010) Group level activity recognition in crowded environments across multiple cameras. In: Seventh IEEE international conference on advanced video and signal based surveillance, pp 56–63

Chen M (2010) Long term activity analysis in surveillance video archives

Chen Y-K, Lin Y-T, Kung SY (1996) A feature tracking algorithm using neighborhood relaxation with multi-candidate pre-screening. In: IEEE international conference on image processing, pp 513–516

Denman S, Fookes C, Sridharan S(2009) Improved simultaneous computation of motion detection and optical flow for object tracking. In: Digital image computing: techniques and applications, pp 175–182

Feris R, Petterson J, Siddique B, Brown L, Pankanti S (2011) Large-scale vehicle detection in challenging Urban surveillance environments. In: IEEE workshop on applications of computer vision, pp 527–533

Foroughi H, Yazdi HS, Pourreza H, Javidi M (2008) An eigenspace-based approach for human fall detection using integrated time motion image and multi-class support vector machine. In: IEEE 4th international conference on intelligent computer communication and processing, pp 83–90

Gandhi T, Trivedi MM (2006) Panoramic appearance map (PAM) for multi-camera based person re-identification. In: Proceedings of the IEEE international conference on video and signal based surveillance, p 78

Huang K, Tao D, Yuan Y, Li X, Tan T (2011) Biologically inspired features for scene classification in video surveillance. IEEE Trans Syst Man Cybern B Cybern 41(1):307–313

Huang S-C (2011) An advanced motion detection algorithm with video quality analysis for video surveillance systems. IEEE Trans Circuits Syst Video Technol 21(1):1–14

Javed O, Shafique K, Shah M (2002) A hierarchical approach to robust background subtraction using color and gradient information. In: IEEE proceedings of workshop on motion and video computing

Kim T, Lee S, Paik J (2011) Combined shape and feature-based video analysis and its application to non-rigid object tracking. Inst Eng Technol Image Process 5(1):87–100

Kiryati N, Riklin TR, Ivanchenko Y, Rochel S (2008) Real-time abnormal motion detection in surveillance video. In: IEEE 19th international conference on pattern recognition, pp 1–4

Ko T (2008) A survey on behaviour analysis in video surveillance applications. In: 37th IEEE applied imagery pattern recognition workshop '08, pp 1–8

Lao W, Han J, deWith PHN (2010) Flexible human behavior analysis framework for video surveillance applications. Int J Digit Multimed Broadcast, Article ID 920121, 1–9

Lee BH, Choi I, Jeon GJ (2006) Motion-based moving object tracking using an active contour. In: International conference on acoustics, speech, and signal processing, pp 649–652

Leykin A, Tuceryan M (2007) Detecting shopper groups in video sequences. In: IEEE conference on advanced video and signal based surveillance, pp 417–422

Li W, Zhang Z, Liu Z (2010) Action recognition based on a bag of 3D points. In: IEEE computer society conference on computer vision and pattern recognition workshops, pp 9–14

Lin W, Sun M-T, Poovendran R, Zhang Z (2010) Group event detection with a varying number of group members for video surveillance. IEEE Trans Circuits Syst Video Technol 20(8):1057–1067

Liang Y-M, Shih S-W, Shih AC-C, Mark Liao H-Y, Lin C-C (2009) Unsupervised analysis of human behavior based on manifold learning. In: IEEE international symposium on circuits and systems, pp 2605–2608

Liao S-K, Liu B-Y (2010) An edge-based approach to improve optical flow algorithm. In: Third international conference on advanced computer theory and engineering, vol 6, pp 45–61

Mahadevan V, Vasconcelos N (2009) Segmentation of motion objects from surveillance video sequences using temporal differencing combined with multiple correlation. In: Advanced video and signal-based surveillance, pp 472–477

McIvor AM (1999) Background subtraction techniques. Reveal Ltd, New Zealand

Ogale NA (2006) A survey of techniques for human detection from video

Park J-S, Yoon J-H, Kim C (2005) STable 2D feature tracking for long video sequences. Int J Signal Process Image Process Pattern Recogn 1(1):39–46

Park S, Aggarwal JK (2004) Semantic-level understanding of human actions and interactions using event hierarchy. In: IEEE computer society conference on computer vision and pattern recognition workshops

Peursum P, Bui HH, Venkatesh S, West G (2005) Robust recognition and segmentation of human actions using HMMs with missing observations. Eur Assoc Signal Process J Appl Signal Process 2005:2110–2126

Rahimt HA, Sheikh UU, Ahmad RB, Zaint ASM, Ariffin WNFW (2010) Vehicle speed detection using frame differencing for smart surveillance system. In: International conference on information science, signal processing and their applications, pp 630–633

Roach M, Mason JS, Pawlewski M (2001) Motion-based classification of cartoons. In: Proceedings of 2001 international symposium on intelligent multimedia, video and speech processing, pp 146–149

Robertson security submission (2006) Automatic human behaviour recognition and explanation for CCTV video surveillance

Rougier C, Meunier J, St-Arnaud A, Rousseau J (2011) Robust video surveillance for fall detection based on human shape deformation. IEEE Trans Circuits Syst Video Technol 21(5):611–622

Seki M, Fujiwara H, Sumi K (2000) A robust background subtraction method for changing background. In: Fifth IEEE workshop on applications of computer vision 2000, pp 207–213

Shafie AA, Hafiz F, Ali MH (2009) Motion detection techniques using optical flow. In: World Academy of Science, Engineering and Technology, vol 56

Tabb K, Davey N, Adams R, George S (2004) Detecting, tracking & classifying human movement using active contour models and neural networks. In: Innovations in intelligent systems, Physica (Springer) Verlag, pp 343–360

Techmer A (2001) Contour-based motion estimation and object tracking for real-time applications. IEEE, Infineon Technologies AG, Corporate Research

Xiaofeng L, Tao Z, Zaiwen L (2010) A novel method on moving-objects detection based on background subtraction and three frames differencing. In: International conference on measuring technology and mechatronics automation, pp 252–256

Yoshimitsu Y, Naito T, Fujimura K, Kamijo S (2010) Behavior understanding at railway station by association of locational semantics and postures. In: IEEE international conference on systems man and cybernetics, pp 3033–3038

Yunus RM (2009) Development of algorithm for human non-human object classification. Scholarly paper, University of Teknikal