International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.7, No.3 (2014), pp.99-122
http://dx.doi.org/10.14257/ijsip.2014.7.3.10

Survey on Video Analysis of Human Walking Motion

S. Nissi Paul and Y. Jayanta Singh

Dept. Computer Science Engineering and information Technology
Don Boco College of Engineering and Technology, Assam Don Bosco University
Guwahati, Assam - India
nissi.paul@dbuniversity.ac.in, jayanta@dbuniversity.ac.in

Abstract

In computer vision related applications, video analysis of human walking motion is currently one of the most active research topics. The task of analyzing human walking can be divided into three distinct subtasks – human detection or segmentation, motion tracking and walking pose analysis. Typically, the analysis of the human walking starts with the extraction of motion information, detection of the presence of humans in the sequences of frames and then followed by analysis of events related to walking. This paper presents a survey of different methodologies used for human walking motion analysis, approaches used for human detection or segmentation, various tracking methods, approaches for pose estimation and pose analysis. The common data sets available for building robust, automatic and intelligent systems to understand "walking" motion are also included. Finally, uses of unsupervised techniques for analysing human walking are highlighted. Human walking motion is a subset of a broad topic of human motion analysis.

Keywords: *human walking motion; human detection; segmentation; tracking; pose estimation; pose analysis*

1. Introduction

Extensive research is carried out on human motion in general and human walking is a specialized form of motion. Human walking is categorized by various types of walking and normal walking is usually recognised by rhythmic and repetitive movements. Video sequences of human walking motion can be obtained from various recording devices like, Charge-coupled devices (CCD) cameras, thermal cameras, night vision cameras and so on. To address the problem of analysing walking motion is hard due to the variability of poses and actions human can manifest. The three phases of analysing human walking motion include segmentation, tracking and pose analysis which are outlined in Figure 1.

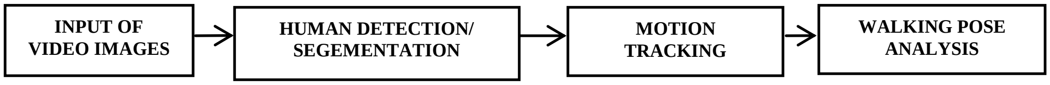


Figure 1. Phases of Analysing Human Walking Motion

Unlike the previous surveys, this paper surveys the tasks associated specifically with walking motion and discusses in detail the approaches and techniques of all the subtasks included in walking motion. The Table I shows some of the popular surveys which discussed various tasks relating to human motion analysis such as human detection, segmentation, motion tracking, activity detection, representation of motion- model-based and model free, types of poses, estimate of poses, comparison of different methods, demerits and merits.

The rest of the paper is organised as follows: Section II, firstly, describes the context of the types of images captured- intensity images and depth images. Secondly, taxonomy relating to marker- based systems and markerless – based systems and thirdly the levels of human walking categorised as atomic movement, actions, interactions and group activities are discussed. Section III, presents, different methodologies found in the literature for human walking motion. Section IV describes in detail the two main approaches used for segmentation, background subtraction as, global, top-down approach and local, bottom-up approach. Details on works on the other post processing operations- Morphological changes, Illumination changes, Removal of shadows, Occlusion handling are also included. Section V deals with the evaluation methods used for supervised and unsupervised techniques of segmentation. Section VI gives an overview of human walking motion tracking methods. Section VII deals with techniques and approaches of analysing human walking poses. Section VIII provides details on the common datasets available for performing pose analysis on walking motion and finally the paper concludes in Section IX with conclusion remarks.

Table I. Popular Surveys of Human Motion Analysis

Authors/ Groups	Year	Area of the survey paper
L. Wang [1]	2003	Human motion and activity understanding
Murat Ekinci [2]	2005	Human motion in indoor and outdoor environments.
Moeslund et al.[3]	2006	Initialization, tracking, pose estimation and recognition
Forsyth et al.[4]	2006	Recovery of human poses and motion from image sequences
Kruger et al. [5]	2007	Higher level of human activity recognition with intention recognition
Ronald Poppe [6]	2007	Model-based /free approaches for motion
Turaga et al. [7]	2008	Higher level of human activity recognition
Weinland et al. [8]	2010	Human motion recognition
Ronald Poppe [9]	2010	Recovery of human poses and motion from image sequences
Agarwal et al. [10]	2011	Human activity analysis
Liu et al. [11]	2012	View-invariant analysis of Human motion detection
L. Chen et al. [12]	2013	Human motion analysis using depth imagery

2. Context and Taxonomy

Widespread research works are performed on the human body movement, face movement analysis, hand gestures, limb movement, gait analysis and on a combination of them. The following section deals with the context and general taxonomy used for applications of human walking motion analysis.

2.1 Video Sequences Obtained as Intensity Images and Depth Images

Most computer vision research is performed on images, and videos captured by RGB cameras where each pixel is the intensity of incoming light. The images contain rich texture and color information and applications are developed similar to how humans directly interpret images depicting a scene. However, in images captured by depth sensors like stereo cameras, time-of-flight (TOF) cameras and structured light sensors, each pixel indicates calibrated distance between the camera and scene and 3-D structure of the scene is depicted in a 2.5 dimensional image. The distance from source to surface at each pixel is measured by the travel of light. The most remarkable advantages of depth images over intensity images are invariance against illumination changes and ease of using less complex approaches for the tasks of background subtraction, segmentation and motion estimation [12]. The figure 2 depicts – RGB image, Depth image and Intensity image

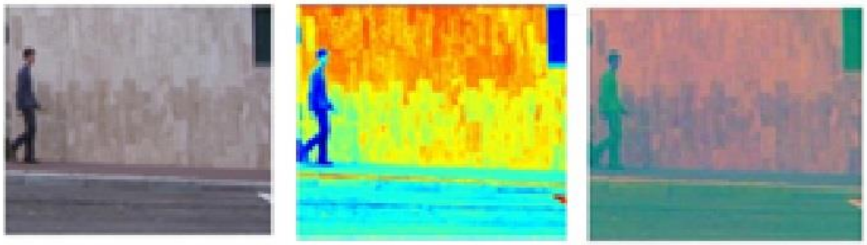


Figure 2. (a) RGB Image (b) Depth Image (c) Intensity Image

2.2 Marker-based and Markerless-based Applications

Figure 3 (a), shows marker–based applications using sensors attached to the human body to capture motion whereas figure 3(b) markerless-based applications refer to the task of full body motion capture without the need of markers or any specialist suits. Marker-based capture of human motion use, three main approaches - electromechanical, electromagnetic and optical sensors.

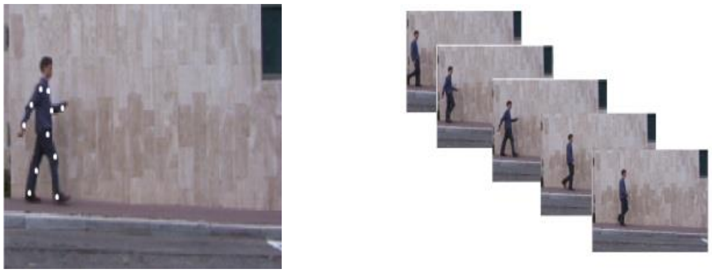


Figure 3: (a) Marker Based Systems (b) Markerless Capture by a Video Camera

Electromechanical systems use suits worn on body and measuring devices constructed from potentiometers and sliding rods. A major drawback is that full range of human motion cannot be expressed specifically in walking and any change in motion. However, accurate results can be obtained only for restrictive movements. The electromagnetic approach improves slightly over electromechanical systems. Sensors are used at key points on the body with an accompanying transmitter to provide information on position and orientation of sensors but still have the disadvantage of wires attached to each sensor. Optical sensors, which are less restrictive, operate on systems with pre-calibrated multi-cameras and infrared reflective balls markers on the body. However, all marker-based systems depend on experts who can handle the devices and are expensive. The current research on human motion has shifted towards non-intrusive mechanism of motion capture in the form of videos, where many computer vision and pattern recognition methods and approaches can be used to understand human walking motion [13].

2.3 Action hierarchy Perspective

Aggarwal and Ryoo [10] categorises human activities into four different levels: 1. Gestures or atomic movement, 2. Actions, 3. Interactions, and 4.Group activities. Gestures are simple movements of a person's body part and are atomic components describing the meaningful motion of a person. "Moving left leg" and "raising a leg" are good examples of gestures. Actions are single-person activities that may be composed of multiple gestures organized temporally, such as "walking," "running" and "kicking." Interactions are human activities that involve two or more persons and/or objects. For example, "two persons fighting" is an interaction between two humans and "a person kicking a stone" is a human-object interaction involving human and object. Group activities are the activities performed by groups of multiple persons and/or objects like, "a group of persons marching," "a group watching a cricket match" and "two groups fighting" are typical examples. Figure 4 depicts all the different levels of human activities.

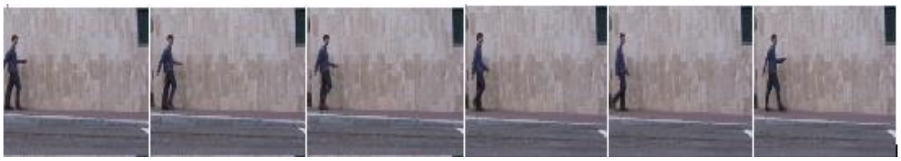


Figure 4. (a) Atomic Movements- Moving Left Leg

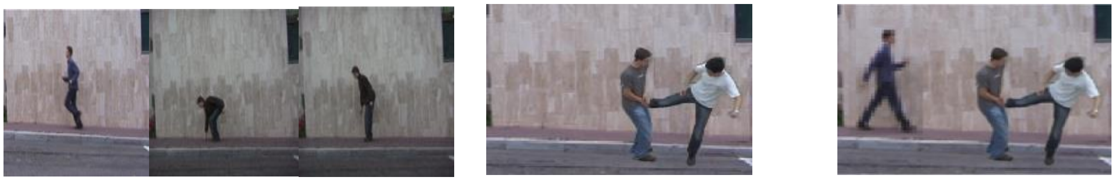


Figure 4 (b) Actions- running, Bending, (c) Interaction: Two People (Kicking) (d) Group Activity

3. Human Motion Methodologies

The area of focus on human walking motion presented in this paper has been a challenge, and many research works have been reported in the past by several researchers. However, the work accomplished is still in infancy and mostly incomplete due to increasing important applications in wide key areas. This section gives an overview of the approaches used for analysing human walking motion in Table II.

Table II. Methodologies of Human Motion Activity

Authors	Methodologies
Rius, I., et al.[14]	Learn various 3D human postures from training sequences, use matching algorithm based on dynamic programming to map different postures and different motion cycles.
Korc, F. et al.[15]	Use 2D articulated models in single view sequences for detecting and tracking of Humans in three steps, (1) Detecting human candidates, (2) Validating the model of a human and (3) Tracking of the model in

	consequent frames. The model is developed with a six-link model and an articulated head of frontal view of a person.
Mikic, I., et al. [16]	Develop an integrated system for automatic acquisition of the human body model and motion tracking using input data acquired from multiple synchronized video streams. Tracking is performed on the 3D voxel reconstructions computed from the 2D foreground silhouettes, the human body model consists of ellipsoids and cylinders.
Ramanan, D. And Forsyth D. A [17]	Appearance based model is built on people in motion and then clustering of candidate body segments is formulated and then the model is used to find all individuals in each frame.
Ning, H., et al. [18]	A motion model is built from the semi-automatically acquired training data and motion constraints are computed by analysing the dependency of joints. Both of them are integrated into a dynamic model in order to reduce the size of the sample set.
Viola, P., et al. [19]	Combine two methodologies by integrating information of appearance with motion information and detection algorithm along with trained sequences. Both motion and appearance information are used to detect a walking person.
Gonzalez, J. J., et al. [20]	A robust feature-based tracking by initializing a standard point-wise tracker method and grouping image points undergoing the same rigid motions is built to track human motions of different body parts without articulation.
Sappa, A. D., et al. [21]	A technique that combines the prior knowledge regarding a person's motion with human body kinematics constraints is developed. It uses an efficient feature point selection and tracking approach to compute feature points' trajectories and then 3D motion models associated with each joint are locally obtained by using key frames, meaning frames where both legs are in contact in the floor.
Wang, L., et al. [22]	Automatic person recognition from body silhouette and gait is performed where background subtraction procedure is combined with a simple correspondence method to segment and track spatial silhouettes of a walking figure. Simple feature selection and parametric Eigen space representation are used to reduce the computational cost during training and recognition.
Yoo, J. H. and Nixon M [23]	The extraction of the gait figure is made using 2D stick figures from the body contour by determining the body points using linear regression and motion tracking with topological analysis. Then, the detection of the gait cycle is computed by symmetry analysis, kinematic analysis and feature extraction to classify the gait patterns.

The methodologies covered in Table II describe various types of representation of human body based on shape models like stick figures, 2D contours or volumetric models, features, kinematic constraints and appearance. Further work is carried out to accomplish various useful tasks using computer vision algorithms and machine learning strategies. Some researchers have combined appearance with motion tracking to get improved results.

4. Human Detection or Segmentation

A computer vision application where human motion analysis is carried out the initial step is to detect moving human. Detecting moving humans initially provides a focus of attention for future tasks. The objective of segmentation is to obtain relevant information about the human figures located in the frames of human motion. It is a process of dividing an image into disjoint regions such a way that the combination of all the segmented regions results in the original image. This division is performed by grouping neighbouring pixels according to criteria of proximity and/or similarity followed by further processing steps on the image focusing on specific regions of interest and not the complete image. Video segmentation is performed on either in grey tones or in colour image sequences. There are various approaches available to detect moving human figures in videos. Commonly used segmentation techniques are given below.

a) Background subtraction technique where the foreground is extracted from the background with a fixed camera at a fixed angle and the background is completely stable.
b) Temporal differencing technique unlike in background subtraction, the background is dynamic. The moving object is detected by taking the difference between two or three consecutive frames and is combined with the figure motion edges [127].
c) Optical flow technique is applied when both camera and humans are in motion, called ego motion [25].

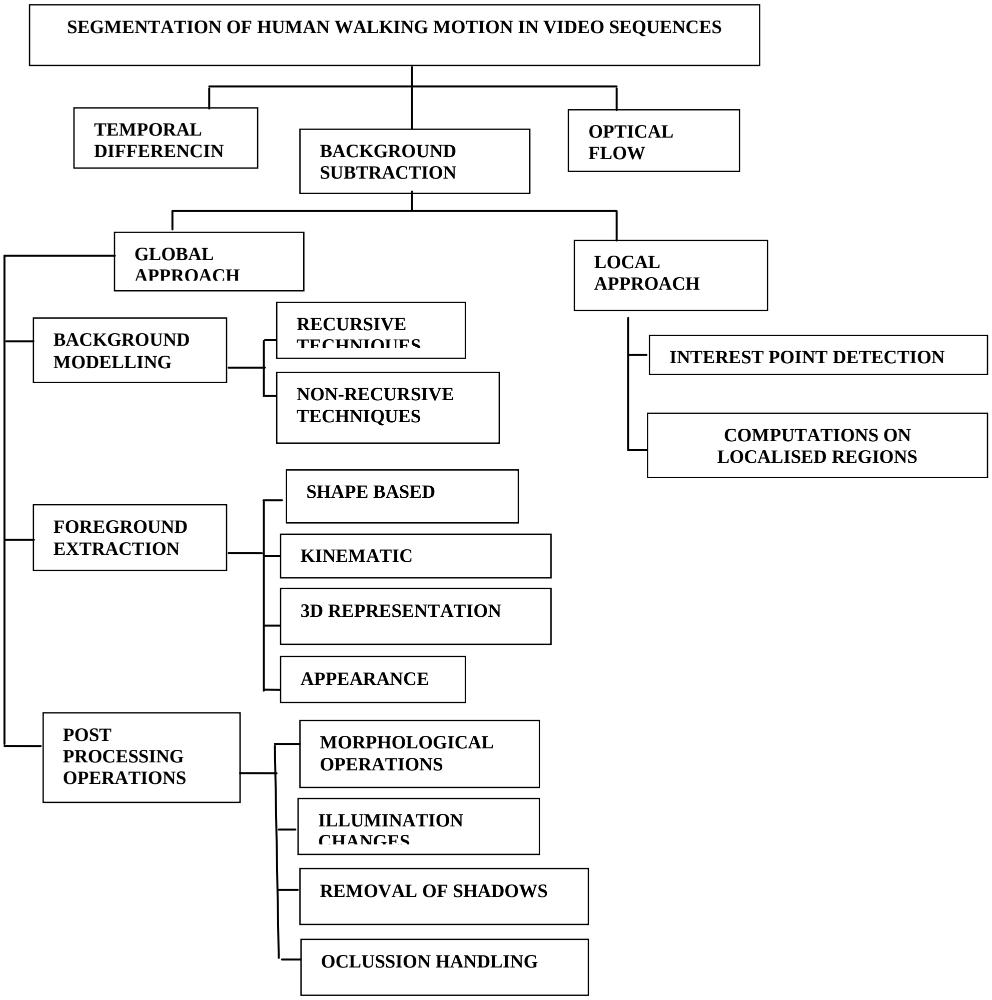


Figure 5. Approaches to Detect Moving Human Figures

This section emphasises on background subtraction and elaborates on different types of background subtraction approaches found in the literature as depicted in Figure 5. Other challenges such as illumination changes, dynamic background, occlusion and removal of shadows are considered as sub problems to address the problem of locating humans and extracting human figures in video sequences.

Broadly, two approaches can be used to segment video images to extract moving humans, *Global*- top down approach and *Local*- bottom up approach. In Global approach the region of interest (ROI) is taken as a whole frame and the moving person are localized using background subtraction methods. However, accurate localization, background subtraction, viewpoints, noises, shadows and occlusions are to be handled carefully to get an effective performance. Where as in Local approach- relevant region of interest (ROI) is created initially by detecting spatial or/and temporal interest points around the moving humans, and local patches are created. After which calculations around the points and patches are combined into a final representation. This approach is less sensitive to noise, partial occlusion and sometimes will not require background subtraction. However, care is to be taken on how to extract the relevant ROI.

4.1. Global-Top Down Approach

Global-Top down approach is commonly known as *background subtraction.* Several background subtraction techniques are found in the literature. The purpose of background subtraction method is to distinguish moving objects from static or from slow moving regions in a scene. Generally, the region of interest is obtained considering the whole frame and is followed by morphological operations to reduce noise, remove shadows and to identify occluded regions. Segmentations using background subtraction are based on motion, depth data, appearance and shape.

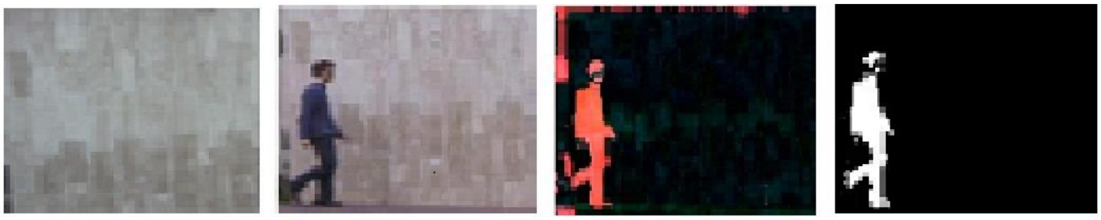


Figure 6. (a)Background (b)Incoming Frame (c)Background Subtraction d) Foreground Extracted

For static backgrounds, the moving regions are detected by taking the difference between the incoming frames and the reference background modelled frame in a pixel-wise fashion. This method is simple but is sensitive to illumination changes, noises, dynamic scenes. The robustness of this technique depends largely on the background model chosen to be as a reference frame. Generally, all background subtraction algorithms use two steps- (1) Background modelling (2) Motion segmentation or foreground extraction. According to study of Prof. Williams [26], a background subtraction must adapt to gradual or fast illumination changes (changing time of day, clouds, *etc.*), motion changes (camera oscillations), high frequency background objects.

4.1.1. Background Modelling

Background modelling is an important stage in the entire background subtraction algorithms. In the literature available background modelling can be divided in two main categories, namely non-recursive and recursive techniques.

4.1.1.1. Non-Recursive techniques

A number of frames (N) are kept in the buffer and estimation is calculated for the background model. These techniques are highly adaptive but suffer from the requirement for large memory and sometimes large buffer.

1. Frame differencing

Here, the background model is the frame, t-1 in which the previous frame is modelled, and the incoming frame is subtracted with a threshold.

$$|I_t – I_{t-1}| > T \qquad (1)$$

Where I_t is the intensity of frame t and T is a fixed threshold. This technique is sensitive to a threshold value and includes the advantages of high adaptability, less memory space and less computational load. . The main disadvantage is only threshold value does not provide an accurate result always and the technique is found to be sensitive to noise. Another main disadvantage is that when human is found to be stationary even for a very short time in the subsequent frames, the human figure becomes background [27].

2. Median filtering

The estimated background value of each pixel in the background model is calculated as the median of that pixel in all frames chosen in the buffer. This is another effective not requiring high computational load but disadvantage is the requirement of (Nxframe size) [28].

3. Linear predictive filter

The background estimate is computed by applying a linear predictive filter on the pixels of the frames in the buffer. The filter coefficients are estimated depending on the sample covariance at each frame time [29].

4.1.1.2. Recursive Techniques

In these techniques buffer of frames is not used instead the background image is updated recursively. The advantage is only one frame will be updated every time a new frame is received. Another advantage is that these techniques can handle multi-modal backgrounds. However, they are computationally complex and sensitive to illumination changes [30].

1.Running average

A simple background modelling algorithm is computed as

$$B_i+1= a\ C_i+ (1-a)\ B_i \qquad (2)$$

Where B stands for background, C_i is the current frame and 'a’ is defined as the learning rate with a typical value of 0.05 [31].

2. Approximated median filtering

The technique was developed by McFarlane and Schofield [32]. Initially, a background estimate is taken and when a pixel in the current frame has a gray value which is larger than the pixel in the background estimate, it is incremented by one. On the other hand when the value of a pixel in the current frame has a value lower than the background estimate, the pixel in the background estimate is decremented by one. When applying this function to the background model, the model converges to an estimate where half the input pixels are greater than the background and the other half are less than the background model. This technique gives a satisfactory result but is slow to adapt to the big changes in the real background.

3. Kalman filtering

This method assumes that the best information of the system state is obtained by estimation. Several approaches are found in literature, and most of them use the luminance intensity and its temporal derivative or intensity and its spatial derivatives. In the most simple variation, we can model the background estimation B(t) as:

$$B(t) = A(t)B(t-1)+K(t)[z(t) – H(t)A(t)B(t-1)\ldots(3)$$

Where A(t) is the system matrix which describes the background dynamics, H(t) is the constant measurement matrix, z(t) is the system input and K(t) is the Kalman gain matrix. Advantage of Kalman filtering is switching between fast and slow adaptation whether the pixel is a foreground or background pixel. Disadvantage is leaving long trails of moving objects.

4. Mixture of Gaussians:

Mixture of Gaussians uses 3 to 5 Gaussian distributions simultaneously. Each pixel is modelled by a mixture of Gaussians that will sum to form a probability distribution function $f(x_t)$ [33].

$$f(x_t)=\Sigma_{i=1}^{k}\omega_{i,t}.\ \eta(x_t - \mu_{i,t}\ \Sigma_{i,t}) \qquad \ldots\ldots\ldots\ (4)$$

Where, f is the probability distribution function, ‘x’, a certain pixel value at time ‘t’ and ‘k’ is a value set between 3 and 5. $\omega_{i,t}$ is an estimate of the weight of the i^{th} Gaussian in the mixture and $\mu_{i,t}$ is the mean value of the i^{th} Gaussian in the mixture. η is the mean of Gaussian function estimate value in next frame. This algorithm is simple and of low complexity. This algorithm works well for stable scenes and segmentation becomes difficult when both background and foreground are complex.

4.1.2. Foreground Extraction

The purpose of foreground extraction is to precisely extract the human figure from the video frames. Binarisation is the process of converting a grayscale image to a black and white image. Binarisation is done using a global thresholding, all pixels are set to a defined value to white and the rest of pixels to black. In the literature, we find different foreground extraction of human figures based on *Shape, kinematic structure, 3-dimensional shape and appearance*. Post processing operations of morphological operations, removal of noises, illumination changes and removal of shadows, occlusions are performed to get refined and accurate human silhouettes.

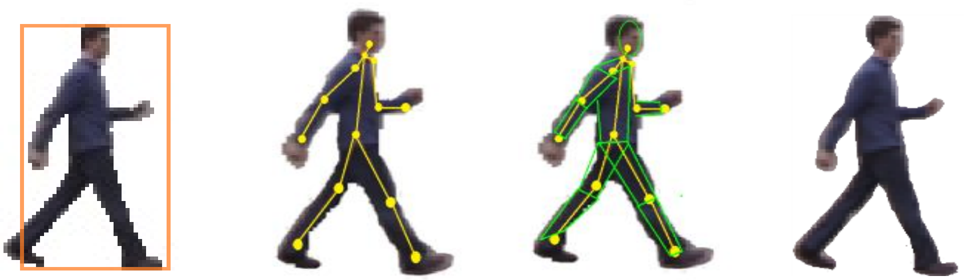


Figure 7. (a) Shape-based (b) Kinematic Structure (c) 3-dimensional Shape (d) Shape Appearance

4.1.2.1 Shape-based Representation

Foreground extraction can be represented in different description of shapes such as point, box, silhouette and blob. Collins *et al.*, [34] represented moving objects as blobs. Kuno and Wantanbe [35] use shape parameters of human silhouettes. Representations such as cylinders, cones, ellipsoids and super-quadrics are found in [36]. Silhouettes and contours are other representations which are commonly in use. Contour representation defines the boundary of an object. The region inside the contour is called the silhouette of the object. Silhouette and contour representations are suitable for complex non-rigid shapes [37]. Skeleton representation is used as shape representation for both articulated and rigid motions. Skeleton is extracted by applying medial axis transform to the human silhouette [38]. Surface representations such as generic mesh models, are described in [39].

4.1.2.2. Kinematic Structure

Kinematic structure, *C*omprises of a fixed number of joints with specific degrees-of-freedom. Kinematic initialization is carried such that the lengths of limb are considered for the purposes analysing human walking motion. Further, several approaches are used to build trajectories by placing 3D markers, finding skeletal symmetry of the postures and anthropometric constraints between ratio of limb lengths are used to allow estimation of the kinematic structure for an unknown scale factor. [40].

4.1.2.3. 3-Dimensional Representation

Human figure representation in 2-dimensional can be extended to 3-dimensional representation. Weinland *et al.*, [41] combine silhouettes from multiple cameras into a 3-dimensional Voxel model. Multiple images over time can be stacked to form a d-dimensional space-time volume.

4.1.2.4. Appearance

Directly persons appearing in the video sequences are observed and statistical models are developed like mixture of Gaussian [42], histograms of colour, texture combination [43]. Texture maps are derived in [44]. Approaches such SVM, AdaBoost were successful to learn body parts available during training. [45]. Lim *et al.*, [46] address the problem of changing appearance of humans due to motion by mapping pixels inside a bounding box [47].

4.1.3. Post Processing Operations

Processing a video stream and extracting foreground human figures from frames is a critical initial step for many computer vision applications and to improve results of background subtraction algorithms few of the post processing techniques given below help to get more accurate extractions of the foreground figures to perform future tasks of tracking and pose analysis.

4.1.3.1. Morphological Operations

Morphological operations provide systematic alterations of the geometrical contents of the human figures extracted from the frames. A Structuring element *s* is used to probe the human figure extracted. The basic operations include *dilation and erosion*. Widely used combinations are *opening, closing*.

Dilation is a process in which the binary image is expanded from its original shape. The dilation operation is defined by $(f \oplus s)$ producing a new binary image g= $f \oplus s$ and a layer of pixels are added to inner and outer boundaries of the regions. *Erosion* is the counter process of dilation. Erosion shrinks the image. The erosion operation is defined by $(f \ominus s)$ producing a new binary image g= $f \ominus s$ and layer of pixels from both the inner and outer boundaries of regions are stripped away [48].

Opening is an operation where first morphological operation, erosion is applied and then followed by dilation operation. Opening smooths the inside of the human figure contour that is detected. *Closing* is the opposite of *opening* dilation operation is followed by erosion. The closing operation fills the small holes and gaps found in a single pixel.

4.1.3.2. Illumination Changes

1. Ridder *et al.*, [49] modelled each pixel value with Kalman filter to compensate illumination variance.
2. Stauffer *et al.*, [33] developed a theoretical framework for adapting to illumination changes.
3. Haritaoglu [50] used a statistical model to represent each pixel with 3 values of minimum, maximum intensity values and the maximum intensity difference between consecutive frames during training.

4.1.3.3. Removal of Shadows

1. Extended Expectation Maximization algorithm: Friedman *et al.*, [51], Malik *et al.*, [52] mixed Gaussian classification model is generated for each pixel with 3 separate predetermined distributions corresponding to background, foreground and shadow. Automatic updating of each distribution according to the likelihood of membership. Shadows are removed successfully.
2. Mc kenna *et al.*, [53] used an adaptive background model with colour and gradient information to reduce influences of shadows and unreliable colour cues.

4.1.3.4. Occlusion Handling

During Occlusions only a portion of each moving human is visible. The classification of Occlusion include- self-occlusions, of a human body, inter-object occlusions, where the humans are occluded by fixed objects such as in case of indoor scenes- tables, chair, objects carried by the person in motion or outdoor scenes by buildings, trees other moving objects like cars, people and occlusion by background scenes. Following approaches were used by various researchers for occlusion handling:

1. Data fusion from multiple cameras
2. Extending to 3-dimensional representation
3. Combine multiple features and predict the estimated pose or action of the moving human.

1. Data fusion from multiple cameras

Huang and Xu [54] used orthogonally placed cameras at approximately similar height and distance to a person. Silhouette from both cameras are aligned at medial axis and an envelope shape is calculated. Cherla *et al.*, [55] used orthogonal cameras and combined features to get view-invariant images. Weinland *et al.*, [56] combine silhouettes from different cameras and perform camera calibration to create 3D- voxel model.

2. Extending to 3-dimensional representation:

The problem of occlusion can be handled by extending a human figure in 2D image sequence and estimate its pose in 3D space. The framework in [57] is extended by incorporating a particle filter to focus on partial occlusions and based on degree of dependencies between predictions and measurements the occlusions are resolved.

3. Combine multiple feature:

Dynamic models or nonlinear dynamic models can be used to combine multiple features and predict the estimated pose or action of the moving human. A linear velocity model is used along with Kalman filter by Beymer and Konolige [58]. Other models used are Silhouette projections, optical flow. Cremers *et al.*, [59] built a shape model from subspace analysis called principal component analysis (PCA) of possible shapes to fill in missing contours.

4.2. Local- Bottom–up Approach

In the Local-Bottom-up approach initially, region of interest is computed, methods are developed to localize the relevant regions of interest across the video frames in the form of patches and are later combined to represent final image. Generally, the methods are invariant to changes of person appearance, partial occlusions, view changes, noises and illumination changes. The figure 6 shows the sequence of operations that can be used in the bottom-up approach.

1. Interest point detection
2. Computations on the localized regions- a) smoothing and dimensionality reduction, b) representations, c) correlations.

4.2.1. Interest Point Detection

Interest point detectors are used to find regions of interest in images. Some of the early and successful methods are Harris interest point detector [59], KLT detector [60] and SIFT detector [61]. Laptev and Lindeberg [62] used Harris corner detection along with 3D for both spatial and temporal domains. Kadir and Brady [63] included methods to calibrate camera motions along with locating 2D salient points and then converting to 3D by computing entropy within each cuboid generated and centres with local maximum energy are selected as salient points. Dollar *et al.*, [64] applied Gabor filtering along spatial and temporal domains and the interest points were calculated using neighbourhood points with local minima. Willems *et al.*, [65] computed salient points with the help of a 3D Hessian matrix. Wang *et al.*, [66] used dense sampling for interest point detection along the space and time dimensions.

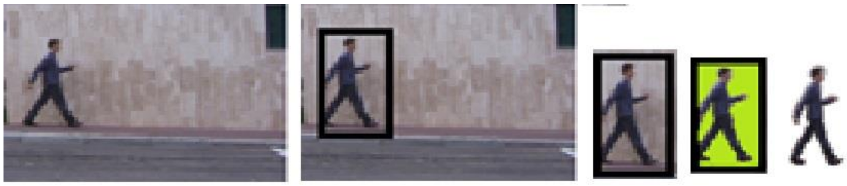


Figure 6. (a) Incoming Frame (b)Interest Point Detection (c) Region of Interest (d) Computations on Region of Interest (e) Extracted Silhouette

4.2.2. Computations on the Localized Regions

Once the region of interest are computed, the following computations can be carried out

4.2.2.1. Smoothing, Dimensionality Reduction

The space and time patches representations can be further processed for smoothing and reducing of state-space search. Schuldt *et al.*, [67] calculate patches of region of interest and compute normalized derivatives in space and time. Niebles *et al.*, [68] applied smoothing and dimensionality reduction using Principal component analysis (PCA). Jhuang *et al.*, [69] use several phases of computations starting with applying Gabor filters to dense flow vectors, next phase is followed by a low max operation then global max operation is applied and finally matching is performed to obtain final set of patches. Comparing patches is usually carried out using codebook where clustering patches and selecting the close related patches as codewords [70].

4.2.2.2. Representations

Grids can be used to bin patches spatially or temporally. Ikizler and Duygulu [71] use spatial grid approach to sample oriented rectangular patches which bin into a grid. Nowozin *et. al.*, [72] used temporal instead of spatial grid. PCA reduced and extracted interest points are mapped on to codebook indices. Laptev and Perez [73] use spatio-temporal grid representation and Bregonzio *et al.*, [74] do not use any local image descriptor rather they look at all interest points within cells of spatio-temporal grid with scales.

4.2.2.3. Correlations

Another important task is to exploit correlations between the interest point patches to select relevant image descriptors for proceeding to next level of understanding images. Scovanner *et al.*, [75] construct a co-occurrence matrix of features is constructed and iteratively merged until all are above a specified threshold. Several other researchers have contributed using supervised, semi-supervised and unsupervised approaches.Supervised learning methods require large number of samples. *Adaptive boosting* and *Support Vector Machines* are commonly used.

5. Evaluation Methods for Segmentation

Evaluation of the segmentation techniques consists of determining whether the regions detected by the algorithms are the actual regions of interest or not according to what is expected in the applications [76]. Usually, evaluation methods are either subjective or objective, relating to specific applications. Sometimes objective and supervised methods use ground-truth reference. Objective and unsupervised methods which do not require comparison are still to be fully implemented. Unsupervised methods have the advantage of self-tuning and can be used for generic images or applications.

5.1. Supervised Methods

Supervised evaluation methods [77, 78] evaluate segmentation algorithms by comparing the resulting segmented image against a manually-segmented reference image, which is often referred to as a ground-truth. The degree of similarity between the human and machine segmented images determines the quality of the segmented image.

One potential benefit of supervised methods over unsupervised methods is that the direct comparison between a segmented image and a reference image is believed to provide a finer resolution of evaluation, and as such, discrepancy methods are commonly used for objective evaluation. However, manually generating a reference image is a difficult, subjective, and time-consuming task. For many applications, it is hard or maybe impossible. Besides, for most images, especially natural images, we usually cannot guarantee that one manually-generated segmentation image is better than another. In this sense, comparisons based on such reference images are somewhat subjective.

A variety of discrepancy measures has been proposed for segmentation evaluation. Most early discrepancy methods evaluated segmented images based on the number of misclassified pixels versus the reference image, with penalties weighted proportional to the distance to the closest correctly classified pixel for that region [79-83]. Another group of discrepancy methods are based on the differences in the feature values measured from the segmented images and the reference image [84-89].

5.2. Unsupervised Methods

Supervised methods evaluate segmented images against a reference image, whereas unsupervised evaluation methods [90], also known as stand-alone evaluation methods do not require a reference image, but instead evaluate a segmented image based on how well it matches a broad set of characteristics of segmented images as desired by humans [91].

Unsupervised evaluation is quantitative and objective. The most important advantage is that it requires no reference image. The ability to work without reference images allows unsupervised evaluation to operate over a wide range of conditions (or systems) and with many different types of images. This property also makes unsupervised evaluation uniquely suitable for automatic control of online segmentation in real-time systems, where a wide variety of images, whose contents are not known beforehand, and need to be processed [92]. Some of the unsupervised evaluation methods included in the survey paper of H. Zhang et al. [92] are given in the Table III.

Table III. Unsupervised Evaluation Methods

Methods	Description
η	Measures both intra- and inter-region variance of the foreground object and the background, allowing the segmentation algorithm to select the threshold that maximizes the inter-region variance [93].
F_{RC}	Global intra-region homogeneity and global inter-region disparity are taken together [94].
Z*eb*	An evaluation criteria used for internal and external contrast of the regions measured in the neighbourhood of every pixel [90].
V_{EST}	Used for video segmentation evaluation. It is a metric measuring the spatial colour contrast along the boundary of each object [95].

6. Tracking Human Motion

Tracking is finding corresponding humans in successive frames over a time period and depicting temporal trajectories or correspondences. For some applications, person tracking is avoided as tracking of humans does, not in itself contribute directly to the performance of the pose estimation. However, tracking is treated as subsequent process by many researchers to carry out tasks relating to temporal aspects of moving humans, like determining prediction, high level knowledge and state of motion. This section describes the techniques used for tracking.

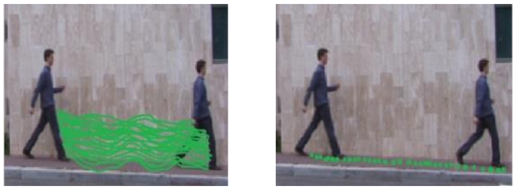


Figure 8. Examples of Tracking

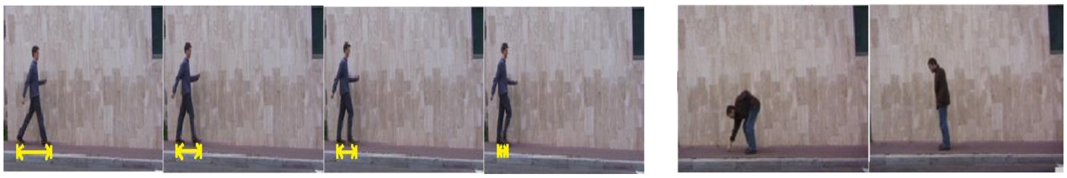
The most widely used mathematical tools for tracking are- Kalman Filter [96], the Condensation algorithm [97], Dynamic Bayesian Network [98]. Tracking strategies can be divided into various categories according to the suitability of different applications of human walking:

1. Tracking of human body parts such as hand, face, and leg [99]
2. Tracking of whole body
3. Total number of views- Single-view [100], Multiple-view [101], and Omni-directional view [102]
4. 2-Dimensional and 3-Dimensional Tracking
5. Tracking environment-indoors and outdoors
6. The number of humans to be tracked -single human, multiple humans and human groups.
7. The camera’s state - moving and stationary [103]

Yilmaz *et al.*, has described three types of tracking representations in [104]. Point correspondence, primitive geometric models and Contour evolutions which are relevant to human walking motion tracking also.

7. Pose analysis of Human Walking Motion

In computer vision tasks, Pose analysis is considered to be the final task in accomplishing understanding of human walking motion. Pose analysis is an interesting and challenging problem due to diversity movement poses, clothing, lighting changes and self- occlusions. Recently, a number of approaches have used depth information to analyse poses and understand their underlying intentions. The pose analysis of human walking motion can be processed in two steps: 1. *Pose estimation* and 2. *Action recognition.*



(a) Understanding Patterns of Walking (b) Recognising Sudden Changes

Figure 8. Pose Estimation and Action Recognition (a) and (b)

7.1. Pose Estimation:

The purpose of pose estimation is to estimate type of articulation and movement a pose is taken during walking. Poses can be constructed by supervised or unsupervised learning. Supervised learning is used where the unknown poses are compared with predefined poses already recorded in the training whereas for unknown poses patterns are not available beforehand, self-organising and self-learning techniques are used. T.B Moesland *et al.*, [105] separated pose estimation algorithms into three categories:

1. *Model-free*, where no explicit priori models are built. Probabilistic assembly of individual parts of the body using classifiers such as 2-D shape, SVM classifiers, AdaBoost and another method is to use example-based where direct mapping of 2-dimensional sequences to 3-dimensional poses are performed.
2. *Indirect,* use model shape and motion for reconstruction of visual-hull or align skeletal models to actual body shape in first frame where human figure is detected for pose estimation as a reference to help in interpretation of measured data.
3. *Direct model,* use an explicit 3-dimensional geometric representation of human shape and kinematic structure and use an analysis-by-synthesis methodology to optimize similarity between estimated and observed images.

Recent contributions include, Kinect pose estimation system [106] based on 3D scene information, and is the only currently available system to our knowledge that reaches satisfactory pose estimation performance in real-world scenarios. Images are depth based and Kinect system enable easy data acquisition for new pose estimation approaches. Singh et al. [107] describe an approach incorporating contextual knowledge into the pose estimation, about human interaction, and significantly improve pose estimation performance based on a combination of Pictorial Structures and a pose estimation approach by Deva Ramanan [108].

Johnson and Everingham [109] proposed a new, large scale data acquisition method and they introduce Pictorial Structures in an extended framework called Clustered Pictorial Structure Models. It can deal with extremely articulated poses and reaches very high pose estimation performance. Another useful contribution is by Wang and Koller [110] where, high level information from Pictorial Structure matching can be combined with low level per pixel segmentation information. An energy function is assembled that combines these different levels of information. Wang and Koller [110] apply relaxed dual composition to include infeasible energy functions in the optimization process. Their concept combines multiple methods and improves their common results.

7.2. Action Recognition:

Pose estimation is followed by *Action recognition* to identify and investigate actions. Goals like distinguishing regular walk from irregular walking movements, determining different types of walking; detecting sudden falls and other applications can be represented. Holistic approaches attempt to recognize a person, find gender or perform simple actions like walking or running. With respect to specific tasks relating to particular body parts it is more advantageous to consider relating body parts instead of taking into account the entire body, according to survey paper of T. B Moeslund *et. al.*, [111].

Nagel [112] suggested a hierarchy of change, event, verb, episode and history. Bobbick [113] proposed different levels of abstraction – movement, activity and action. T. B Moeslund [111] used action hierarchy as- action primitives, actions and activities.

8. Common Datasets

In the recent years many public video datasets are available for performing various research tasks such as segmentation, tracking, recognition, motion analysis and so on. The advantages are that, they save time and resources that there is no need to record new video sequences. A suitable dataset helps in performing tasks to be more reliable and the use of same datasets facilitates comparison of different approaches and gives insight of various methods and algorithms. The following Table IV gives the list of the common public data sets which are useful in analysing human walking motion in the video sequences.

Walking is one of the actions found in all the datasets mentioned in the Table IV. The ground truth and various walking styles, running, bending actions are also included for the purpose of focusing on different practical type of scenes, number of actions [126]. KTH, Weizmann datasets were recorded to study new algorithms to improve performance human actions of which one set is only for human walking movement.

CAVIAR is a project developed to answer the question “Can rich local image descriptions and other image sensors, selected by a hierarchical visual attention process and guided and processed using task, scene, function and object contextual knowledge, improve image-based recognition process?”. The datasets include 9 activities of which one belongs to human walking.

ETISEO was created to improve video surveillance algorithm robustness, focusing human related activities, walking, running and other related activities.

Table IV. Data Sets for Action Recognition- Human Walking Motion

Type of Recording environment, camera movement	Datasets
Indoor and outdoor, simple and static background. Static camera movement	KTH [114], Weizmann [115]
Indoor and outdoor, complex without static background and illumination conditions are not controlled. Static camera movement	CAVIAR [116], ETISEO [117],CASIA Gait recognition A [118], CASIA Gait recognition B [118], MSR Action [119]
Indoor, multi-view recording. Static camera movement	IXMAS [120], MuHAVi [121]
Outdoor, multi-view recording. Static camera movement	CASIA [122]
Online Repository. Several camera movements	VISOR [123], VIRAT [124]
Web Videos. Several camera movements	HMDB51 [125]

CASIA action, CASIA Gait recognition A, CASIA Gait recognition B are three different datasets developed to research biometrics and intelligent surveillance consisting of various sequences of human activities of walking, running, bending, etc. The intention was to study how to deal with big changes in the view angle without using 3-dimensional models. MSR Action dataset was created to study the behavior of recognition algorithms in presence of clutter and dynamic backgrounds and varying types of actions.

IXMAS dataset was created to investigate how to build spatio-temporal models of human actions that could support categorization and recognition of simple action classes independent of view-point and body sizes. MuHAVi was developed for the purpose of evaluating silhouette-based human action methods. It provides a realistic challenge for illumination changes and segmentation.

VISOR, VIRAT are two large datasets collected in natural scenes showing people in various viewpoints, resolutions and background clutter. HMDB51, videos were collected from various sources from movies and actions are categorised into five types: general facial actions, facial actions with object manipulation, general body movements, body movements with object interaction and body movements for human interaction.

9. Conclusion Remarks

In this paper, we presented an extensive survey of various concepts and techniques focused on analyzing human walking motion, which has an increasing importance among the computer vision community applications relating to visual surveillance, video retrieval and human-computer interaction. Unsupervised methods offer unique advantages over the supervised methods of operating over a wide range of conditions and suitable for automatic control of online segmentation in real time. Without the manually segmented reference frame or ground truth the evaluations, some of the content or features that are not possible to be known beforehand can easily be handled.

The introduction to pose estimation and action recognition of human walking motion opens a large scope of applications. Tracking may or may not be required in all cases. Poses of human can be varying when we change the interested regions of different parts of the body. Details on some of the common data sets are also provided in this survey. We believe that this paper will give valuable insight into the research topic of human walking motion.

References

[1] L. Wang, W. Hu and T. Tan, "Recent developments in human motion analysis", Pattern Recognition, vol. 36, no, 3, **(2003)**, pp. 585–601.

[2] M. Ekinci, "Sihouette based Human Motion Detection and Analysis for real time automated Video surveillance", Turk J Elec Engineering, vol. 2, no. 13, **(2005)**.

[3] M., "A Survey of advances in vision-based human motion capture and analysis", Computer Vision and Image Understanding, vol. 104, **(2006)**, pp. 90-126.

[4] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien and D. Ramanan, "Computational studies of human motion part 1: tracking and motion synthesis", Foundations and Trends in Computer Graphics and Vision, vol. 1, no. 2, **(2006)**, pp. 77–254.

[5] V. r Krüger, D. Kragic, A. Ude and C. Geib, "The meaning of action: a review on action recognition and mapping", Advanced Robotics, vol. 21, no. 13, **(2007)**, pp. 1473–150.

[6] R. Poppe, "Vision-based human motion analysis: an overview", Computer Vision and Image Understanding (CVIU), vol. 108, no. 1-2, **(2007)**, pp. 4–18.

[7] P. Turaga, R. Chellappa, V. S. Subrahmanian and O. Udrea, "Machine recognition of human activities: a survey", IEEE Transactions on Circuits and Systems for Video Technology, vol. 18, no. 11, **(2008)**, pp. 1473–1488.

[8] D. Weinland, R. Ronfard and E. Boyer, "A Survey of vision-based methods for action representation, segmentation and recognition", Computer Vision and Image Understanding, vol. 115, **(2010)**, pp. 224-241.

[9] R. Poppe, "A survey on vision-based human action recognition", Image Vision Comput., vol. 28, **(2010)**, pp. 976-990.

[10] J. Agarwal and M. Ryoo, "Human activity analysis: A review", ACM Comput. Surv., vol. 43, **(2011)**, pp. 16:1-16:43.

[11] X. Ji and H. Liu, "Advances in View-invariant Human Motion Analysis: a Review", IEEE Transactions on systems, Man, Cybernetics, vol. 40, no. 1, **(2010)**.

[12] Y. M. Lui, "A least squares regression framework on manifolds and its application to gesture recognition in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)", **(2012)**, pp. 13-18.

[13] L. Chen, H. Wei and J. Ferryman, "A Survey of human motion analysis using depth imagery", Pattern Recognition Letters, vol. 34, **(2013)** 1995-2006.

[14] B. Joseph and Uxbridge Ub Ph, "Markerless Based Human Motion Capture: A Survey", Vision and VR **(2004)**.

[15] I. Rius, "Automatic Learning of 3D Pose Variability in Walking Performances for Gait Analysis", International Journal for Computational Vision and Biomechanics, vol. 1, no. 1, **(2007)**, pp. 33-43.

[16] F. Korc and V. Hlavac, "Detection and Tracking of Humans in Single View Sequences Using 2D Articulated Models", Human Motion: Understanding, Modelling, Capture and Animation, Springer, **(2007)**.

[17] I. Mikic, "Human Body Model Acquisition and Tracking Using Voxel Data", International Journal for Computater Vision, vol. 53, no. 3, **(2003)**, pp. 199-223.

[18] D. Ramanan and D. A. Forsyth, "Finding and Tracking People from the Bottom Up", IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Wisconsin, **(2003)**.

[19] H. Ning, "People tracking based on motion model and motion constraints with automatic initialization", Pattern Recognition, vol. 37, **(2004)**, pp. 1423-1440.

[20] P. Viola, M. J. Jones and D. Snow, "Detecting Pedestrians Using Patterns of Motion and Appearance", International Journal for Computer Vision, vol. 63, no. 2, **(2005)**, pp. 153-161.

[21] J. J. Gonzalez, "Robust tracking and segmentation of human motion in an image sequence", in International Conf. on Acoustics, Speech and Signal Processing, Hong Kong, **(2003)**.

[22] A. D. Sappa, "Prior Knowledge Based Motion Model Representation", Electronic Letters on Computer Vision and Image Analysis, vol. 5, no. 3, **(2005)**, pp. 55-67.

[23] L. Wang, "Silhouette Analysis-Based Gait Recognition for Human Identification", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 12, **(2003)**, pp. 1505-1516.

[24] J. H. Yoo and M. Nixon, "Markerless Human Gait Analysis via Image Sequences", International Society of Biomechanics XIXth Congress, Dunedin NZ, **(2003)**.

[25] C. W. Chu, O. C. Jenkins and M. J. Mataric, "Marker less Kinematic Model and Motion Capture from Volume Sequences", Proceedings of IEEE Computer Vision and Pattern Recognition, Wisconsin, USA, **(2003)**.

[26] T. Y. Tian, C. Tomasi and D. J. Heeger, "Comparision of approaches to egomotion, Computer Vision and Pattern Recognition", Proceedings CVPR'96, IEEE Computer Society Conference, **(1996)**, pp. 315-320.

[27] Prof. W. H. Press, "Gaussian Mixture Models and EM Methods", The University of Texas at Austin, CS 395T, Spring, **(2008)**.

[28] K. P. Karmann and A. Brandt, “Moving object recognition using an adaptive background memory”, V. Cappellini (Ed.), time-Varying Image Processing and Moving Object Recognition, Elsevier, Amsterdam, The Netherlands, vol. 2, **(1990)**.

[29] D. R. K. Brownrigg, “The weighted median Filter”, Commun. ACM, vol. 27, **(1984)**.

[30] G. O. Glentis, “An efficient affine projection algorithm for 2-D FIE adaptive filtering and linear prediction”, Signal Processing, vol. 86, no. 1, **(2006)**, pp. 98-116.

[31] S. S. Cheung and C. Kamath, “Robust techniques for background subtraction in urban traffic video”, EURASIP Journal on Applied Signal Processing, vol. 1, **(2005)**, pp. 2330-2340.

[32] C. R. Wren, A. Azarbayejani, T. Darell and A. P. Pentland, “Pfinder: real-time tracking of the human body”, IEEE PAMI, vol. 19, no. 7, **(1997)**, pp. 780-785.

[33] N. McFarlane and C. Schofield, “Segmentation and tracking of piglets in images”, Machine Vision and Applications, vol. 8, no. 3, pp. 187-193, **(1995)**.

[34] C. Stauffer and W. Grimson, “Adaptive background mixture models for real-time tracking”, CVPR99, vol. 2, no. 252, **(1999)**.

[35] R. T. Collins, “A system for video surveillance and monitoring: VSAM final report”, CMU-RI-TR-00-12, Technical Report, Carnegie Mellon University, **(2000)**.

[36] Y. Kuno, T. Watanabe, Y. Shimosakoda and S. Nakagawa, “Automated detection of human for visual surveillance system”, Proc. of Intl. Conf. on Pattern Recognition, **(1996)**, pp. 865-869.

[37] T. B. Moeslund and E. Granum, “A survey of computer vision-based human motion capture”, Computer Vision and Image Understanding, vol. 81, no. 3, **(2001)**, pp. 231–268.

[38] A. Yilmaz, K. Shafique and M. Shah, “Target tracking in airborne forward looking imagery”, Image Vision Computing, vol. 21, no. 7, **(2003)**, pp. 623–635.

[39] A. Ali and J. Aggarwal, “Segmentation and recognition of continuous human activity”, IEEE Workshop on Detection and Recognition of Events in Video, **(2001)**, pp. 28–35.

[40] S. Ilic and P. Fua, “Generic deformable implicit mesh models for automated reconstruction”, ICCV Workshop on Higher-Level Knowledge in 3-D Modeling and motion Analysis, Nice, France, **(2003)**.

[41] L. Herda, R. Urtasun, P. Fua and A. Hanson, “An automatic method for determining quaternion field boundaries for ball-and-socket joint limits”, International Conference on Automatic Face and Gesture Recognition, Washington DC, USA, **(2002)** May 20–21.

[42] C. Barron and I. A. Kakadiaris, “On the improvement of anthropometry and pose estimation from a single uncalibrated image”, Machine Vision and Applications, vol. 14, no. 4, **(2003)**, pp. 229–236.

[43] D. Weinland, R. Ronfard and E. Boyer, “Free viewpoint action recognition using motion history volumes”, Computer Vision and Image Understanding, vol. 104, **(2006)**, pp. 249–257.

[44] N. Paragios and R. Deriche, “Geodesic active regions and level set methods for supervised texture segmentation”, Int. J. Comput. Vision, vol. 46, no. 3, **(2002)**, pp. 223–247.

[45] D. Comaniciu, V. Ramesh and P. Meer, “Kernel-based object tracking”, IEEE Trans. Patt. Analy. Mach. Intell., vol. 25, **(2003)**, pp. 564–575.

[46] J. Carranza, C. Theobalt, M. Magnor and H.-.P. Seidel, “Free-viewpoint video of human actors”, ACM SIGGRAPH, **(2003)**, pp. 565-577.

[47] A. Micilotta, E. Ong and R. Bowden, “Detection and tracking of humans by probabilistic body part assembly”, British Machine Vision Conference, Oxford, UK, **(2005)**.

[48] H. Lim, V. I. Morariu, O. I. Camps and M. Sznaier, “Dynamic appearance modeling for human tracking”, Computer Vision and Pattern Recognition, New York, USA, **(2006)**, pp. 17–22.

[49] C. Tsai, “Intelligent Post –processing via bounding box-based morphological operations for moving objects detection”, Advanced research in Applied Artificial Intelligence, Lecture Notes in Computer Science, vol. 7345, **(2012)**, pp. 647-657.

[50] C. Ridder, O. Munkelt and H. Kirchner, “Adaptive background estimation and foreground detection using Kalman-filtering”, **(1995)**.

[51] I. Haritaoglu, D. Harwood and L. S. Davis, “w^4: Real-Time Surveillance of People and their activities”, IEEE Trans. PAMI., vol. 22, **(2000)**, pp. 809-830.

[52] N. Friedman and S. Russell, “Image segmentation in video Sequence: A probabilistic approach in the thirteeth conf. in Uncertainity in artificial Intelligence”, Brown University, **(1997)**.

[53] J. Shi and J. Malik, “Normalized cuts and image segmentation”, IEEE transactions on PAMI, **(2000)**.

[54] S. J .McKenna, S. Wang and Y. Chen, “Human tracking by adaptive Kalman filtering and multiple kernels tracking with projected gradients”, Proceedings of IEEE fifth ACM/IEEE International conf. on distributed smart cameras (ICDSC), Ghent, Belgium, vol. 23-26, **(2011)**, pp. 1-6.

[55] X. Huang, Y. Sun, D. Metaxas, F. Sauer and C. Xu, “Hybrid image registration based on configural matching of scale-invariant salient region features”, Computer Vision and Pattern Recognition Workshop, CVPRW'04,IEEE, **(2004)**, pp. 167-177.

[56] S. Cherla, K. Kulkarni, A. Kale and V. Ramasubramanian, “Towards fast, view-invariant human action recognition”, Proceeding of Computer Vision and Pattern Recognition Workshops, 2008. CVPRW '08. IEEE, **(2008)**.

[57] D. Weinland, E. Boyer and R. Ronfard, “Action recognition from arbitrary views using 3D examplars”, Proc. ICCV, **(2007)**.

[58] L. Sigal, M. Isard, H. Haussecker and M. Black, “Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation”, Int. J. Comput. Vis., vol. 98, **(2012)**, pp. 15–48.

[59] J. Lee, R. Sandhu and A. Tanenbaum, “Particle filters and Occlusion Handling for rigid 2D- 3D Pose Tracking”, Comt. Vis., Image Understanding, vol. 117, no. 8, **(2013)** August 1, pp. 922-93.

[60] D. Beymer and K. Konolige, “Real-time tracking of multiple people using continuous detection”, IEEE International Conference on Computer Vision (ICCV) Frame-Rate Workshop, **(1999)**.

[61] D. Cremers, T. Kohlberger and C. Schnorr, “Non-linear shape statistics in mumford-shah based segmentation”, European Conference on Computer Vision (ECCV), **(2002)**.

[62] C. Harris and M. Stephens, “A combined corner & edge detector”, 4th Alvey Vision Conference, **(1988)**, pp. 147-151.

[63] I. Laptev and T. Lindeberg, “Space–time interest points”, Proceedings of the International Conference on Computer Vision (ICCV’03), Nice, France, vol. 1, **(2003)**, pp. 432–439.

[64] T. Kadir and M. Brady, “Scale saliency: a novel approach to salient feature and scale selection”, Proceedings of the International Conference on Visual Information Engineering (VIE), Guildford, United Kingdom, **(2003)**, pp. 25–28.

[65] P. Dollár, V. Rabaud, G. Cottrell and S. Belongie, “Behavior recognition via sparse spatio-temporal features”, Proceedings of the International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS’05), Beijing, China, **(2005)**, pp. 65–72.

[66] G. Willems, T. Tuytelaars and L. J. Van Gool, “An efficient dense and scale-invariant spatio temporal interest point detector”, Proceedings of the European Conference on Computer Vision (ECCV’08) – part 2, Lecture Notes in Computer Science, Marseille, France, no. 5303, **(2008)**, pp. 650–663.

[67] H. Wang, M. Muneeb Ullah, A. Kläser, I. Laptev and C. Schmid, “Evaluation of local spatiotemporal features for action recognition”, Proceedings of the British Machine Vision Conference, London, United Kingdom, **(2009)**.

[68] R. C. Schuldt, I. Laptev and B. Caputo, “Recognising human actions: a local SVM approach”, Proceedings of the International Conference on Pattern Recognition (ICPR’04), Cambridge, United Kingdom, **(2004)**, pp 32-36.

[69] J. C. Niebles, H. Wang and L. Fei-Fei, “Unsupervised learning of human action categories using spatial–temporal words”, International Journal of Computer Vision (IJCV), vol. 79, no. 3, **(2008)**, pp. 299-318.

[70] H. Jhuang, T. Serre, L. Wolf and T. Poggio, “A biologically inspired system for action recognition”, Proceedings of the International Conference on Computer Vision (ICCV’07), Rio de Janerio, Brazil, **(2007)** October, pp. 1-8.

[71] N. Ikizler and P. Duygulu, “Histogram of oriented rectangles: a new pose descriptor for human action recognition”, Image and Vision Computing, vol. 27, no. 10, **(2009)**, pp. 1515-1526.

[72] S. Nowozin, G. Bakır and K. Tsuda, “Discriminative subsequence mining for action classification”, Proceedings of the International Conference on Computer Vision (ICCV’07), Rio de Janeiro, Brazil, **(2007)** October, pp. 1-8.

[73] I. Laptev and P. Pérez, “Retrieving actions in movies”, Proceedings of the International Conference On Computer Vision (ICCV’07), Rio de Janeiro, Brazil, **(2007)** October, pp. 1–8.

[74] M. Bregonzio, S. Gong and T. Xiang, “Recognising action as clouds of space–time interest points”, Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR’09), Miami, FL, **(2009)** June, pp. 1-8.

[75] P. Scovanner, S. Ali and M. Shah, “A 3-dimensional SIFT descriptor and its application to action recognition”, Proceedings of the International Conference on Multimedia (MultiMedia’07), Augsburg, Germany, **(2007)** September, pp. 357–360.

[76] H. Zhang, “Image segmentation evaluation: A survey of unsupervised methods”, Computer Vision and Image Understanding, vol. 110, **(2008)**, pp. 260–280.

[77] L. Yang, F. Albregtsen, T. Lonnestad and P. Grottum, “A supervised approach to the evaluation of image segmentation methods”, Computer Analysis of Images and Patterns, **(1995)**, pp. 759-765.

[78] S. Chabrier, H. Laurent, B. Emile, C. Rosenburger and P. Marche, “A comparative study of supervised evaluation criteria for image segmentation”, EUSIPCO, **(2004)**, pp. 1143–1146.

[79] W. A. Yasno, J. K. Mui and J. W. Bacus, “Error measure for scene segmentation”, Pattern Recognition, vol. 9, **(1977)**, pp. 217–231.

[80] J. Weszka and A. Rosenfeld, “Threshold evaluation techniques”, IEEE Transactions on Systems, Man and Cybernetics, vol. 8, no. 8, **(1978)**, pp. 622–629.

[81] S. Lee, S. Chung and R. Park, "A comparative performance study of several global thresholding techniques for segmentation", Computer Vision, Graphs and Image Processing, vol. 52, **(1990)**, pp. 171–190.

[82] Y. Lim and S. Lee, "On the color image segmentation algorithms based on the thresholding and fuzzy c-means techniques", Pattern Recognition, vol. 23, **(1990)**, pp. 935–952.

[83] M. Wollborn and R. Mech, "Refined procedure for objective evaluation of video generation algorithms", **(1998)**.

[84] K. Strasters and J. Gerbrands, "Three-dimensional image segmentation using split, merge and group approach", Pattern Recognition Letters, vol. 12, **(1991)**, pp. 307–325.

[85] N. Pal and D. Bhandari, "Image thresholding: some new techniques", Signal Processing, vol. 33, no. 2, **(1993)**, pp. 139-158.

[86] Y. Zhang and J. Gerbrands, "Objective and quantitative segmentation evaluation and comparison", Signal Processing, vol. 39, **(1994)**, pp. 43–54.

[87] P. Correia and F. Pereira, "Objective evaluation of video segmentation quality", IEEE Transactions on Image Processing, vol. 12, no. 2, **(2003)**, pp. 186-200.

[88] M. Van Droogenbroeck and O. Barnich, "Design of statistical measures for the assessment of image segmentation schemes", Proceedings of International Conference on Computer Analysis of Images and Patterns, vol. 280, **(2005)**.

[89] F. Ge, S. Wang and T. Liu, "Image-segmentation evaluation from the perspective of salient object extraction", Proceedings of IEEE Internatioanl Conference on Computer Vision and Pattern Recognition, vol. I, **(2006)**, pp. 1146–1153.

[90] S. Chabrier, B. Emile, H. Laurent, C. Rosenberger and P. Marche, "Unsupervised evaluation of image segmentation application to multispectral images", Proceedings of the 17th international conference on pattern recognition, **(2004)**.

[91] P. Correia and F. Pereira, "Stand-alone objective segmentation quality evaluation", JASP 2002, vol. 4, **(2002)**, pp. 389–400.

[92] H. Zhang, "Image segmentation evaluation: A survey of unsupervised methods", Computer Vision and Image Understanding, vol. 110, **(2008)**, pp. 260–280.

[93] N. Otsu, "A threshold selection method from gray-level histograms", IEEE Transactions on Systems, Man and Cybernetics, vol. 9, no. 1, **(1979)**, pp. 62–66.

[94] C. Rosenberger and K. Chehdi, "Genetic fusion: application to multicomponents Image segmentation", Proceedings of ICASSP-4Istanbul, Turkey, **(2000)**.

[95] C. E. Erdem, B. Sanker and A. M. Tekalp, "Performance measures for video object segmentation and tracking", IEEE Transactions on Image Processing, vol. 13, **(2004)**, pp. 937–951.

[96] G. Welch and G. Bishop, "An introduction to the Kalman Alter", from http://www.cs.unc.edu, UNC-Chapel Hill, TR95-041, **(2000)** November.

[97] M. Isard and A. Blake, "Condensation—conditional density propagation for visual tracking", Int. J. Comput. Vision, vol. 29, no. 1, **(1998)**, pp. 5–28.

[98] V. PavloviWc, J. M. Rehg, T.-J. Cham and K. P. Murphy, "A dynamic Bayesian network approach to Agure tracking using learned dynamic models", Proceedings of the International Conference on Computer Vision, **(1999)**, pp. 94–101.

[99] M. H. Yang and N. Ahuja, "Recognizing hand gesture using motion trajectories", Proc. of IEEE CS Conference on Computer Vision and Pattern Recognition, **(1999)**, pp. 468-472.

[100] C. Barrón and I. A. Kakadiaris, "Estimating anthropometry and pose from a single uncalibrated image", Computer Vision and Image Understanding, vol. 81, no. 3, **(2001)**, pp. 269-284.

[101] E. J. Ong and S. Gong, "Tracking 2D-3D human models from multiple views", Proc. of International Workshop on Modeling People at ICCV, **(1999)**.

[102] T. Boult, "Frame-rate multi-body tracking for surveillance", DARPA Image Understanding Workshop, Monterey, Calif. San Francisco: Morgan Kaufmann, **(1998)** November.

[103] L. Wang, "Recent developments in human motion analysis", Pattern Recognition, vol. 36, **(2003)**, pp. 585-601.

[104] A. Yilmaz, O. Javed and M. Shah, "Object tracking: A survey", ACM Comput. Surv., Article 13, vol. 38, no. 4, **(2006)** December, pp. 45.

[105] T. B. Moesland, "A survey of advances in vision-based human motion capture and analysis", Computer Vision and Image Understanding, vol. 104, **(2006)**, pp. 90-126.

[106] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake, "Real-Time Human Pose Recognition in Parts from Single Depth Images", IEEE Conference on Computer Vision and Pattern Recognition, **(2011)**.

[107] V. K. Singh, F. M. Khan and R. Nevatia, "Multiple Pose Context Trees for estimating Human Pose in Object Context", IEEE Conference on Computer Vision and Pattern Recognition Workshops, **(2010)**, pp. 17–24.

[108] S. Singh, A. Gupta and A. Effros, "Unsupervised discovery of mid-level discriminative patches", ECCV, Berlin, Heidelberg, Springer-Verlag, **(2012)**.

[109] D. Ramanan, "Learning to parse images of articulated bodies", Ad-vances in Neural Information Processing Systems, **(2007)**, pp. 1129–1136.

[110] S. Johnson and M. Everingham, "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation", Proceedings of the 21st British Machine Vision Conference, I, **(2010)**.

[111] H. Wang and D. Koller, "Multi-level inference by relaxed dual decomposition for human pose segmentation", IEEE Conference on Computer Vision and Pattern Recognition, **(2011)**, pp. 2433–2440.

[112] T. B. Moeslaund, A. Hilton and V. Kruger, "A survey of advances in vision based human motion capture and analysis", Computer vision and image understanding, vol. 104, **(2006)**, pp. 90-126.

[113] H. H. Nagel, "From image sequences towards conceptual descriptions", Image and Vision Computing, vol. 6, no. 2, **(1988)**, pp. 59-74.

[114] A. Bobbick, "Movement, Activity and Action: the role of knowledge in the perception of motion", Philosophical Transactions of Royal Society of Landon, vol. 352, **(1997)**, pp. 1257-1265.

[115] C. Schuldt, I. Laptev and B. Caputo, "Recognizing human actions: a local SVM approach", International Conf. on Pattern Recognition, **(2004)**, pp. 32-36.

[116] L. Gorelick, M. Blank, E. Shectman, M. Irani and R. Basri, "Actions as shape-time shapes", Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, **(2007)**, pp. 2247-2253.

[117] R. B. Fisher, "Pets04 surveillance ground truth dataset", Proc. Sixth IEEE Int. Work, on Performance Evaluation of Tracking and Surveillance (PETS04), **(2004)**.

[118] A. T. Nghiem, F. Bremond, M. Thonnat and R. Ma, "New evaluation approach for video processing algorithms", WMVC 2007 IEEE Workshop on Motion and Video, computing, **(2007)**.

[119] Centre for Biometrics and security Research, Casia gait database, **(2011)**.

[120] J. Yuan, Z. Liu and Y. Wu, "Discriminative subvolume search for efficient action detection", IEEE Conf. on Computer Vision and Pattern Recognition, **(2009)**.

[121] INRIA, Inria xmas motion acquisition sequences (ixmas), **(2011)** November.

[122] S. Singh, S. A. Velastin and H. Ragheb, "Muhavi: a multicamera human action video dataset for evaluation of action recognition methods", 2nd Workshop on Activity Monitoring by Multi-Camera Surveillance Systems (AMMCSS), **(2010)**.

[123] Center for Biometrics and Security research, Casia gait database, **(2011)** November.

[124] Kitware, Virat video dataset, http://www.kitware.com, **(2012)** January.

[125] The Imagelab Laboratory of the University of Modena and Reggio Emilia, Visor (video surveillance online repository), **(2011)** November.

[126] Serre lab, Hmdb: a large video database for human motion recognition, http://www.serre-lab.clps.brown.edu/resources/HMDB/index.htm, **(2011)** November.

[127] J. M. Chaquet, E. J. Carmona and A. Fernandez-Caballere, "A Survey of video datasets for human action and activity recognition", Computer Vision and Image Understanding, vol. 117, **(2013)**, pp. 633-659.

[128] K. K. Kim, S. H. Cho, H. J. Kim and J. Y. Lee, "Detecting and Tracking Moving Object Using an Active Camera", Proceedings of IEEE 7th International Conference on Advanced Communication Technology (ICACT), 21–23, Phoenix Park, Dublin, Ireland, vol. 2, **(2005)**, pp. 817–820.

Authors

S. Nissi Paul is a currently pursuing research in Artificial Intelligence and computer Vision from the Department of Computer Science & Engineering and Information Technology, Don Bosco College of Engineering and Technology of Assam Don Bosco University, Assam, India. She has completed M.Phil (Comp. Sc.) from Bharatidasan University in 2005.

Dr Y. Jayanta Singh is working as Associate Prof. and Head of Department of Computer Science & Engineering and Information Technology, Don Bosco College of Engineering and Technology of Assam Don Bosco University, Assam, India. He has received Ph.D.(Comp. Sc. and IT) from Dr Babasaheb Ambedkar Marathwada University, Maharashtra in 2004. He has worked with Swinburne University of Technology (AUS) at (Malaysia campus), Misurata University(North Africa), Skyline University (Dubai), Keane Inc (Canada) etc. His areas of research interest are Real Time Distributed Database, Cloud Computing, Digital Signal processing, Expert Systems etc. He has published several papers in International and National Journal and Conferences. He is presently executing AICTE sponsored Research Project.