THE UNIVERSITY OF WARWICK

University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

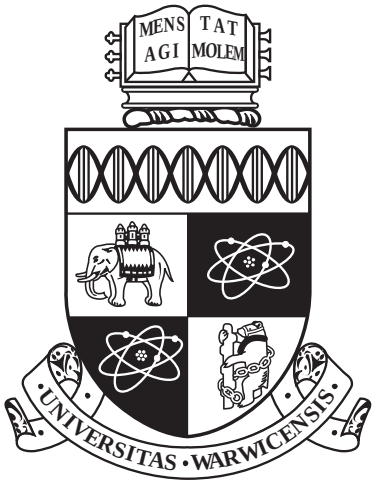
A Thesis Submitted for the Degree of PhD at the University of Warwick

http://go.warwick.ac.uk/wrap/69575

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



Marker-less Human Body Part Detection, Labelling and Tracking for Human Activity Recognition

by

FAISAL AZHAR

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

School of Engineering and Department of Computer Science

March 2015

THE UNIVERSITY OF
WARWICK

Abstract

This thesis focuses on the development of a real-time and cost effective marker-less computer vision method for significant body point or part detection (i.e., the head, arm, shoulder, knee, and feet), labelling and tracking, and its application to activity recognition. This work comprises of three parts: significant body point detection and labelling, significant body point tracking, and activity recognition. Implicit body models are proposed based on human anthropometry, kinesiology, and human vision inspired criteria to detect and label significant body points. The key idea of the proposed method is to fit the knowledge from the implicit body models rather than fitting the predefined models in order to detect and label significant body points. The advantages of this method are that it does not require manual annotation, an explicit fitting procedure, and a training (learning) phase, and it is applicable to humans with different anthropometric proportions. The experimental results show that the proposed method robustly detects and labels significant body points in various activities of two different (low and high) resolution data sets. Furthermore, a Particle Filter with memory and feedback is proposed that combines temporal information of the previous observation and estimation with feedback to track significant body points in occlusion. In addition, in order to overcome the problem presented by the most occluded body part, i.e., the arm, a Motion Flow method is proposed. This method considers the human arm as a pendulum attached to the shoulder joint and defines conjectures to track the arm

since it is the most occluded body part. The former method is invoked as default and the latter is used as per a user's choice. The experimental results show that the two proposed methods, i.e., Particle Filter and Motion Flow methods, robustly track significant body points in various activities of the above-mentioned two data sets and also enhance the performance of significant body point detection. A hierarchical relaxed partitioning system is then proposed that employs features extracted from the significant body points for activity recognition when multiple overlaps exist in the feature space. The working principle of the proposed method is based on the relaxed hierarchy (postpone uncertain decisions) and hierarchical strategy (group similar or confusing classes) while partitioning each class at different levels of the hierarchy. The advantages of the proposed method lie in its real-time speed, ease of implementation and extension, and non-intensive training. The experimental results show that it acquires valuable features and outperforms relevant state-of-the-art methods while comparable to other methods, i.e., the holistic and local feature approaches. In this context, the contribution of this thesis is three-fold:

- Pioneering a method for automated human body part detection and labelling.
- Developing methods for tracking human body parts in occlusion.
- Designing a method for robust and efficient human action recognition.

Acknowledgments

I owe a great thanks to my supervisors, Dr. Tardi Tjahjadi at Image Processing and Expert Systems (IPES) Laboratory, School of Engineering, and Prof. Chang-Tsun Li at Image Processing Laboratory, Department of Computer Science, University of Warwick. It is a pleasure to have worked under their supervision and I consider myself lucky for having been their student.

Thanks to Graduate School for awarding me full time Warwick Postgraduate Research Scholarship, and special thanks to my supervisor Dr. Tardi Tjahjadi for arranging the School of Engineering Bursary, which allowed me to fulfil my dream of pursuing a PhD research in computer vision and machine learning.

Thanks to Prof. Alison Rodger at Molecular Organisation and Assembly in Cells (MOAC), Department of Chemistry, University of Warwick, for providing moral support and motivation to complete my research in time. She is a wonderful person, and I have been inspired by her unconditioned dedication to students.

Thanks to my colleagues from Image Processing and Expert Systems (IPES) Laboratory, School of Engineering, and Image Processing Laboratory, Department of Computer Science, with whom I enjoyed insightful discussions on research topics. Thanks to Mr. Faisal Khan who has been like a brother during my first year of PhD. Thanks to my closest friend Mr. Salman Shahid who gave me wonderful company during this PhD journey.

And thanks to my parents, i.e., Zarina Azhar and Azhar Ali, my wife, i.e., Elena Marchis and my brother, i.e., Ahsan Azhar for their love and support.

Declarations

I hereby declare that I have produced this thesis without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This thesis has not previously been presented in identical or similar form to any other publication unless otherwise specified. The thesis work was conducted from 2010 to 2014 under the supervision of Dr. Tardi Tjahjadi at Image Processing and Expert Systems (IPES) Laboratory, School of Engineering, and Prof. Chang-Tsun Li at Image Processing Laboratory, Department of Computer Science, University of Warwick.

Coventry, United Kingdom.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Videos are cheaply available and open up opportunities for developing computer vision based applications. Among so many potential applications my work will focus on the recognition of human activities because this can enable applications such as unusual activity detection, surveillance, home-based rehabilitation, behaviour recognition, location estimation, etc. The fact that humans are the most captured objects in the majority of the videos provides a strong motivation for automated analysis and interpretation of human activities. Therefore, this thesis focuses on developing computer vision methods for detecting, labelling and tracking significant body points or parts (i.e., the head, arm, shoulder, knee, and feet), and recognizing human activities.

1.1 Justification for the research

Computer vision methods provide automated, low cost, efficient and effective solutions to detect, label, and track human significant body parts, and recognize human activities [1–3]. These methods do not require subject cooperation, large experimental set-up time, specialized environment, etc., and thus can be used for various applications.

A real-time, accurate, fully automated, universal (applicable to different age, gender, ethnicity, etc.), and complete method that is able to detect, label and track human significant body points, and then utilize them for the task of human activity recognition does not exist. This is because most of the previous computer vision methods to detect, label and track human significant body points are either computationally expensive or require an intensive training phase. Also, the methods that

are computationally inexpensive or do not require training are not accurate. In addition, these methods are not always fully automated and often require some manual initialization. Moreover, the methods that use arbitrary predefined body models might not be applicable to humans with different anthropometric proportions.

In this context, this research thesis aims to fill the above-mentioned gap in the literature by investigating novel computer vision methods in order to develop a real-time, accurate, fully automated, universal, and complete human body part detection, labelling and tracking framework for human activity recognition.

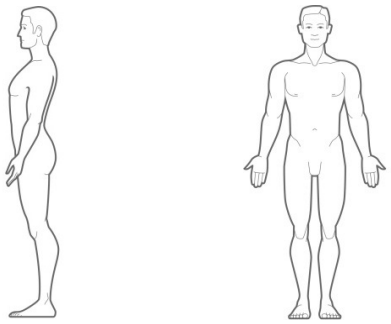
1.2 Research problem and objectives

Computer vision methods use marker-less techniques to detect and label significant human body points. The previous research work on marker-less significant body point detection can be broadly divided into the model-based (prior model) or model-free (no prior model) approaches [3, 4]. The former approach requires a fitting procedure, manual annotation, and numerous predefined models which are time consuming processes, while the latter tends to be less accurate. The arbitrary predefined models might not always be a proper fit for the human subjects as the human body proportions vary with respect to age, ethnicity, gender, etc. However, the empirical studies on the human anthropometry [5,6] allow definition of more accurate human body proportions that can cover the majority of the world population. So far, anthropometry has only been used in a semi-automated manner to detect and label human body parts for merely stand postures [7, 8] since its application in complex activities is not an easy task.

In order to address the above-mentioned drawbacks of the previous marker-less methods, the objectives of the first part of this thesis, i.e., Human Body Part Detection and Labelling, are as follows:

- To investigate the application of the human anthropometry (measurement of human body proportions) and kinesiology (study of human movement) in order to define more accurate human body models.

- To explore a novel, efficient, robust, and fully automated marker-less method that does not require explicit model fitting and manual annotation to detect and label human significant body points in various activities observed from a profile view.



(a) (b)

Figure 1.1: (a) Profile view and (b) Front view.

When observed from the profile view, the human activities such as Walk, Run, Bend, etc., might contain considerable rapid motion and self-occlusion of the human body parts e.g., arms and legs. Also, the human body can attain various postures and perform numerous activities due to its high dimensionality, i.e., degrees of freedom of its motion. Hence, the foreground segmentation of human body is affected and might contain artefacts that will result in false detection of significant body points. Therefore, the profile view in Fig. 1.1 (a) is chosen over the front view in Fig. 1.1 (b) since it presents a more challenging scenario to label and track significant body points and human activity recognition. A robust method for significant body point detection and labelling should be able to recover the positions of the body parts during occlusion. Thus, it is imperative to incorporate a tracking method that deals with occlusion, variation of illumination, rapid motion, etc. The non-Gaussian, i.e., multimodal distribution, and non-linear, i.e., the system is a function of polynomial degree higher than one, assumption of the Particle Filter method [9, 10] make them suitable for visual tracking.

The Particle Filter draws samples/particles from the uniform distribution and assigns them equal weights. It then uses a model that represents the current system to predict the new state. Finally, the new state is updated based on the measurement, i.e., observation, to re-assign weights to the particles. However, the standard Particle Filter struggles to predict accurately when there are no measurements, i.e., observation of significant body points, in the image.

In order to address the above-mentioned challenges and the inability of the standard Particle Filter method to track in occlusion, the objectives of the second part of this thesis, i.e., Human Body Parts Tracking, are as follows:

- To examine new ways to enhance the capability of the standard Particle Filter to track during occlusion.
- To apply the pendulum physics in order to develop a new tracker for predicting the arm which is the most occluded significant body point and thus most challenging body part to track.
- To analyze whether the significant body point detection and labelling is improved by using a tracking method.

Significant body points can be utilized for various tasks such as activity recognition, motion analysis of sit-stand for elderly people, realistic animation of human body models, surveillance, etc. Human activity recognition methods can be broadly divided [1, 3, 11] into: holistic (a), local feature (b), and model-based (c) or model-free (d). The holistic method uses shape or optical flow information, while the local feature method uses descriptors of local regions to define an activity. The extraction of shape and optical information from each frame of video sequence is a computationally expensive procedure. The learning of local descriptors require intensive training phase in order to perform accurate recognition. In contrast, the model-based approach fits a predefined model to human silhouette while the model-free uses body characteristics such as orientation, proportion, motion etc., to recognize activities. They are computationally inexpensive in comparison to holistic and local feature methods but lack accuracy. Also, many human activity recognition methods [12–17], cannot accurately discern, without intensive training, similar activities such as walk, run, jump, etc. This is due the fact that the feature space for very similar activities includes considerable overlaps. Previous methods such as relaxed hierarchy [18], only deal with a two overlap class separation problem in the spatial domain and hence are not applicable to multiple overlaps in the spatio-temporal domain.

In order to address these above-mentioned drawbacks of the previous activity recognition methods, the objectives of the third part of this thesis, i.e., Human Activity Recognition, are as follows:

- To explore the use of the significant body points in order to build innovative feature descriptors that enable to discern human activities.
- To investigate a novel relaxed hierarchy based method which tackles the multiple overlaps problem in the feature space for efficient and robust human

activity recognition.

1.3 Major contributions

The main contributions of this work are as follows.

- This is the first work to provide both quantitative and qualitative evaluation of significant body point detection. The quantitative evaluation of significant body point detection, labelling and tracking has not been done in most of the relevant previous works [19–22].
- This is also the first work to perform the ground truth mark-up of significant body points on both the Weizmann and MuHAVi data sets for quantitative evaluation. There was no state-of-the-art data set available that contained ground truth significant body points.
- The novel proposition of the Implicit Body Models (IBMs) that are derived by combining the science from Anthropometry, Kinesiology, and Biomechanics studies. IBMs contain the knowledge of the body part positioning, range of motion of human body parts and understanding of type of motion. The knowledge from the Implicit Body Models are utilized to robustly detect and label significant body points and to achieve real-time efficiency. In contrast to previous works, it does not require an explicit fitting procedure and a manual annotation.
- An innovative Particle Filter method based on the temporal Markov chain framework to perform prediction during occlusion. The proposed Particle Filter utilizes the temporal information of the previous observation and estimation (kept in memory) via a feedback to predict human body parts in occlusion. It predicts more accurately in occlusion than the standard Particle Filter.
- A new motion flow prediction method specifically designed for arm since it is the most occluded limb. It considers the human arm as a pendulum attached to the shoulder joint producing curvilinear motion and derives linear equations from the pendulum physics to predict arm in occlusion.
- The significant body point detection, labelling and tracking proposed in this work is a low cost solution to VICON motion capture technology and does

not require subject cooperation. It automatically determines significant body points, create a 2D stick body model and extracts motion of the limbs.

- This significant body point detection, labelling and tracking proposed in this work is also an alternative to KINECT and has been shown to work on both low and high resolution videos without any depth information.
- The method in [21] is extended by introducing two features, i.e., the leg power and torso power, in addition to the leg angle and torso angle to create a robust feature descriptor for recognizing very similar activities.
- A hierarchical relaxed partitioning system method that combines relaxed hierarchy and hierarchical strategy methods and uses an innovative majority voting scheme to discern easily confused activities with real-time speed and without intensive training. Most of the previous methods [13, 15, 16, 22, 23] confuse very similar activities and require either computationally expensive feature extraction or intensive training to overcome this issue.

1.4 Outline of the thesis

The outline of the entire thesis is as follows.

Chapter 2 covers the detailed literature review on human motion analysis and tracking approaches. It also described the general approaches to human activity recognition. In addition, it explains and illustrates the experimental data sets used in this thesis.

Chapter 3 describes the use of the anthropometry and kinesiology information to develop novel implicit body models. It explains the proposed marker-less approach which uses computer vision methods based on implicit body models to detect and label human significant body points, i.e., the head, arm, shoulder, knee, and feet in various human activities. Next, it presents the procedure to construct 2D stick figures from the detected and labelled significant body points. The accuracy of the proposed method is established by evaluating its ability to detect and label significant body points in various activities of two different resolution data sets, i.e. low (180 x 144) and high (720 x 576).

Chapter 4 details the improvements made in the standard Particle Filter method for visual tracking and presents two tracking methods, i.e., the Particle Filter with memory and feedback, and motion flow, to predict significant body

points during occlusion. It describes how the proposed Particle Filter addresses the limitations of the standard Particle Filter to track in occlusion. In addition, it introduces the concept of using a pendulum for the human arm prediction based on the motion flow. The accuracy of the proposed methods is established by evaluating their ability to robustly predict the significant body points in occlusion or in missed detections. The impact of the tracking methods on the performance of the significant body point detection and labelling is also demonstrated in this chapter.

Chapter 5 presents the proposed hierarchical relaxed partitioning system solution for human activity recognition. It explains the process of building feature descriptors by using the human significant body points. In addition, it details a hierarchical relaxed partitioning system method for human activity recognition. The discerning ability of the feature descriptors is shown on the training data set. The accuracy of this proposed method is authenticated by evaluating its ability to discern various very similar human activities that have significant multiple overlaps in the feature space.

Chapter 6 concludes the entire thesis by highlighting the efficiency of the proposed computer vision methods for human activity recognition. It suggests the implications of the research undertaken and its various applications. It also speculates on the future directions and developments.

Chapter 2

Related Work and Datasets

2.1 Human body part detection

Human body part detection involves estimation of the location and orientation of joints of a human body. This section focuses as to why out of the two broad approaches to human body part detection, i.e., marker-based, and marker-less; the latter is preferred over the former and builds up a discussion of pros and cons of the two approaches. This section also explains the model-based technique of the marker-less human motion analysis approach and why this technique has been chosen in lieu of the model-free based approach.

2.1.1 Marker-based approach

The marker-based approach estimates human body motion by determining coordinates of a set of markers attached on particular points of the human body, as shown in Fig. 2.1 (a) and (b) [1–3]. The coordinates of a set of active or passive markers attached on the anatomical landmarks of human body whose spatial trajectories are to be estimated and computed by a stereo-photogrammetric method [1, 24]. The joint kinematics is estimated by reconstructing the 3D position of the attached markers and conjecturing the fundamental human body model, as shown in Fig. 2.1 (c) [2]. In the recent years, the marker-based human motion capture approach has been used commercially for biometrics (gait recognition), special effects in motion pictures, clinical and rehabilitative fields, etc. [1]. The growing significance of healthcare for elderly and disabled persons could be seen in the enormous concentration of European Commission research on the area of ambient assisted living for the ageing

(a) (b) (c)

Figure 2.1: Marker-based approach. (a) An actor wearing a suit with reflective infra-red markers, i.e., the small white balls in the middle of the black patches. The motion of the actor is recorded by several cameras, (b) HumanEva data set subject performing an activity, and (c) the corresponding model fitting that detects body parts such as head, torso, arms, legs, etc. [2].

society [25,26]. Every year the number of casualties and injuries amongst the ageing and disabled is increasing especially in household incidents while performing routine but difficult activities. Thus, applications such as surveillance, animation, and assisted living bring new challenges to the marker-based approach [4, 20, 25, 26].

2.1.1.1 Pros and Cons

The commercially available marker-based approaches are accurate with a root mean square error below 6mm for 3D reconstruction of the position of markers [4]. The existing technologies use sensors to prevent injuries to persons [25] or to generate an alarm to a surveillance team in case of suspicious/ abnormal beaviour. In order to achieve this, the subject needs to wear an electronic sensor that keeps record of his or her movement. The difficulty with using sensors is that they are required to be worn at all times which is not possible for any outdoor applications e.g., surveillance, sports etc. An elderly person may forget to wear the sensor due to his or her age when going outside [25], while sensor fitting is unsuited for surveillance because it requires subject cooperation. Moreover, attaching markers is not only a time consuming exercise, but it also restricts the movement of the subject. In addition, it is not easy to use equipment for people of all ages and this requires inter-session repeatability of measurement. The marker based approach requires expensive specialized hardware, environment and is intrusive for the subject. Furthermore, the sensors may be affected by the environment and may generate false alarms [4, 20, 25].

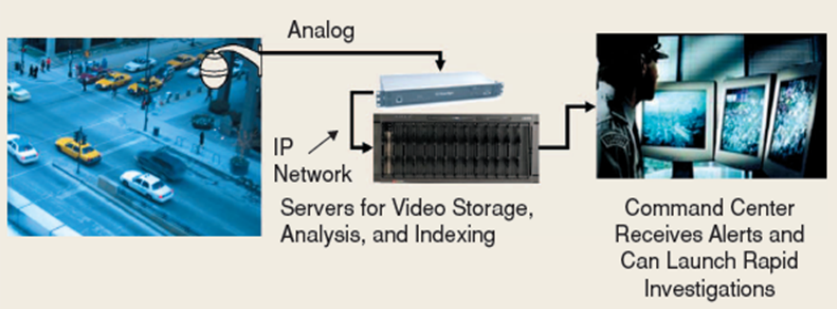


Figure 2.2: Marker-less approach. Cameras are used to relay information about the subjects, i.e., humans, cars, etc., and servers store the videos. The video analysis software provides real-time alerts [28].

2.1.2 Marker-less approach

Marker-less approaches are employed by several researchers to make up for the limitations of the marker based approaches [1, 3]. These are also available commercially for private and public offices, defence installations, as well as domestic usage; therefore they are deemed preferable for this type of research. In the marker-less approach, cameras are used to relay information about the movements and whereabouts of the subjects, as shown in Fig. 2.2. A video analysis system is used in such systems to recognize human behaviours, anomalies, etc. [27]. Recently, smart camera based systems have been proposed for surveillance, assisted living, behaviour recognition, etc. [28]. These systems comprise of cameras, video storage servers, and a command centre. The video from a camera is converted into internet protocol stream by video encoders and accumulated on a server by a video managing framework for controlling video storage. The event videos are stored in database with appropriate indexing with respect to the camera and event attributes for rapid retrieval. The video analysis mechanism executes on the server and provides real-time alerts for user defined incidents and allows swift search of specified events, as shown in Fig. 2.2 [28].

The deployment of smart camera based systems requires sophisticated technology, configuration and tuned alarm systems, and privacy protection mechanisms [28, 29]. The current sophisticated smart camera based systems comprise of the following methods: plug-and-play analysis, object recognition and tracking, object and colour categorization, alert description and identification, database incident indexing, and seek and retrieval [29]. At the core of the smart camera node, statisti-

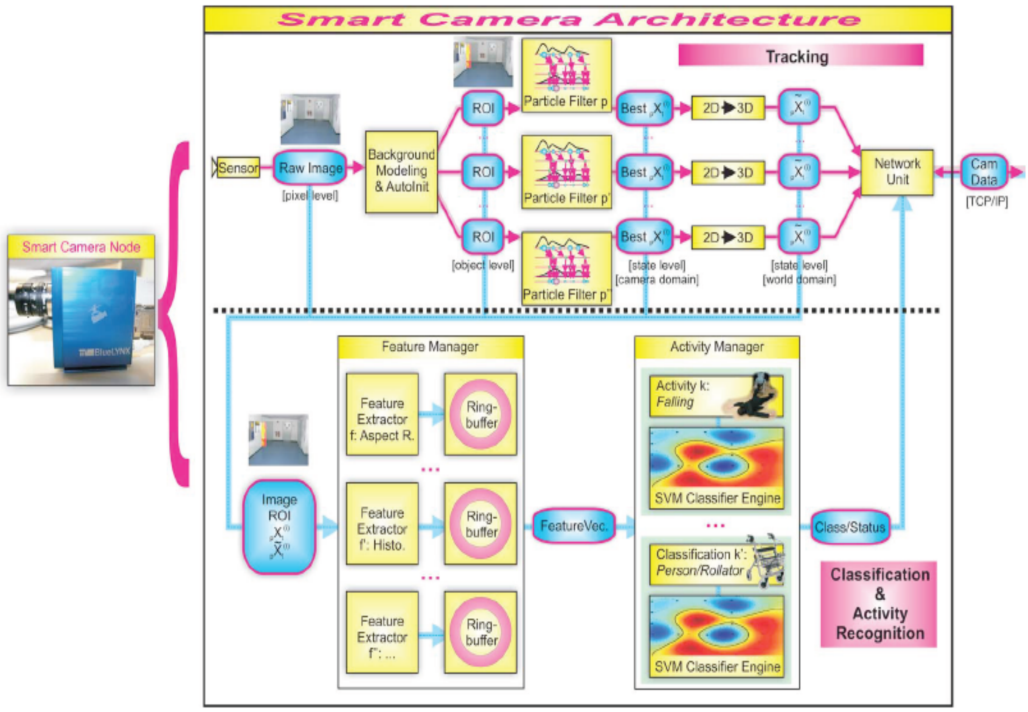
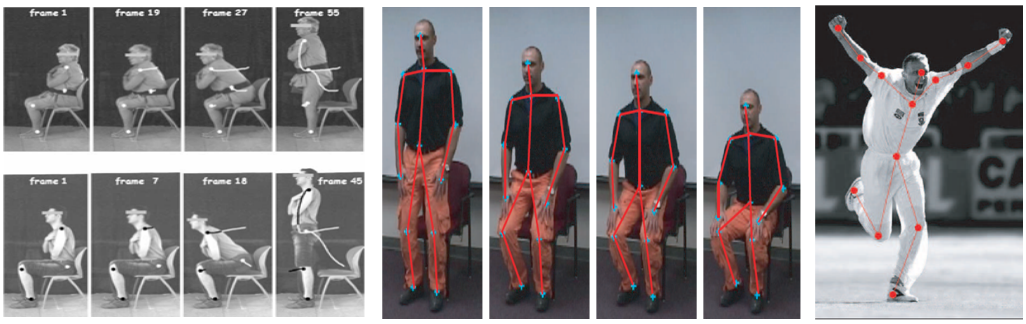


Figure 2.3: Smart camera node architecture (reprinted from [29]).

cal methods are utilized to differentiate foreground moving objects from background, tracking methods link the motion of the objects over time to generate a trajectory, and features are extracted from the region of interest to recognize activities by using a classifier engine as shown in Fig. 2.3 [29].

The commercially available smart camera based systems come with a monitor along with features such as a text message on a personal digital assistant, an email, and alarm generation [29]. A graphical user interface allows the user to set specified criteria, boundaries, and define regions of interest, etc. The video data storage is managed in these systems by recording video in case of an alarm generation due to an abnormal activity. The basic paradigm of these systems is to hunt for relevant video from a huge video data, correlate events of multiple cameras, and correlate events to other information [3]. Thus, a smart surveillance system provides efficient location of video of required incidents, fast tracking of perpetrator with multiple cameras, and explores scouting activities of perpetrators prior to the incident. The highly advanced systems also provide geo-coded mapping tools to allow the person



(a) (b)

Figure 2.4: Marker-less model approach applications. (a) Sit to stand motion analysis between young and elderly person [4], (b) Stick figure generation for stand to sit activity and (c) Stick figure construction for sport activities [30].

in charge to pin-point the location of the activity [3, 27, 29].

2.1.2.1 Pros and Cons

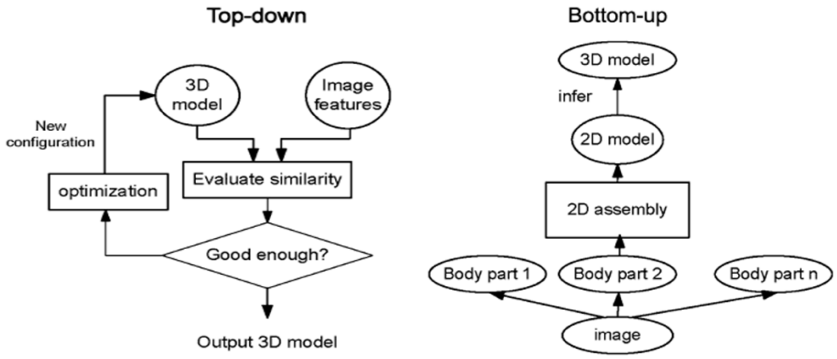
In the marker-less approach, the usage of cameras is to provide information on multiple events occurring concurrently. It alleviates the inconvenience of wearing and remembering to wear a sensor [3]. The marker-less approaches present several advantages such as cost effectiveness, use of conventional cameras, no requirement of particular attire and ease of application to numerous fields, e.g., surveillance, sports, animation, and assisted living etc. The biggest advantage in using cameras, as a means of monitoring and providing information, is the production of richer semantic information. The approaches using cameras for monitoring subjects can be easily extended for several users. The only limitation of camera based monitoring is that it requires sophisticated computer vision algorithms to track and identify the scenario occurring in a video. This makes the algorithms complex and computationally expensive. However, the current hardware advancements have made it possible to implement sophisticated computer vision algorithms that are efficient. It is harder to generate a stick figure for joint estimation and tracking, etc., using cameras as shown in Fig. 2.4 [27, 31]. The cameras used for monitoring have limited view and require good resolution to apply computer vision algorithms. Also, a single camera is not enough to keep track of persons for example in public spaces and private houses. Thus, multiple cameras are needed for complete monitoring [3, 29]. This makes the task of monitoring subjects more complex because the computer vision

algorithms need to perform inter-camera communication and coordination. Despite the complexity of the task, researchers have used camera based systems which use marker-less approach as a tool for analysing human subjects. The prime task is to provide an alert in case of an anomaly such as intruder, restricted access to an area, and in case of fall/injury [4, 29].

2.1.2.2 Classification of Marker-less approach

The marker-less approaches can broadly be classified into the model-free and model based approaches [3, 4]. The model-free approach does not require a prior model while the model based approach uses prior models. The model-free approaches use low level features on human silhouette such as contour, convex hull, edges, etc., to locate region of interest. In [20, 21, 32, 33], the local maximum of the distance curve of human contour is used to construct a star shape. The star shape yields the extremities, i.e., body parts, of the human contour. The method in [34] and [22] extends the method in [21] by creating two star and variable star, respectively. The method in [35] applies heuristic rules to the human contour in order to detect body parts. In [20], skin colour is combined with multiple contour and convex hull based cues to detect human body parts. The model-free approach is computationally inexpensive because it does not require any fitting of predefined models on the human body. However, it does not accurately locate the human body parts.

The model-based approaches use two major methods, i.e, Top-down and Bottom-up, for model-based estimation, as shown in Fig. 2.5 [36]. The Top-down method in Fig. 2.5 (a) is an analysis-by-synthesis approach that compares a pre-stored human body model with the image observation. It is prone to self occlusions, computationally bulky, and requires manual initialization. The Bottom-up method in Fig. 2.5 (b) locates and assembles individual body parts onto a human body. The manual initialization is not required but these methods are not accurate enough. An amalgamation of these two model based estimation methods is proposed by researchers for robustness. The model-based method in [32] creates 2D stick figures by using a Poisson equation solution and negative minimum curvature to locate the torso, head, hand and feet. The Poisson equation solution considers silhouette contour as a boundary and computes the random walk of all the points that are inside the silhouette till they hit the boundary [58]. In [37], pre-stored labelled body models are matched to the outline of human subjects to detect body parts. A predefined skeleton model in [38] is connected to dominant points along the con-



(a) (b)

Figure 2.5: Marker-less model based approach. (a) Top-down method, and (b) Bottom-up methods [36].

vex hull of a silhouette contour to detect human body parts. The method in [19] also uses a model-based method to detect and label human body parts by using dominant convex hull points. In [4], Gauss-Laguerre transform based method is proposed to analyse Sit to Stand motion between young and old by manually selecting shoulder, hip, knee and ankle joint as shown in Fig. 2.4 (a). A predefined model is matched to the selected joints in order to examine their trajectories. It uses monocular vision and is extendible to stereo vision marker-less configurations. In [39], a 2D torso model detects the torso and skin colour is used to detect hands. The method in [40] computes silhouette skeleton and decomposes it into segments that represents human body parts. A graph that captures the topology of these segments is created and matched with a pre-stored 3D model of human skeleton to label human body parts. In [30], the given joint locations (based on a predefined model) in the training videos are matched to a test video based on anthropometric constraints (e.g., joint locations and linkage) in order to detect and track human body parts, as shown in Fig. 2.4 (b) and (c). This research work shows the potential of human anthropometry to detect body parts in same activities observed from different viewpoints. The model-based approach has been considered for human body part detection by most of the researchers due to its accuracy. For this very reason, the present research is intended to be based on model based human body part detection and labelling [3, 36].

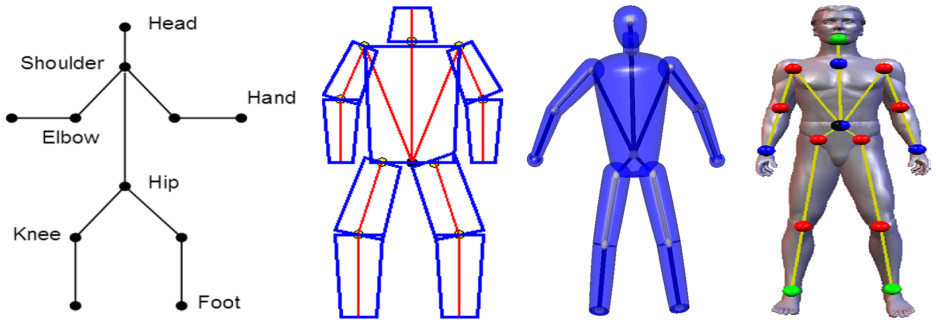


Figure 2.6: Models for tracking human body parts. (a) Stick figure model, (b) 2-D model, (c) 3D volumetric model, and (d) 3D surface model [3].

2.2 Human body part tracking

The model-based approaches use one of the following: a stick figure model, 2-D model (rectangle or contour), 3D volumetric model, and 3D surface model, as shown in Fig. 2.6 [3], to track body parts by fitting them to a 2D or 3D data of the subject. The articulated human body model such as stick figure provides rich information on human motion analysis as shown in Fig. 2.6 (a). It is an effective way of representing the physical human body structure and constraining its motion. The effectiveness of the articulated models for tracking has been shown both in 2D space and 3D space. In some methods, a 3D model is fitted onto a 2D image for 3D joint angle estimation [3, 41]. The volumetric 3D models represent the human body parts via cylinders and super-quadratics, as shown in Fig. 2.6 (c) [41]. A distance metric that minimizes the error between the observed body parts and the 2D or 3D model is used to determine the best fit. The 3D articulated model based tracking approaches are widely used because of the 3D nature of the human body. The 3D articulated models in Fig. 2.6 (c) and (d) provide richer information and are more suitable to track the human body. However, they also require specialized environment and accurate data from calibrated cameras [42]. The 3D model based tracking approach is also computationally expensive and hence not suitable for real-time applications. Thus, many researchers use 2D models instead of 3D models for tracking as shown in Fig. 2.6 (b). However, they are vulnerable to artefacts, occlusions, etc. Therefore, 2D models present a challenge to develop a robust human body part tracking method.

The human body is represented using a state vector that represents the model

(2D or 3D) parameters. This state vector is estimated by fitting and tracking the articulated body model on the human silhouette. Each state parameter represents one degree of freedom, e.g., joint angle of the human model [43]. The dimensionality of the state vector increases with the number of parameters used to define the model. A more complex model contains more parameters which in turn increase the computational complexity. Thus, several methods such as principal component analysis based dimensional reduction have been proposed to reduce the state vector by adding constraints [44]. However, these approaches limit the posture space and are not appropriate for universal motion analysis system [43].

Human body tracking is an estimation process which is performed from one frame to another by using a single or multiple hypotheses. In the following section, the Kalman Filter which is based on single hypothesis (system being modelled has one object, i.e., unimodal distribution) and the Particle Filter method which uses multiple hypotheses (multiple objects of the system can be modelled concurrently, multimodal distribution) is described.

2.2.1 Single hypothesis tracking

The single hypothesis tracking methods comprise of the Kalman filtering, and local-optimization (an iterative procedure to minimize a distance function e.g., how far a sample is from the mean of all samples). The Kalman Filter which was first introduced in 1960 has been applied to various applications [9]. It is based on three underlying assumptions: (a) the system being modelled is linear, i.e., the state parameters have unimodal distribution, (b) measurements contain white noise, and (c) noise is Gaussian in nature [9, 45]. Given a history of measurements of a system, the Kalman Filter is used to build a model for the state of the system that maximizes the a posteriori probability of those previous measurements. This means that the newly constructed model is based on the previous model with its uncertainty and the new measurements with its uncertainty has the highest probability of being accurate. In general, the Kalman Filter uses the following state description.

$$x_k = Fx_{k-1} + Bu_k + w_k \tag{2.1}$$

Here, x_k is an n-dimensional vector of state components and F is an $n-\text{by}-n$ transfer matrix, u_k is a vector of control inputs, B relates the control inputs to the state change, and w_k is random noise represented as a Gaussian distribution N with

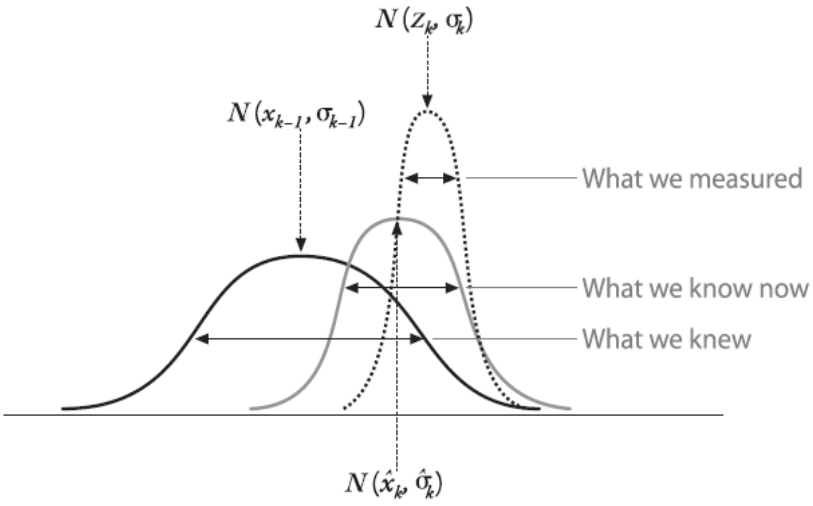


Figure 2.7: Combining prior knowledge $N(x_{k-1}, \sigma_{k-1})$ with the measurement observation $N(z_k, \sigma_k)$ to estimate the result $N(\hat{x}_k, \hat{\sigma}_k)$ [9].

zero mean. In general, the measurement (e.g., speed of a car) of the state variable x_k is computed using

$$z_k = Hx_k + v_k. \tag{2.2}$$

Here, H is a matrix of measurements and v_k is the measurement error represented as a Gaussian distribution. Finally, the Kalman gain $K = \sigma_k^2/(\sigma_k^2 + \sigma_{k+1}^2)$, with measurement error σ, is used to predict the updated value for x_k as follows.

$$x_k = x_k^- + K(z_k^- - Hx_k^-) \tag{2.3}$$

In order to achieve correct estimation $N(\hat{x}_k, \hat{\sigma}_k)$, the Kalman Filter starts with what is known $N(x_{k-1}, \sigma_{k-1})$, then obtains the new information about it $N(z_k, \sigma_k)$, and finally, decides to change what is known based on how certain it is about the old and new information by using a weighted combination of the old and the new as shown in Fig. 2.7 [9]. The first assumption, i.e., system being modelled is linear, of the Kalman Filter restricts its applicability to non-linear systems. Thus, an extended Kalman Filter was proposed to cope with this limitation of the standard Kalman Filter [9, 45]. It is a non-linear version of the standard Kalman Filter that attempts to handle non-linearities by linearising the relevant processes. The extended Kalman Filter fails when the initial estimate is incorrect or the sys-

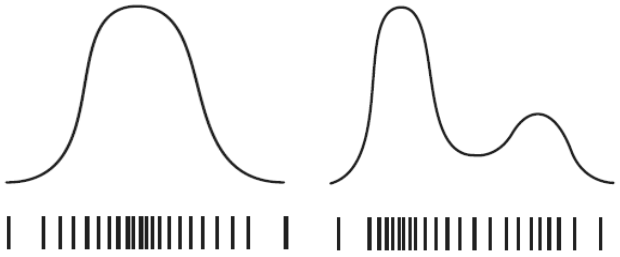


Figure 2.8: (a) The unimodal (Gaussian) distribution that can be represented by the Kalman Filter and (b) multimodal (non-Gaussian) distribution that cannot be represented by Kalman Filter but can be represented by a set of particles whose density approximates the represented distribution [9].

tem is incorrectly modelled. The Kalman Filter works well when the system being modelled has a unimodel (Gaussian) probability distribution, i.e., single hypothesis. However, in most real world applications this assumption does not hold true due to the presence of occlusions or cluttered background that yield multimodal (non-Gaussian) distribution [9, 45].

2.2.2 Multiple hypotheses tracking

The Kalman Filter cannot represent multiple hypotheses simultaneously due to the underlying assumption that the probability distribution of the system being modelled is unimodal Gaussian as shown in Fig. 2.8 [9]. Although, a set of Kalman filters can be used to propagate multiple hypotheses, they are suitable only for linear motion and, hence, are not effective for human motion which is nonlinear due to joint acceleration. Thus, a more advanced method known as the Particle Filter [9, 10] addresses these limitations of the Kalman Filter and extended Kalman Filter. The Particle Filter introduces a new parameter, i.e., the number of hypotheses (particles), that the Filter maintain at any given time. The collection of these individual hypothesis (particles) represent parametrized Gaussian probability distributions of the Kalman Filter.

The main idea in the Particle Filter is to approximate the posterior distribution $p(x_t|z_t)$ of target state at time t by a weighted sample (particle) set $S = \{(s_t^{(n)}, \pi_t^{(n)})\}_{n=1}^{N}$. Each of N particles has the state $s_t^{(n)}$ (which represent the hypothetical state of the object being tracked) and its associated weight or sampling probability $\pi_t^{(n)}$. The weights are normalized such that $\sum_n^N \pi_t^{(n)} = 1$. The posterior

density $p(x_t|z_t)$ and the observation density $p(z_t|x_t)$ are often non-Gaussian.

Algorithm 2.2.1: PARTICLE FILTER ALGORITHM(x, z, s, π)

Construct a new weighted particle set $S = \{(s_t^{(n)}, \pi_t^{(n)})\}_{n=1}^N$ for time t from the old weighted particle set $S = \{(s_{t-1}^{(n)}, \pi_{t-1}^{(n)})\}_{n=1}^N$ at time $t-1$.
Select N particles from the set $S = \{(s_{t-1}^{(n)}, \pi_{t-1}^{(n)})\}_{n=1}^N$ to give $S = \{(s_{t-1}^{'(n)}, 1/N)\}_{n=1}^N$.
Predict each particle using the dynamic model $p(x_t|x_{t-1}) = s_{t-1}^{'(n)}$ to give $\{(s_{t-1}^{'(n)}, 1/N)\}_{n=1}^N$.
Measure and weight the particles as $\pi_t^{(n)} \propto p(z_t|x_t = s_t^{'(n)})$ to give $S = \{(s_t^{(n)}, \pi_t^{(n)})\}_{n=1}^N$. Normalize $\pi_t^{(n)}$ so that $\sum_n^N \pi_t^{(n)} = 1$.
Estimate the tracking result for time t as $E[x_t] = \sum_{n=1}^N \pi_t^{(n)} s_t^{(n)}$.

Particle filtering has three operational steps: sampling (selection), prediction, and observation. In the sampling step, N particles are selected from the prior probability according to the set $S = \{(s_t^{(n)}, \pi_t^{(n)})\}_{n=1}^N$. In the prediction step the dynamic model $p(x_t|x_{t-1})$ is used to predict the state of the selected particles. In the observation step, the weights of predicted particles are recomputed using the observation model $p(z_t|x_t)$. The new state is estimated based on the newly weighted particle set. PFs can cope with non-linear dynamics and non-linear observations, by maintaining multiple hypotheses. Managing a multi-modal density allows PFs to handle clutter and recover from failures in visual tracking. The standard particle filtering algorithm is described in Algorithm. 2.2.1.

The number of particles required for robust tracking is relatively large (e.g., 50 or 100) depending on the complexity of the system being modelled [9,10]. Thus, various improvements have been proposed to enhance the standard Particle Filter to deal with increased complexity and reduce computational burden. The method in [46] uses sample importance re-sampling in which the particles are drawn from prior and assigned importance weights. Next, the particles are drawn from this importance weighted particles set. In [47–50], the standard Particle Filter is enhanced to reduce the search space (for detailed explanation see Chapter 4). In [51], the uncertainty in the state model of the Particle Filter is adapted and balanced for visual tracking. The method in [42] combines the Kalman and Particle Filter to tracking lower body parts, i.e., the leg, by using a predefined 2D articulated model.

The methods in [29, 43, 52] incorporate colour information to enhance the standard Particle Filter to achieve robust tracking. In [53], mean shift method [54] which computes local maximum is embedded with Particle Filter for tracking. A continuously adaptive mean shift method in [55] has been proposed to guide the Particle Filter for robust and efficient tracking (see Chapter 4 for further explanation).

In [56], a gravity optimised Particle Filter method was proposed which is based on Newton's law of universal gravitation. It uses the concept of gravity along with weighted particles to attract nearby particles that are close to the local maximum of the current observation. The new set of particles are replicated at the location nearer to where the particles are supposed to move. This process results in increased sampling efficiency and a reduction in the number of particles required for tracking. This method was applied to track the fingers of human hand while performing a linear motion, i.e., up and down bending of the finger. Thus, its ability to track non-linear motion and high dimensional articulated models are further research issues.

2.3 Activity recognition

This section reviews the state-of-the-art methods for human activity recognition. To this aim, the existing research work on human activity recognition is categorised into the holistic, local feature, and human model-based (prior model) or model-free (no prior model) approach [1, 3, 4, 11]. The holistic approach localises humans in videos and subsequently learns activity models that capture local and global characteristics without any notion of body parts. The local feature approach extracts descriptors from local regions in a video to learn activity models, without any knowledge about human positioning and human body parts. The human model-based approach fits a 2D or 3D model to locate human body parts and consequently extract information such as body part positioning, trajectory, etc., for activity recognition.

2.3.1 Holistic approach

The holistic approach uses shape (silhouette) and optical flow information to recognize activities. In [57], the human actions are represented by motion energy images and motion history images, as shown in Fig. 2.9 (a). The motion energy images are binary mask that signify regions of motion, and the motion history images are their corresponding weighted representations with respect to the point in time when they

(a) (b) (c)

Figure 2.9: Holistic approach. (a) Motion energy images and motion history images [57], (b) Actions as space-time shape (from left to right) for Two Hand Wave, Walk, and Run activities [58] and (c) 3D shape context descriptor (from left to right) for Bend and Skip activities [59].

(a) (b) (c) (d) (e)

Figure 2.10: Holistic approach. (a) Bounding box, (b) Scaled and aligned bounding boxes, (c) Optical flow, (d) Accumulation Regions and (e) Action descriptor.

occurred. The more recent images are given higher weight. In [58], human actions are considered as three dimensional (3D) silhouettes in the space-time volume, as shown in Fig. 2.9 (b). The space-time shapes are computed from the video scene by utilizing background subtraction. The properties of the solution to the Poisson equation are used to extract features such as local space-time saliency, action dynamics, shape structure and orientation. A similar method in [59] determines the 3D shape context, as shown in Fig. 2.9 (c), for action recognition.

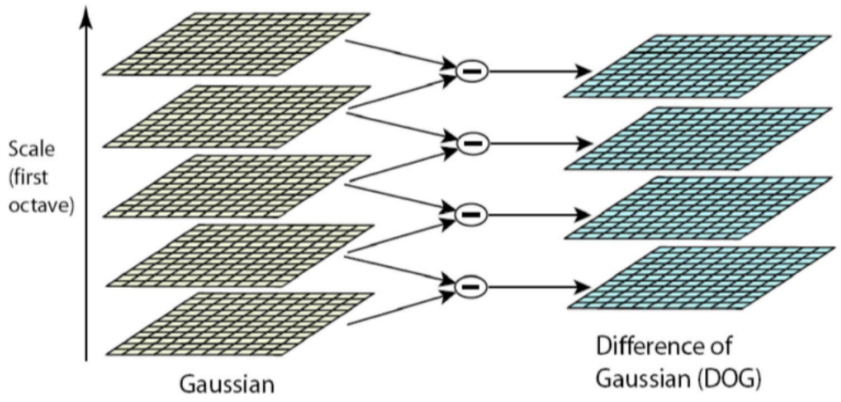
Some of the other similar shape and optical flow based methods include [12–14, 60]. In [12], an action descriptor is proposed based on aggregated local motion estimates for human action recognition as shown in Fig. 2.10. First, the bounding boxes are extracted to localise the subject performing human actions. Next, these bounding boxes are scaled and aligned, and the optical flow is estimated for every two frames. Finally, the optical flow is accumulated over a fixed number of regions to create an action descriptor. A nearest neighbour classifier is

use to recognise human actions. The method in [60] proposes a 3D motion context descriptor for human action recognition. First, motion images similar to [57, 59] are obtained from the video sequences. Next, a motion context representation is created for each human action by using the motion images. Subsequently, a 3D motion context descriptor is formed for each motion context representation. Finally, all the 3D motion context descriptors are aggregated to generate one 3D motion context descriptor to represent an action. The human actions are recognised by using probabilistic latent semantic analysis and support vector machine. In [13], a shape-motion prototype-based method is presened for action recognition. In the training phase, it extracts shape-motion descriptors to learn action prototypes which are represented via a binary hierarchical tree. In the testing phase, the shape-motion descriptor is used to recognize human actions via tree-based prototype matching and look-up table indexing. In [14], a learning-based method is proposed which uses time series of optical flow motion features for human action recognition. In the learning stage, the optical flow motion features extracted from the given action sequences are concatenated to construct motion curves. Each human action is represented by a cluster of motion curves which are clustered by using a Gaussian mixture model. In the recognition stage, the cluster of optical flow motion curves of the probe sequence is matched to the learned motion curves using a similarity function which computes the minimum distance between the motion curves. The shape and optical flow based methods are computationally expensive.

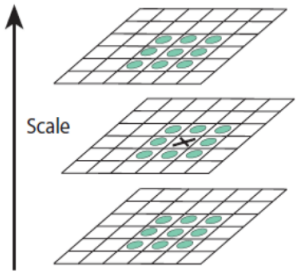
2.3.2 Local feature approach

The local feature approach uses a feature detector and feature descriptor to extract unique attributes for human activity recognition. The feature detector determines interest points such as corners, edges, etc. The feature descriptor encodes shape and motion information in a local neighbourhood around the interest points. In [61], a space-time interest point detector is proposed which detects local variations in both space and time. It has been shown to be able to detect events such as detection and pose estimation of walking people. In [62], various feature descriptors such as histogram of optical flow, histogram of 3D gradient, extended Speed-Up Robust Feature (SURF), etc., are compared.

The Scale Invariant Feature Transform (SIFT) descriptor [63] has been widely used for recognition tasks. The SIFT descriptor is based on determining the interest points (keypoints) in an image and computing a description about them using



(a)



(b)

Figure 2.11: (a) Difference-of-Gaussian is convolved with image for each scale space and (b) Maxima and minima of the difference-of-Gaussian images by comparing a sample point (pixel) in 3x3 region at a scale above and below [63].

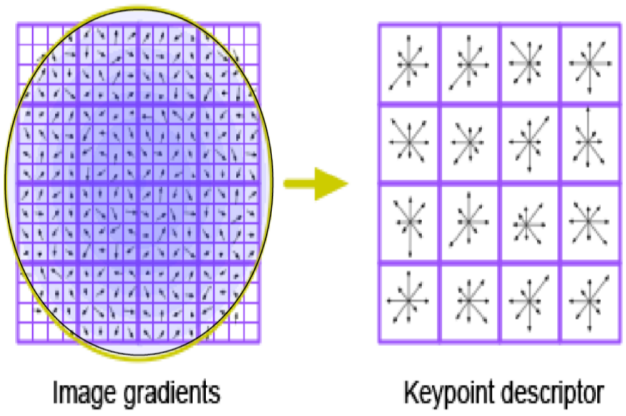
their neighbourhood pixels as shown in Fig. 2.11 and Fig. 2.12. It is computed as follows. First, the image $I(x, y)$ is convolved with variable-scale Gaussian $G(x, y, \sigma)$ to determine a scale space $L(x, y, \sigma)$ of an image.

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \tag{2.4}$$

where $*$ is the convolution operation in x and y, and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\frac{-(x^2+y^2)}{2\sigma^2}}. \tag{2.5}$$

To efficiently detect keypoint locations in scale space, the difference-of-Gaussian



(a)

Figure 2.12: SIFT descriptor computed from 16x16 neighbourhood represented by using 4x4 quadrants described as 8 orientations, i.e., 4x4x8=128, feature vector [63].

function is convolved with the image as

$$D(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y). \tag{2.6}$$

where k is a constant multiplicative factor of scale. The keypoint is determined by comparing each sample point of the difference-of-Gaussian images with its eight neighbours, i.e., in a 3x3 region, in the scale above and below, i.e., 26 neighbours. A keypoint is selected if it is larger or smaller than all the neighbours. A keypoint descriptor is created by first computing the gradient $m(x, y)$ and orientation $\theta(x, y)$,

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \tag{2.7}$$

$$\theta(x, y) = \arctan \frac{((L(x, y+1) - L(x, y-1))}{(L(x+1, y) - L(x-1, y)))}. \tag{2.8}$$

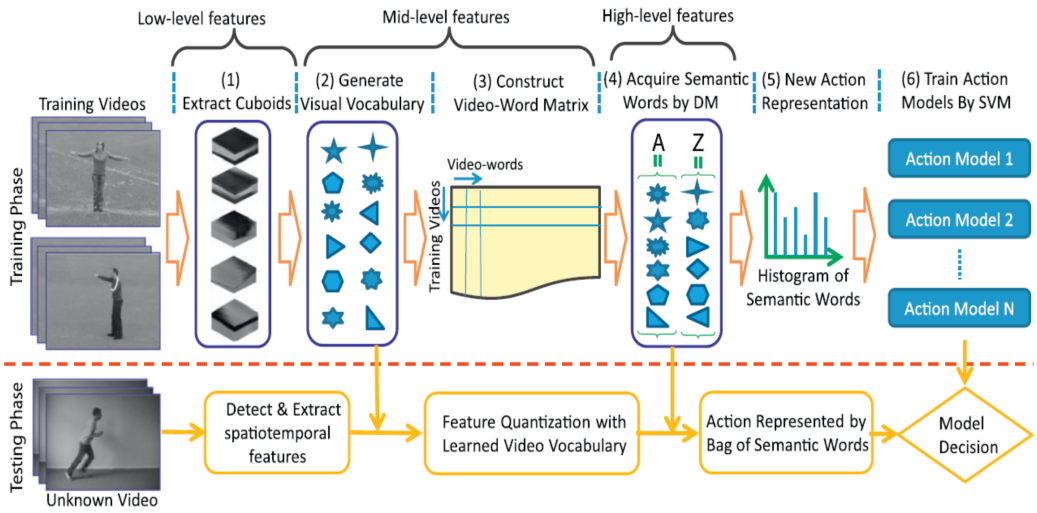
of each image sample point (pixel) in a 16x16 neighbourhood of pixels around the keypoint. The orientations in the 16x16 neighbourhood is accumulated into 4x4 quadrants where each quadrant is represented using a 8 orientation histogram. This creates the 4x4x8= 128 element feature vector, i.e., SIFT descriptor, for each keypoint as shown in Fig. 2.12.

Recently, researchers focused more on the bag of word or bag of features methods based on local features for activity recognition [15, 64–67]. This method involves the following steps: (a) feature extraction, (b) learning a visual vocabulary

(dictionary), (c) quantifying features using visual vocabulary, and (d) represent an activity by frequency of visual words. The 3D (SIFT) descriptor is proposed in [64] for action recognition. The concept is similar to applying multiple 2D SIFT descriptors [68] to several frames of a video sequence to create one 3D SIFT with its 3D sub-volumes. A bag of words method using the proposed 3D SIFT is used to represent each action. A word co-occurrence based criteria is used for human action recognition. The histograms of gradient and optical flow descriptors are presented in [65] to determine local motion and appearance. The histograms are accumulated in the space-time neighbourhood of the interest points [61] by dividing the local region into a grid of cells. A spatio-temporal bag of features representation is constructed for human action classification. In [66], two local descriptors, i.e., SIFT and cuboids, are used to represent each action by using a bag of words method. A multi-class support vector machine is used for classifying human actions. In [15], the kinematic features from the optical flow extracted from videos are converted into kinematic modes using principal component analysis. These kinematic modes are then used in a bag of kinematic mode representation for human action recognition. In [67], a novel method is proposed to learn semantic vocabulary (bag of words) for efficient and robust human action recognition as shown in Fig. 2.13. In the training phase, low-level spatio-temporal features are extracted around interest points in videos. These spatio-temporal features are clustered to obtain traditional video word vocabulary which is represented as video-word matrix, i.e., mid-level features. Next, diffusion maps (see [69] for detail) are used to create semantic words, i.e., high-level features. A new action representation is formed by computing histogram of semantic words, i.e., bag of semantic words. The training videos are used to learn action models by using support vector machine. In the testing phase, for an unknown video the same procedure is repeated to generate bag of semantic word which is compared with the learned action model for human action recognition.

2.3.3 Model-free or Model-based approach

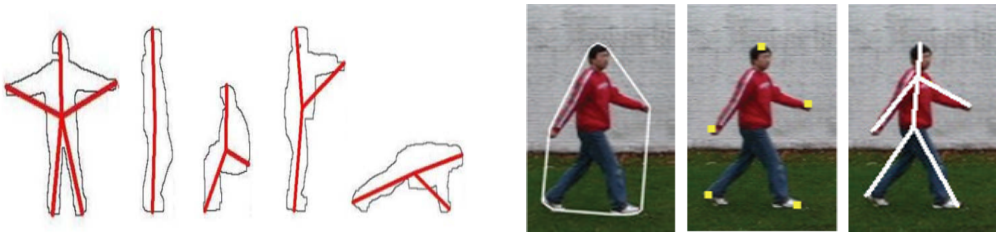
The model-free or model-based approach detects humans and then extracts shape and motion information from the silhouette contour to recognize human activities. The model-free approach in [21, 22, 34] uses star based methods to represent various human postures, as shown in Fig. 2.14 (a), and subsequently extracts features for human activity recognition. In [21], a one-star model is created to represent human posture. A human motion analysis method is proposed to extract two motion cues,



(a)

Figure 2.13: (a) Bag of semantic words for human action recognition [67].

i.e., the leg frequency and torso angle, for recognising the Walk and Run activity. This method uses the discrete Fourier transform of the filtered and autocorrelated leg frequency to discern the Walk and Run activity. The method in [34] proposes a two-star model to extract five features for detecting fence climbing action, i.e., x coordinate of centroid, y coordinate of centroid, y coordinate of centroid above fence, two or more extreme points above fence and two or less extreme points under fence. A hidden Markov model (HMM) is trained to recognise fence climbing action based on these five features. In [22], a variable-star method is proposed to robustly extract the extremities of the human contour. Subsequently, the human contour is evenly divided into twelve sectors to compute an shape context descriptor which is simply a vector indicating if there is an extremity in each sector. Finally, the feature vectors built from the detected extremities are used by the HMM for human action recognition. A similar method that combines skin colour information and various cues from human contour is proposed in [20]. In [70], convex deficiencies, i.e., the difference between the human contour and its convex hull, are proposed to represent human actions. The centroids of the convex deficiencies over time is grouped to extract five features. A human action is recognised by matching the similarity between two sets of 5D feature vectors. These methods work in real-time, however, they lack good accuracy.



(a) (b)

Figure 2.14: (a) Model-free approach that uses extremities as limb points [21,22,34], (b) Model based approach that uses a pre-defined body model to locate limbs [71].

The model based approach in [19, 33, 38] fits a body model, as shown in Fig. 2.14 (b), and then extracts features from this fitted model for activity recognition. In [33], a negative minimum curvature, i.e., points of maximally concave extremities, are used to locate the head. Next, the Poisson equation is used to determine the torso. Finally, a 8D feature descriptor extracted from the body model is utilized with the hidden Markov model for activity recognition. The method in [71] uses motion and shape features extracted from the fitted body model with the continuous hidden Markov model for event based analysis of human activities. The shape features include area and the ratio of the bounding box containing the subject.

The holistic methods that extracts shape and optical flow information are computationally expensive and require intensive training. In addition, both shape and motion information are required for accurate recognition of very similar activities. The local feature methods require intensive training, which makes them unsuitable for real world applications. Furthermore, they need large number of image frames to learn enough information to distinguish between very similar activities. In contrast, the model based or model free approach are more efficient than the holistic and local feature approaches but are less accurate for human activity recognition. The model-based approach is more accurate as compared to the model-free approach. However, it requires a fitting procedure and manual initialization which are computationally expensive. In addition, the highly accurate detection of human body parts in various activities that contain mild and severe self-occlusion continues to be a challenging issue. Furthermore, both approaches confuse very similar activities. Therefore, in this research an efficient and robust human body part detection is explored to recognise very similar human activities.

Datasets	Weizmann	MuHAVi
Key features	• Low resolution videos (180 x 144) • Background illumination variation • Full body view • 9 Subjects of different height and built • Imperfect silhouettes • 10 Routine activities	• High resolution videos (720 x 576) • Background illumination variation • Full body view • 2 Subjects of different height and built • Perfect silhouettes • 9 Routine and non-routine activities
Challenges	➢ Profile view creates **self occlusion** of limbs ➢ **Rapid movements** of limbs ➢ Very **similar activities** ➢ Used rigorously, but most methods **confuse** similar activities	➢ Activities with mild and severe **self occlusion** ➢ **Rapid change** of posture ➢ **Similar activities**

Figure 2.15: Complementary features and different challenges of Weizmann and MuHAVi datasets.

2.4 Datasets

The Weizmann and MuHAVi data sets are selected for SBP labelling and tracking because of their complementary features (e.g., low versus high resolution etc.) and the different challenges (e.g., rapid movements of limbs versus rapid change of posture et) as summarized in Fig. 2.15. In the past few years several publicly available human activity data sets have emerged that provide various challenges e.g., very similar activities, illumination variation, varying clothing, complex backgrounds, multiple actors, person-to-person interaction, human object interaction, multiple views etc. (see [72] for details on datasets). Each of the publicly available human activity data set contains one or more of the above-mentioned challenges. In addition, the human activity data sets also varies with respect to application scenario e.g., industrial setting (overhead camera generating top view), assisted living, surveillance etc. Therefore, the state-of-the-art human activity data set varies with respect to the type of challenge it presents and the application scenario. As identified in the literature review most of the human activity recognition methods confuse very similar activities. Both data sets contain easily confused activities and self occlusion of limbs, background illumination variation, varying clothing and full body view of subject. The MuHAVi data set also contains different views. These challenges make both data set suitable for human activity recognition.

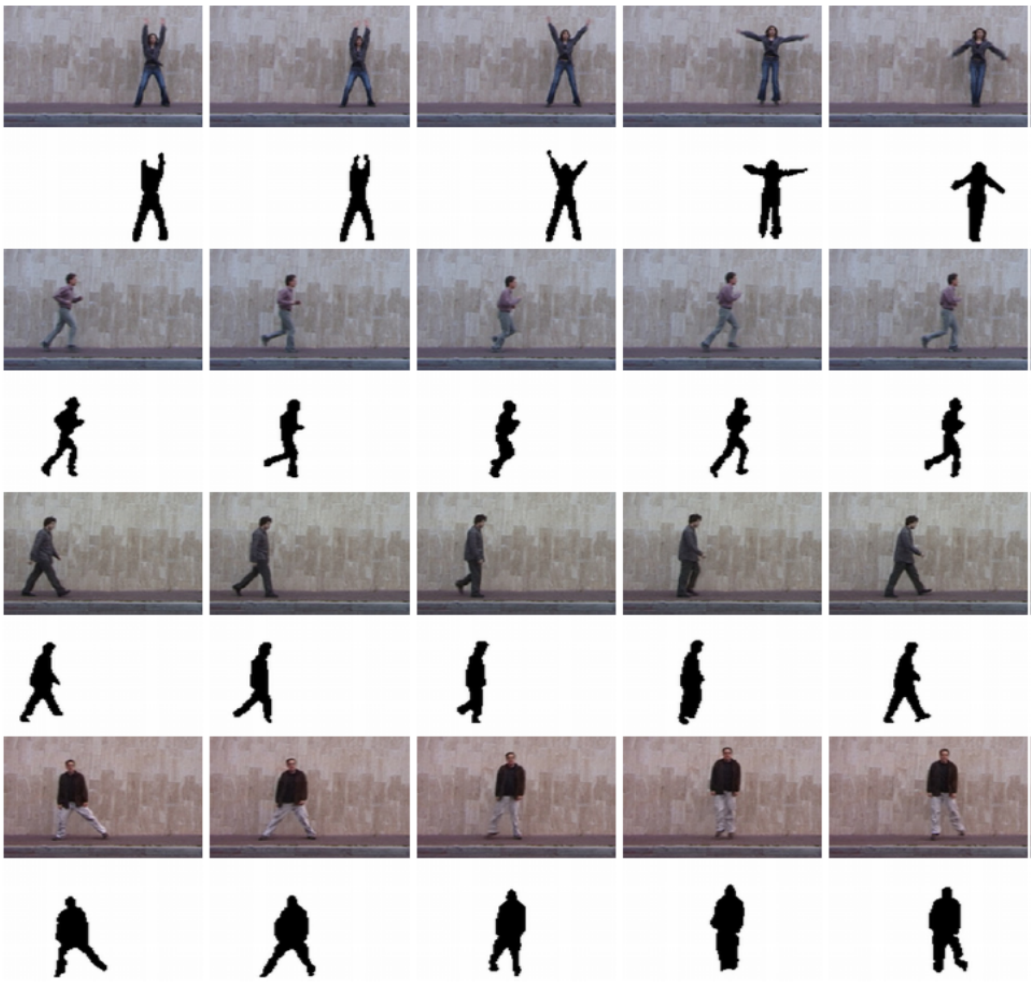


Figure 2.16: Weizmann data set. Jack, Run, Walk and Side from top to bottom row [58].

2.4.1 Weizmann data set

The Weizmann data set [58] comprises of ninety low-resolution 180×144 video sequences of various subjects performing daily activities, i.e., Walk, Run, Side, Jump, Skip, Pause Jump, Bend, Jack, Two Hand Wave and One Hand Wave. Each video sequence consist of about 80 to 120 frames. An example of video sequences and extracted silhouettes from the Weizmann data set is shown in Fig. 2.16. The silhouettes of Weizmann data set are good on average, however, they contain imperfect silhouettes in many activities.

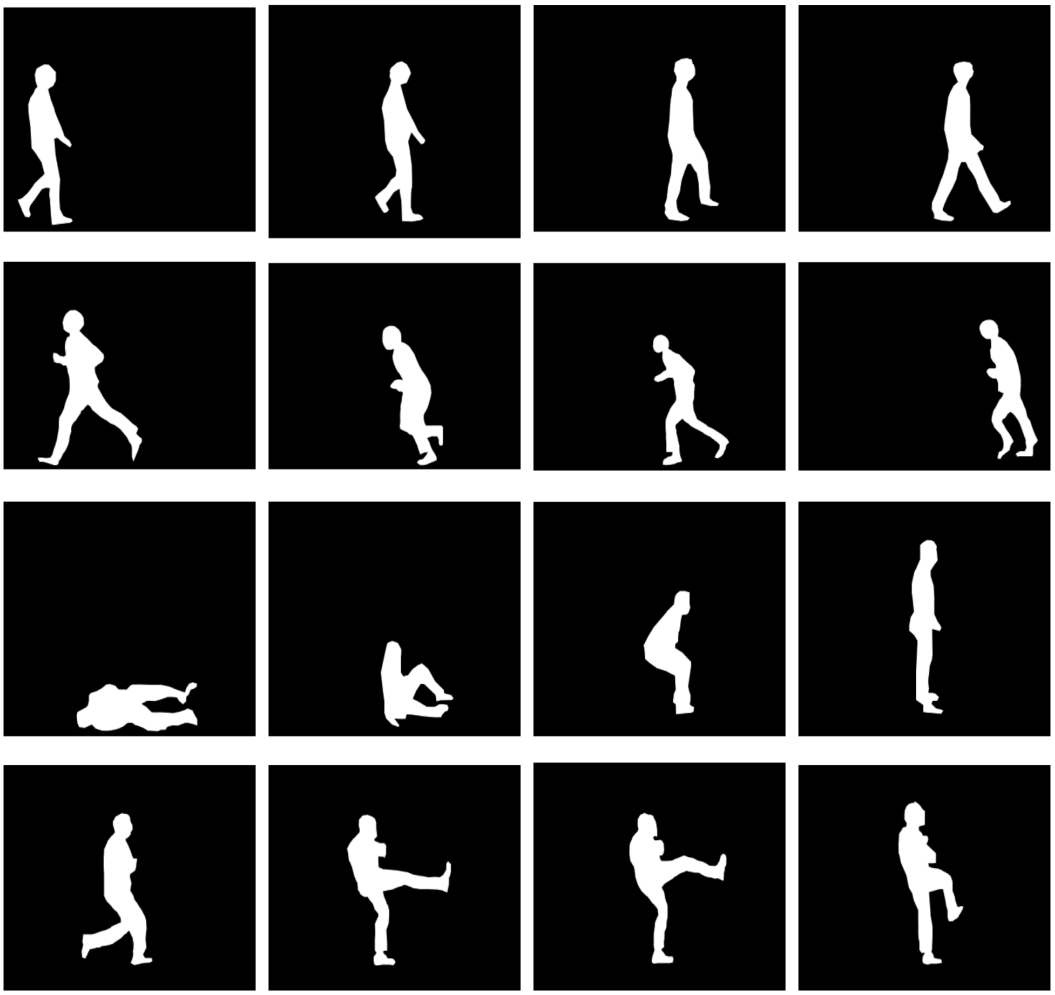


Figure 2.17: MuHAVi data set. Walk, Run, Collapse and Kick from top to bottom row [73].

2.4.2 MuHAVi data set

MuHAVi data set [73] comprises of eight high resolution 720×576 primitive activity classes, i.e., Collapse, Standup, Walk, Run, Turn, Guard-to-punch, Guard-to-kick, Punch, and Kick, of two actors with two samples with two different views (camera 3 and camera 4), i.e., a total of eight samples per activity. Each video sequence consist of 50 to 80 frames. An example of these activities is shown in Fig. 2.17. The two views, i.e., camera 3 and camera 4, is shown in Fig. 2.18

Figure 2.18: MuHAVi data set two views, i.e., camera 3 (left column) and camera 4 (right column) [73].

Chapter 3

Human Body Part Detection and Labelling

3.1 Introduction

The marker-less approach to human motion analysis uses video-based methods to detect and track positions of significant body points (SBPs) located at the convex points, i.e., the local maxima, of the silhouette contour. Applications include tracking, stick figure generation, animation for cartoons and virtual reality, imitation of human action by robots and action recognition for assisted living, surveillance, etc., [4,20]. The approach offers advantages, e.g., cost effectiveness, no requirement of particular attire and ease of application [27,31]. The marker-less approach to human motion analysis can broadly be classified into the model-based and model-free approaches. The model-based approach employs a prior model. The model-free approach estimates the motion of regions that enclose relevant anatomical landmarks without prior information about the subject's shape [4]. The former requires fitting, manual annotation, and predefined models which are time consuming while the latter tend to be less accurate.

This chapter presents a marker-less method, which uses Implicit Body Models (IBMs), that does not require a manual annotation of SBPs, training phase (learning a classifier), or fitness procedure as illustrated in Fig. 3.1. IBMs provide anthropometric, geometric, and human vision inspired constraints for labelling SBPs in activities observed from a profile view and performed by subjects of differing anthropometric proportions. The whole human body is considered as an inverted

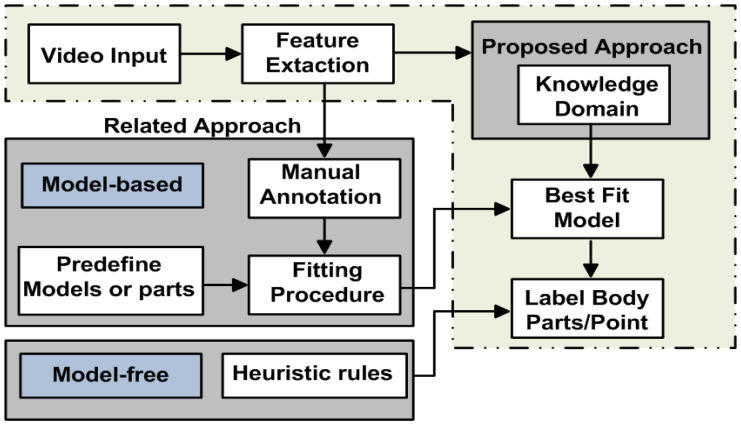


Figure 3.1: Block diagram of the proposed method versus related approaches.

pendulum model and ellipse fitting is used to compute the global angle in order to classify the Stand, Sit, and Lie postures. The contour moments are used to find the angle between the principal and vertical axis to provide cues for selecting the best IBM. The convex hull [9] of the contour is utilized to determine the locations of SBPs across time. The versatility of the proposed method is demonstrated in a number of challenging activities on the low and high resolution video data sets.

3.2 Literature review

3.2.1 Model-free approach

The body segmentation and posture estimation method in [20] is model-free and locates convex points on the contour at the local maxima of the distance curve of the silhouette contour pixels. The principal and minor axes of the human body, their relation with the silhouette contour, relative distance between convex points, and convex point curvature are used as rules to label convex points as SBPs. This method uses the head point to determine the location of feet, however, an inaccurate head point localization may lead to inaccurate feet point. It also ignores the knee point and does not present quantitative evaluation of labelled SBPs. The Star skeletonization method [21] is also model-free and recognises Walk and Run from the frequency of leg and torso angles during motion. It does not label local maxima as SBPs.

3.2.2 Model-based approach

A model-based modified Star skeleton method [32] produces stick figures from monocular video sequences and is extended in Connectivity Based Human body Modelling (CBHM) [33] by using a modified solution of the Poisson equation to obtain torso size and angle. It uses negative minimum curvature to locate the head, and nearest neighbour tracking to find the hand and feet. The local maximum method used in [20, 21, 32, 33] to identify extremities within the distance curve is sensitive to silhouette contour and these extremities are not always identified due to self occlusion. Furthermore, a smooth distance curve and self occlusion may result in missed local maxima. The method in [38] selects dominant points along the convex hull on a silhouette contour and utilises prior knowledge of body-ratio within the head, and the upper body and lower body segments to identify SBPs. The body parts are connected to a predefined skeleton model via its centre to adapt it to the subject's posture. However, the criteria for labelling convex points as SBPs are not clearly presented in [38]. This method is extended in [71] for activity analysis and 3-dimensional (3D) scene reconstruction.

First Sight [37] produces stick body parts of a subject performing complex gymnastic movements by matching a pre-stored labelled body model with an outline of a current image of the subject. The method in [74] generates an elaborate stick figure by a manual selection of anatomical landmarks, body ratios, ratio pruning, and an initial stick figure.

The W4 system [19] classifies a posture into Stand, Sit, Crawl, or Lie, then classifies the postures into front/back, and left-side, and right-side perspectives using vertical and horizontal projection histograms of its silhouette. SBPs are identified using the vertices of convex and concave hulls on the silhouette contour. A topological model is projected onto the contour to label SBPs. The quantitative accuracy of the labelled SBPs is not presented. This system is computationally expensive. In [75], the Discrete Fourier Transform (DFT) is applied to the vertical and horizontal histograms of the silhouette. A neural fuzzy network is then used to infer postures from magnitudes of significant DFT coefficients and length-width body ratio. SBPs are not labelled in [75]

In [39], a 2D model combined with the Particle Filter is used to detect the torso, and colour information is used to detect the hands. A posture is recognized by the nearest mean classifier that assigns to observations the label of the class whose mean is closest to the observation. However, initial camera calibration and

use of 500 particles to track only torso and hand limit its application in real time. The method in [35] uses heuristic rules with contour analysis to locate SBPs, and employs colour information and the Particle Filter for robust feature tracking. It has only been applied to subjects in the Stand posture. The segmentation of a silhouette contour length into portions is inadequate for activities such as Walk, Crawl, and Bend due to variations in contour lengths. The use of a Particle Filter with 1000 particles also decreases the speed of computation.

In [76], a part appearance map and an anthropometry-based spatial constraint graph cut are used to locate scope of body parts such as torso, head, arms, and legs. In [77], human body is segmented into parts, and pose is estimated using a combination of joint pixel-wise and part-wise formulation. Each pixel is assigned to an articulated model using a histogram of gradients. This model is segmented into body parts using a given set of joint positions. However the locations of body parts are not evaluated in these methods.

The pose estimation framework in [78] uses a two layered random forest classifier to localise joints. The first layer classifies the body parts, and the second incorporates the body parts and their joint locations to estimate the pose. In [79] articulated body parts are detected by first finding the torso and then performing a fitness procedure to locate the remaining body parts. It is computationally expensive with no occlusion handling ability.

The recent introduction of the low-cost depth camera has motivated researchers to utilise depth images. In [80], the 3D pose is estimated from a single depth image. The human body is divided into a set of parts and a random forest is employed to compute the probability of each pixel belonging to each part. The 3D joint locations are then independently estimated from these probabilities. A similar method in [81] is applied to video images from multiple views. Random forest is used to assign every pixel a probability of being either a body part or background. The results are then back-projected to a 3D volume. Corresponding mirror symmetric body parts across views are then found by using a latent variable, and a part-based model is used to find the 3D pose. In [82], a local shape context descriptor is computed from edges obtained from depth images to create a template descriptor of each body part category, i.e., head, hand, and foot. A multivariate Gaussian model is employed on the template descriptor to compute the probability of each category. A greedy algorithm then finds the best match to identify the body parts. The use of multi-view and depth images are not within the scope of this thesis.

3.3 Foundation of proposed framework

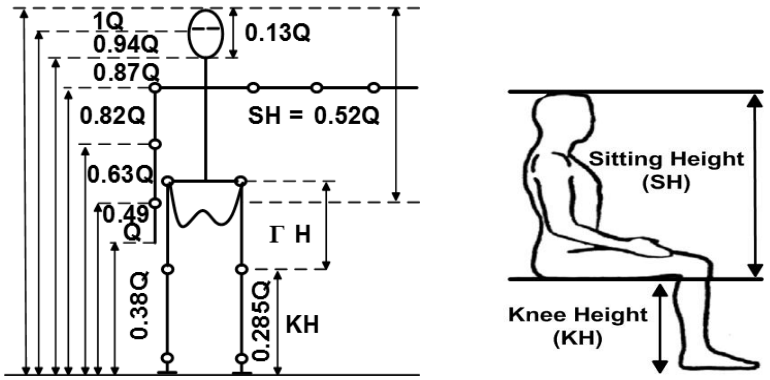
The human body has no fewer than 244 degrees of freedom [83] and can attain a variety of postures due to its high dimensionality. Anthropology reveals that body dynamics are affected by age, ethnicity, class, family custom, sex, talent, circumstance, and preference [84–86]. However, empirical studies have revealed that these variations are not arbitrary [86, 87]. Moreover, human actions are also influenced by psychology, society, and culture. Thus, the sheer range and complexity of human actions make developing automated SBPs labelling algorithm a daunting task.

Human body proportion has been widely studied with applications in engineering, ergonomics, and computer vision [86]. By using the 5^{th}-95^{th} percentile values of body proportion, 90 percent of the world population can be covered [5, 6]. Anthropometry has only been used to detect significant body points in the Stand posture in a semi-automated manner, since its application in complex actions is not an easy task [7, 8]. Anthropometric transformations do not conform to any known laws, it is thus not possible to formally define invariant properties. A functional definition of anthropometric transforms is presented combining anthropometric, geometric, kinesiology, and human vision (heuristic) inspired constraints, to provide six IBMs for robust labelling and tracking of SBPs. The six IBMs cover most actions, activities, and range of motion performed by human from a profile view (see Section 3.5).

In this chapter, SBPs are labelled as Head (H), Shoulder (S), Arm (A), Knee (K), Feet (F). Each SBP abbreviation can be considered as a vector which has a 2D position, i.e., $H = (x^H, y^H)$, $A = (x^A, y^A)$ and $F = (x^F, y^F)$. Here, the superscripts represent the abbreviations of SBPs. The current and previous position of a SBP is denoted as $H(t) = (x^H_t, y^H_t)$ and $H(t-1) = (x^H_{t-1}, y^H_{t-1})$ respectively. Subscript refers to a specific entity, e.g., x_c, x_{cv} and x_{nr} represent the x coordinate of a centre, convex point, and normalised convex point, respectively.

3.3.1 Implicit Body Models (IBMs)

Several anthropometric studies reveal that in the Stand posture the head length is approximately one-eighth the total length of the human body [85, 88, 89]. The body segment length as a fraction of human body height (1Q) is shown in Fig. 3.2 (a), where $8 \times 0.13Q \approx 1Q$ [89]. These ratios are used to provide ranges of eight segments



(a) (b)

Figure 3.2: (a) Body segment lengths as a fraction of the body height (1Q); (b) Sitting height measured form head to seated buttocks [88].

to label SBPs in the Stand posture. The human body maintains an approximate Stand posture in activities such as Walk, Run, Skip, etc. However, these activities induce motion in the vertical plane of the human body which is compensated for by selecting a longer range from the eight segments providing accurate labelling and tracking of SBPs. Thus, the Stand body model is divided into seven segments (G1-G7) as shown in Fig. 3.3 (a) (see Section 3.4.1.4).

Anthropometric studies show that in the Sit posture the thigh becomes horizontal to the ground and human body height decreases (i.e., head length is not one-eighth the total length of human body) [6, 88] as shown in Fig. 3.2 (b). As a result, the Sit posture cannot be divided into eight segments based on empirical anthropometric studies. Note that the body part positioning, (i.e., Head, Shoulder, Arms, Knee, and Feet above each other, respectively) is somewhat maintained in the Sit posture [88]. This problem is resolved by finding the relationship between the segmentation of the Sit and Stand posture based on anthropometric studies [6,88,89]. According to Fig. 3.2 (a) and (b)

$$\Gamma H = 1Q - SH - KH = 1Q - 0.52Q - 0.285Q = 0.195Q \tag{3.1}$$

where ΓH and KH are respectively the thigh length and knee height in the Stand posture. SH is the sitting height (i.e., measured from head to buttocks) in the Sit posture [88].

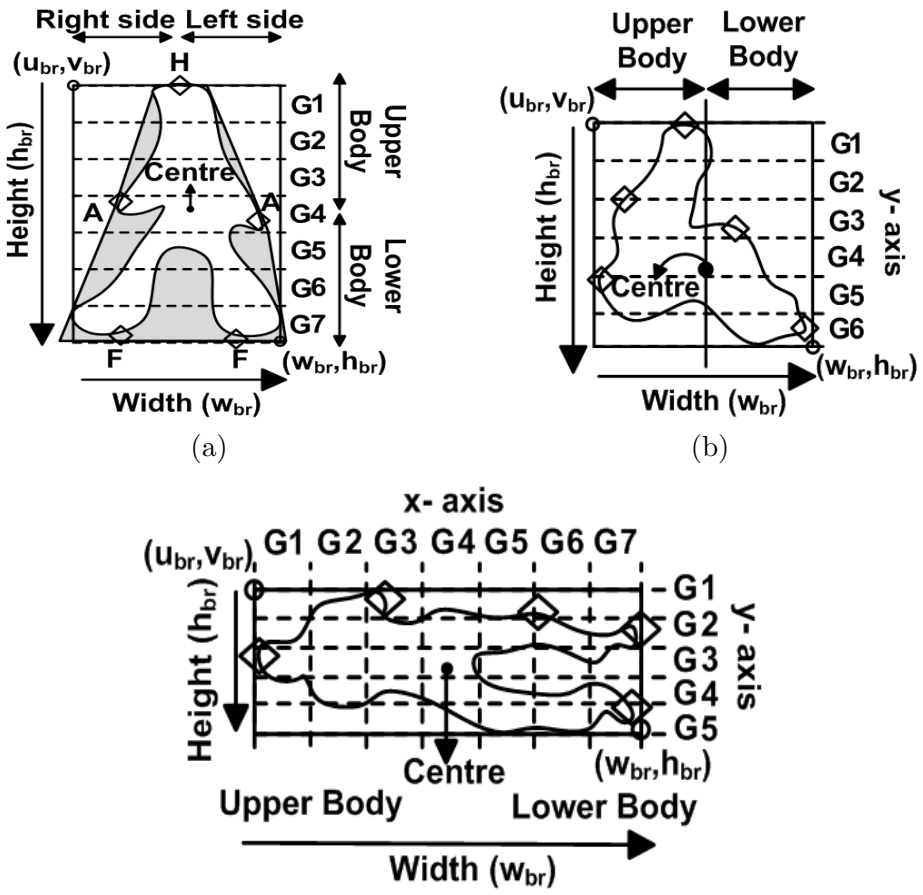


Figure 3.3: IBMs for Head (H), Arm (A), and Feet (F) SBP labelling and anthropometry based segmentation [G1-G7] (see Section 3.4.1.4 Table 3.3) of silhouette contour using bounding rectangle minimum (u_{br}, v_{br}) and maximum points (w_{br}, h_{br}) for: (a) Stand (α activities in Table 3.1, convex hull in shaded region); (b) Sit; and (c) Lie.

The number of segments is

$$N_{seg} = \frac{8(1Q - \Gamma H)}{Q} = \frac{8(1Q - 0.195Q)}{Q} \approx 6. \tag{3.2}$$

By substituting (Eq. 3.1) in (Eq. 3.2), for the Sit posture N_{seg} should be six, hence, the Sit body model is divided into six horizontal segments (G1-G6) as shown in Fig. 3.3(b). The Lie body model is considered as the Stand body model rotated by 90° based on geometry, thus it is divided into seven vertical segments (G1-G7). The lie body model is further divided into five horizontal segments (G1-G5) to account for head leaning [90, 91] in the sagittal plane as shown in Fig. 3.3(c). These three IBMs can be used to label SBPs in cyclic activities (e.g., Walk, Side, and Skip), and in the Stand, Sit and Lie postures. In all of these activities, anthropometric body proportions and part positioning are somewhat maintained. However, in activities such as Bend, Wave, Punch, and Kick, the anthropometry based positioning of body parts/points is not maintained, i.e., the hand goes above/near the head (in Wave, Punch) or below the knee (in Bend), and the feet go above the knee and centre of contour (in Kick) [5, 90–92].

The IBMs are defined based on a range of motion obtained from anthropometric [5,91,92] and kinesiology studies [90], human geometry and vision constraints. They are used to label and track SBPs in activities that do not exactly maintain anthropometry (see Section 3.4.1.4 and Section 3.4.3.4 for details). The Wave IBM in Fig. 3.4(a) covers a range of motion of shoulder, arm, and elbow. The Kick IBM in Fig. 3.4(b) covers a range of motion of knee and leg. The Sit body model slightly overlaps with the bend posture. Finally, the Bend IBM in Fig. 3.4(c) covers a range of motion of trunk. These models cover a diverse range of motions of the shoulder, hand, arm, elbow, knee and hip mentioned in kinesiology studies and as shown in Fig. 3.5 [90].

3.3.2 Inverse pendulum and contour moments

Humans are bipeds and locomote over the ground with the majority of the body mass located two third of the body height above the ground. Due to this reason the whole human body can be represented as an inverted pendulum which is capable of moving in anterior-posterior (forward-back movement) and medial-lateral (side-to-side movement) directions as shown in Fig. 3.6 (a) and Fig. 3.6 (b) [93–99]. In a simple pendulum, it is assumed that motion happens only in two dimensions, i.e.,

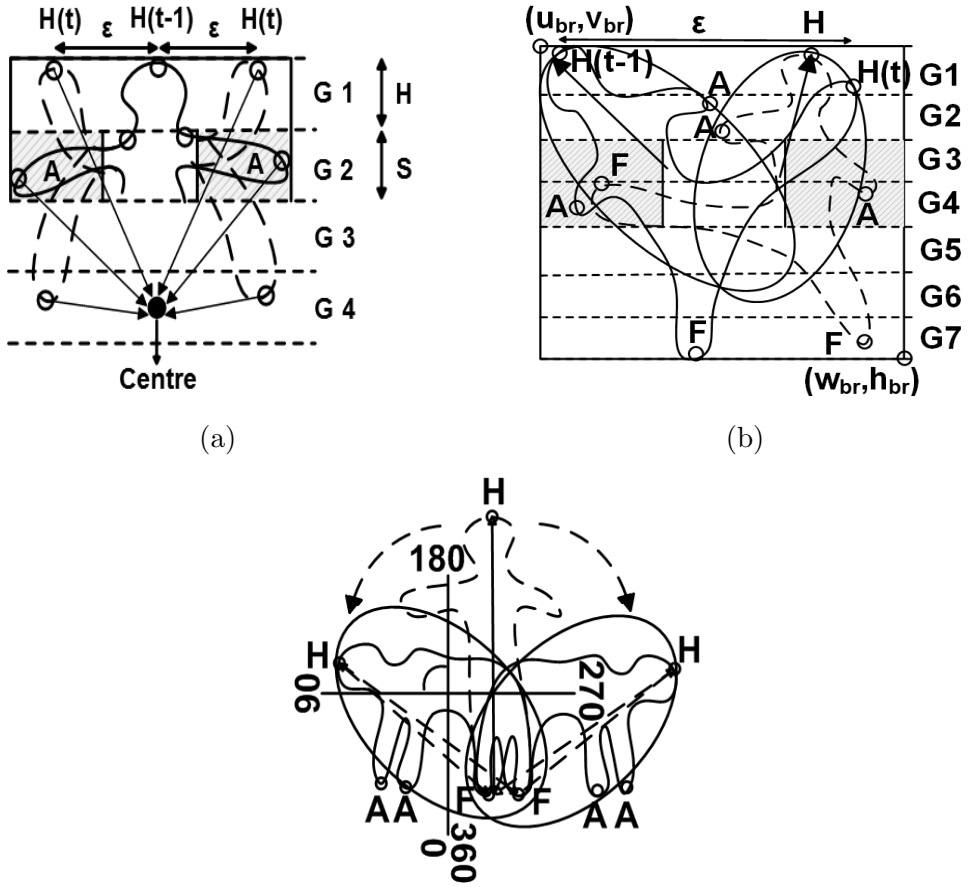


Figure 3.4: IBMs based on cues in Section 3.4.1.4 with Smart Search Algorithm (see Section 3.4.3.4) for locating and labelling Head (H), Arm (A), and Feet (F) SBPs in β activities (see Table 3.1): (a) Wave; (b) Kick and (c) Bend.

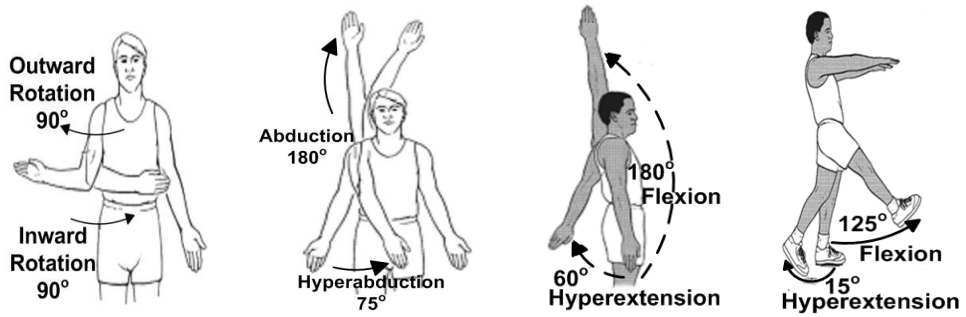
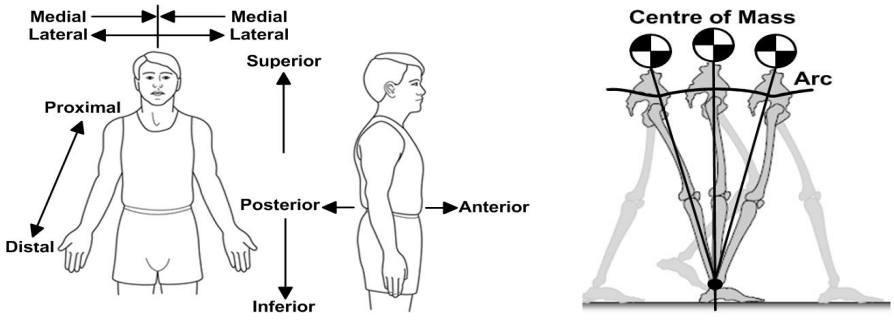


Figure 3.5: Front and Side view: (a) Elbow range of motion, (b)-(c) Arm range of motion and (d) Leg range of motion based on anthropometric and kinesiology studies [90–92].



(a) (b)

Figure 3.6: (a) Body planes and orientation based on anatomy [91, 92] and (b) Human body inverse pendulum model draws an arc in Walk motion [93].

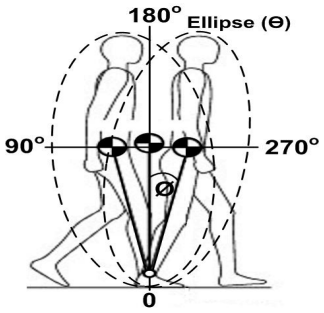


Figure 3.7: The global angle θ and angle ϕ from the vertical axis of the inverse pendulum human body model.

the point of mass does not draw an ellipse but an arc. This conjecture allows us to apply an inertia ellipse (referred in this thesis as 2D ellipse fitting procedure) on the inverted pendulum human body model as shown in Fig. 3.7.

The global angle θ and angle ϕ of the human body from the vertical, respectively, are computed using ellipse fitting and contour moments. The contour moment of an image $f(x, y)$ is defined as [100, 101]

$$m_{pq} = \sum_{-\infty}^{\infty} \sum_{-\infty}^{\infty} x^p y^q f(x, y) dx dy \tag{3.3}$$

where are respectively the x-order and y-order (whereby order means the power to which the corresponding component is taken in the integral) moment of the contour, and x and y are coordinates. The centre of the ellipse enclosing the human body is

an approximation of the centre (x_c, y_c) the human contour mass, i.e.,

$$x_c = \frac{m_{10}}{m_{00}}, y_c = \frac{m_{01}}{m_{00}} \tag{3.4}$$

where m_{10}, m_{01}, and m_{00} are respectively the first and zero order spatial moments. The centre (x_c,y_c) is used to calculate the central moment

$$\hat{\sigma}_{pq} = \sum_{-\infty}^{\infty}\sum_{-\infty}^{\infty}(x - x_c)^p(y - y_c)^q f(x, y)dxdy. \tag{3.5}$$

The global angle of the human body is the angle of the axis with the least moment of inertia in degree or radian as shown in Fig. 3.7, i.e.,

$$\theta = \frac{1}{2}\tan^{-1}\frac{2\hat{\sigma}_{1,1}}{\hat{\sigma}_{2,0} - \hat{\sigma}_{0,2}} \tag{3.6}$$

where $\hat{\sigma}_{1,1}$ is the first order central moment, and $\hat{\sigma}_{2,0}$ and $\hat{\sigma}_{0,2}$ are the second order central moments [100, 101]. The angle of the human body from the vertical using contour moments is computed as $\phi = |90 - \theta(180/3.14)|$. Both the global angle and the angle of human body from the vertical vary over time t, i.e., $\theta(t)$ and $\phi(t)$.

3.4 The proposed framework

A split approach is developed to find the best IBM for labelling the convex points on a silhouette contour as SBPs. Fig. 3.8 shows the main components and work flow of the proposed framework. A hierarchical categorization of human posture (Stand, Sit, Lie), movements (Right to left, Left to Right, Stand to Lie, Lie to Stand) and the human body itself (Upper body and lower body, Right side and left side) is done. Stand, Sit, and Lie postures are categorized by considering the human as an inverse pendulum and using contour moments. In the Stand, Sit and Lie postures, Upper body and Lower body, and Right side and Left side are respectively distinguished based on the transverse and sagittal planes as shown in Fig. 3.3.

Initially the Stand to Lie or Lie to Stand movement is ascertained (see Section 3.4.1). The human posture is categorised in Stand to Lie and Lie to Stand movements by using the global angle. Right to Left, Left to Right, and no movement are discerned based on the subject's location in the first frame. In Stand to Lie, for Stand, the movement is further divided into α and β (see Table 3.1).

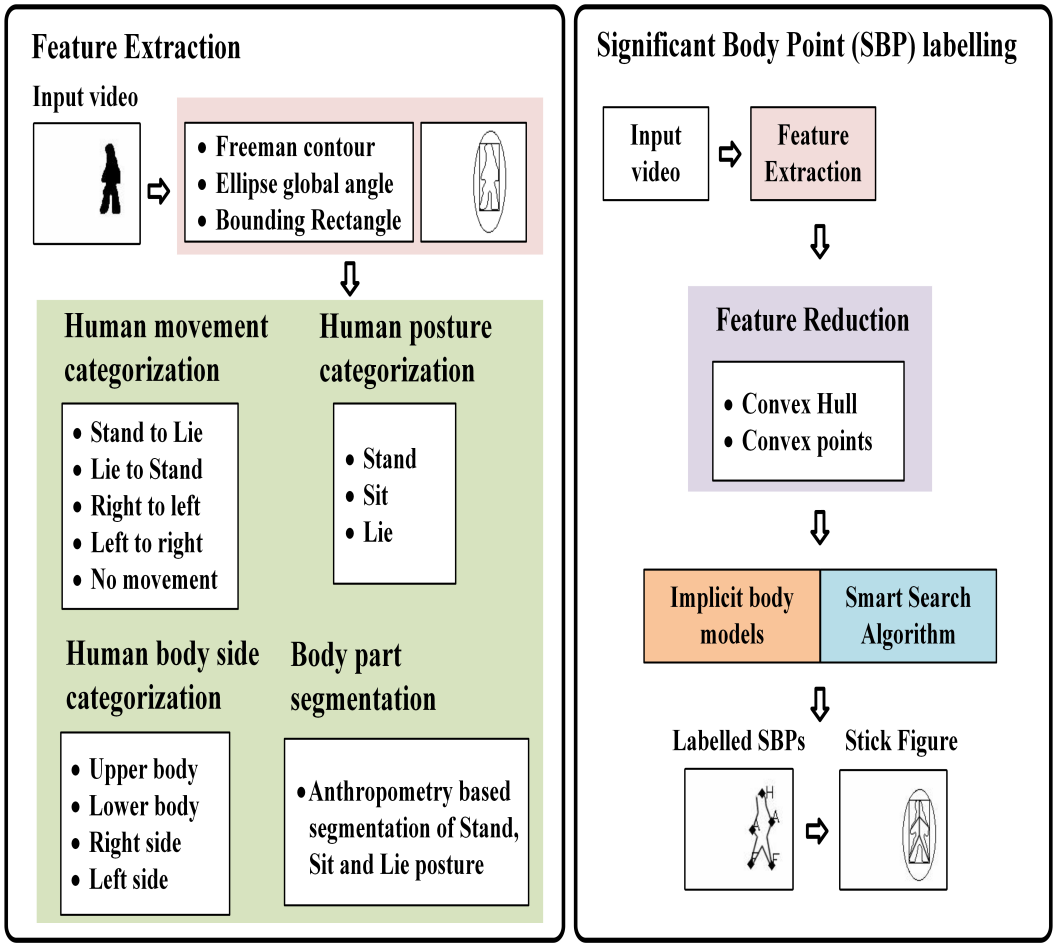


Figure 3.8: The components and work flow of the proposed framework for Significant Body Point (SBP) labelling.

Table 3.1: Acronyms for activities.

Type	Activities (α)
1	Walk
2	Run
3	Skip
4	Side
5	Jump
6	Turn

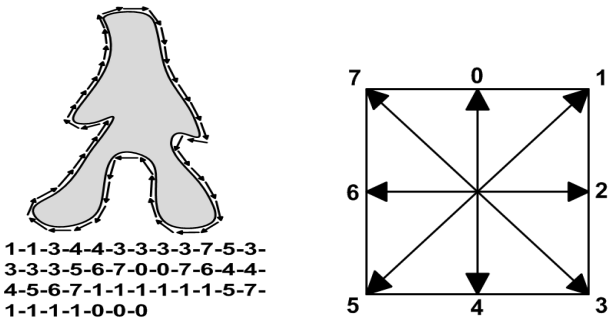
Type	Activities (β)
7	Jump-in-place-on-Two-Legs/Pause Jump
8	Bend
9	One Hand Wave
10	Two Hand Wave
11	Jack
12	Standup
13	Collapse
14	Kick
15	Punch
16	Guard-to-Kick
17	Guard-to-Punch

α refers to activities with Right to Left or Left to Right movement, e.g., Walk, Run, Skip, Side, Jump, Turn. β refers to activities in which the subject remains almost at the same place and has Right side or Left side motion, e.g., Jump-in-place-on-two-legs, Bend, One Hand Wave, Two Hand Wave, Jack, Standup, Collapse, Kick, Punch, Guard-to-Kick, Guard-to-Punch.

The global angle and the bounding rectangle are respectively used in α and β to select the best IBM for labelling anatomical landmarks. β is further categorized into $\dot{\beta}$ and $\ddot{\beta}$ (see Section 3.4.1.4) to select the appropriate IBM. For any action, the convex points of a human contour are normalized with respect to the bounding rectangle and then filtered. The criteria summarized in Section 3.4.3 from the proposed IBMs are used to label these convex points as SBPs in Stand to Lie, Lie to Stand, α, and β movements. Particle Filter (or Motion flow) is used for prediction during occlusion. Finally, the SBPs are connected to generate stick figures for various actions and activities.

3.4.1 Silhouette feature extraction

As in [102] a contour is traced using the Freeman chain code (using 8-way connectivity) [103] as shown in Fig. 3.9 on the silhouettes of the Weizmann [58] and Multi-camera Human Action Video (MuHAVi) data sets [73] (see Section 3.5). A



(a) (b)

Figure 3.9: (a) Freeman Chain Code contour (b) Chain direction.

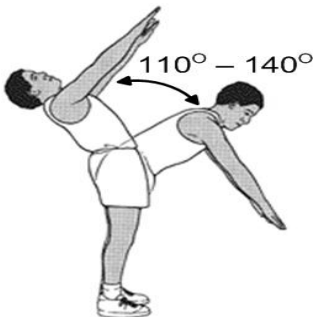


Figure 3.10: Trunk extension and flexion range based on biomechanical basis [92]) of human movement.

least-squares fitness procedure is used to compute the ellipse global angle $\theta(t)$ based on (Eq. 3.6) that best approximates the contour.

3.4.1.1 Human movement categorization

The maximum flexion and extension range of the trunk in the Stand posture, i.e., 140°, as shown in Fig. 3.10 [92], is used to set the initial global angle θ_{start} parameters such that $255 - 115 = 140°$. This initial global angle is only checked in the first frame of the input video sequence. It is a metric to ascertain the preliminary state of the subject's posture by determining whether the body movement starts from Stand, i.e., Stand to Lie, or from Lie, i.e., Stand to Lie, according to

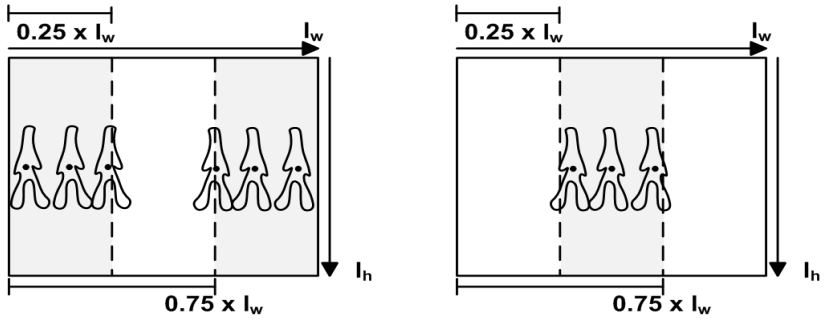
$$\gamma 3 = \begin{cases} \text{Stand} & \text{if } 115 \leq \theta_{start} \leq 255 \end{cases} \tag{3.7}$$

$$\gamma 4 = \begin{cases} \text{Lie} & \text{if } 115 \not\leq \theta_{start} \not\leq 255 \end{cases} \tag{3.8}$$

Table 3.2: Acronyms for body movement and body side.

Type	Body movement (γ)
1	Right to Left
2	Left to Right
3	Stand to Lie
4	Lie to Stand

Type	Body side (δ)
1	Upper body
2	Lower body
3	Right side
4	Left side



(a) (b)

Figure 3.11: (a) α significant movement from right to left, and left to right; (b) β no significant movement .

where body movements $\gamma 3$ and $\gamma 4$ are described in Table 3.2.

α and β are respectively determined as shown in Fig. 3.11 using

$$\alpha = \left\{ \ \gamma 1 | 0.25 I_w > x_c \text{ or } \gamma 2 | x_c > 0.75 I_w \right. \tag{3.9}$$

$$\beta = \left\{ \ 0.25 I_w < x_c < 0.75 I_w. \right. \tag{3.10}$$

where body movements $\gamma 1$, and $\gamma 2$ are described in Table 3.2. I_w and I_h are the frame width and frame height, respectively.

3.4.1.2 Human posture categorization

Standard deviation of the global angle has been used to discriminate human shapes, posture based events, and activities [104]. In [20], the difference in angle between the principal and vertical axes is used to detect SBPs but not for posture classification.

Stand, Sit, and Lie postures are categorized by considering human as an inverse pendulum and using contour moments. Biomechanical analysis of human

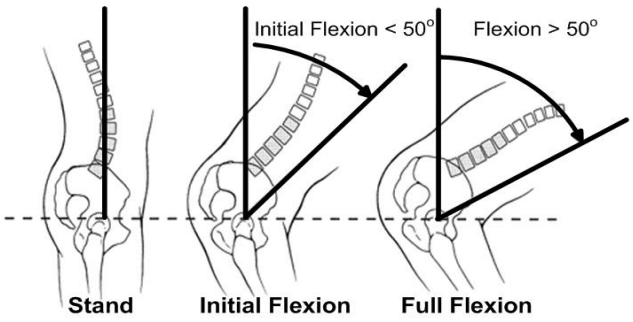


Figure 3.12: Biomechanical analysis of trunk flexion due to rotation of lumbar vertebrae and pelvic [92].

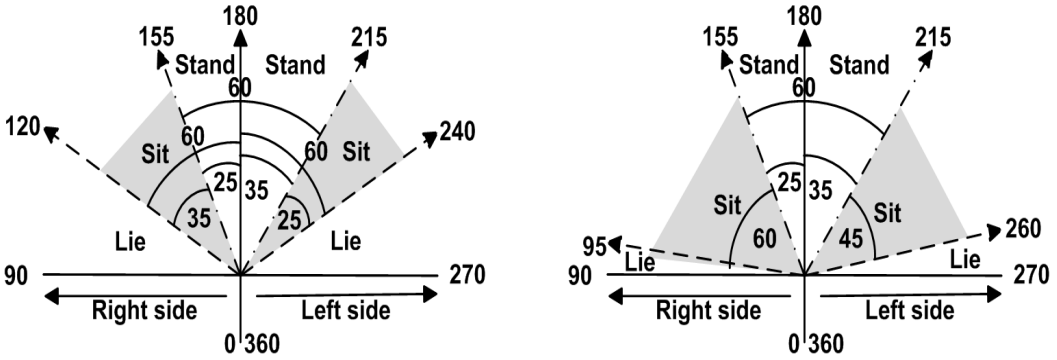


Figure 3.13: Stand, Sit, and Lie posture classification using ellipse global angle $\theta(t)$ (see Section 3.4.1.2) in movements from: (a) Stand to Lie and (b) Lie to Stand.

spine show that a complete flexion of the whole trunk occurs due to a rotation of the lumber vertebrae and pelvis, when the difference between the vertical and axis of human body rotation is greater than 50° [92] as shown in Fig. 3.12 [92]. A 60° variation in global angle is set to differentiate between the Stand and Lie posture for Stand to Lie.

The reference global angle for Stand is set to 180° in Fig. 3.13. A flexion of more than 60° from the reference in clockwise or anti-clockwise direction is considered as the Lie posture, i.e., Lie $= 180 \pm 60 = 120°$ or 240°. The human body can flex and extend at a range of $110 - 140°$ [92] while maintaining a somewhat Stand posture as shown in Fig. 3.10. This yields a variation of 40-70° from the reference global angle with an average of 55°. Thus, the range of angle for the Stand posture is set to be $215 - 155 = 60°$, i.e., Stand $= 180 + 35 = 215°$ or $180 - 25 = 155°$ as

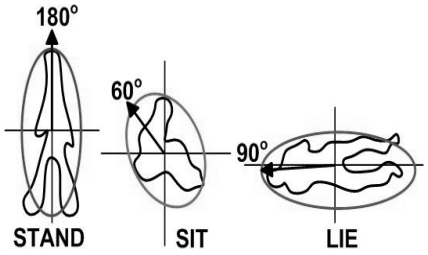


Figure 3.14: Stand, Sit, and Lie posture orientation and categorization concept.

shown in Fig. 3.13 (a). The disproportionate division of this range is to cater for the clockwise and anti-clockwise directions leaning ability of the human body while in the Stand posture. Sit posture is categorised in the remaining range of angle for clockwise and anti-clockwise directions. It also encompasses intermediate posture such as Bend, manoeuvre from Sit to Lie, and *vice versa*.

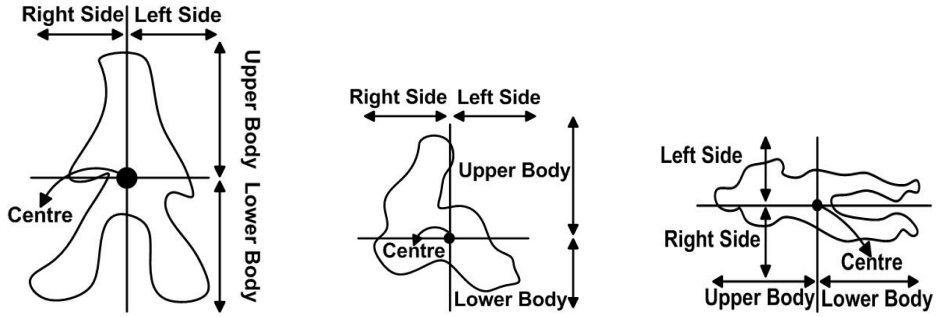
The range of global angle for Stand in Lie to Stand Fig. 3.13 (b) is kept the same as Stand to Lie, i.e., $215 - 155 = 60°$. However, in trying to stand from Lie, the body leans forward and the subject remains in intermediate posture (Sit) for a longer duration. Thus, a global range of 60° is set for the Sit posture in Lie to Stand, i.e., $155 - 95 = 60°$. The Lie posture is categorized in the remaining range of global angle for clockwise and anti-clockwise directions. Fig. 3.13 illustrates the resulting division of ellipse quadrant used to categorise postures for Stand to Lie and Lie to Stand. A mirror reflection of Fig. 3.13 is used for the opposite direction of Right side and Left side for Stand to Lie and Lie to Stand. Fig. 3.14 shows the Stand, Sit, and Lie posture orientation and categorization concept. Thus, the IBM for α activities is selected based on these ranges of global angle.

3.4.1.3 Human body side categorization

The human body side is categorized into Upper body and Lower body, and Right side and Left side based on centre location as shown in Fig. 3.15 using

$$\begin{aligned} &\text{Stand, Sit}|\delta1 < y_c \;\&\; \delta2 > y_c \;\&\; \delta3 < x_c \;\&\; \delta4 > x_c \\ &\text{Lie}|\delta1 < x_c \;\&\; \delta2 > x_c \;\&\; \delta3 > C_y \;\&\; \delta4 < y_c \end{aligned} \tag{3.11}$$

where body sides $\delta1$, $\delta2$, $\delta3$ and $\delta4$ *are* described in Table 3.2.



(a) (b) (c)

Figure 3.15: Human body side categorization (a) Stand, (b) Sit, and (c) Lie.

3.4.1.4 Body part segmentation

The ellipse fitting procedure used in [20] provides approximations, i.e., not all the body contour points are enclosed by the ellipse as illustrated in Fig. 3.7. The bounding rectangle is used to enclose contour, and obtain its minimum and maximum points, i.e., $P_{min} = (u_{br}, v_{br})$ and $P_{max} = (w_{br}, h_{br})$. u_{br} and v_{br} are respectively the starting x and y coordinates of the bounding rectangle. w_{br} and h_{br} are respectively the width and height of the bounding rectangle. These points represent the size of the silhouette contour, and are used to divide the body into segments [G1-G7] using anthropometric information [85] (see Section 3.4.3) defined for IBMs in each of the Stand, Sit and Lie postures as illustrated in Fig. 3.3. The difference between two segments (which depends on the number of segments N_{seg}) is

$$D_{seg} = (P_{max} - P_{min})/N_{seg} \tag{3.12}$$

where N_{seg}=7,6,5 and D_{seg}=30,21,22 pixel for horizontal segmentation of Stand, Sit and Lie, respectively, and N_{seg}=7 and D_{seg}=30 pixel for vertical segmentation of Lie. h_{br} and v_{br}, and w_{br} and u_{br} are used in (Eq. 3.12) for horizontal and vertical segmentation, respectively. The normalised segments $G[g]$ are determined using

$$G[g+1] = D_{seg} \times (g+1)/(P_{max} - P_{min}), \forall g \in 0 : N_{seg} \tag{3.13}$$

where $g = 0$ and $g = N_{seg}$ respectively correspond to the minimum and maximum points of the bounding rectangle as shown in Fig. 3.3. Table 3.3 shows the normalised segmentation values for the Stand, Sit, and Lie posture fixed for all the experiments.

The bounding rectangle along with the angle $\phi(t)$ from the vertical and global

Table 3.3: Normalised segment values for Stand, Sit and Lie IBM.

Model	G1	G2	G3	G4	G5	G6	G7
Stand	0.147	0.295	0.443	0.591	0.738	0.886	1
Sit	0.164	0.328	0.492	0.656	0.742	1	-
Lie	0.194	0.388	0.582	0.776	1	-	-

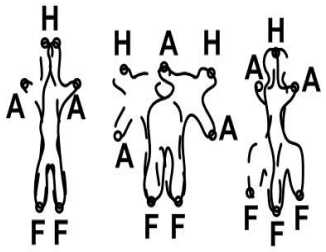


Figure 3.16: The intermediate human body postures.

angle $\theta(t)$ are used to provide cues towards selecting the best IBM for β movements. β is divided into $\dot{\beta}$ and $\ddot{\beta}$ respectively for $0.7h_{br} > w_{br}$ and $0.7h_{br} < w_{br}$. Thus,

$$\beta = \begin{cases} \text{Wave} & \text{if } \dot{\beta} \text{ and SSA} \\ \text{Kick} & \text{if } \ddot{\beta} \text{ and } 2 \leq \phi(t) \leq 15 \text{ and SSA} \\ \text{Bend} & \text{if } \ddot{\beta} \text{ and } 170 > \theta(t) > 190 \\ & \text{and } |H - F| < 1.5D_{seg} \text{ and SSA.} \end{cases} \tag{3.14}$$

The intermediate postures shown in Fig. 3.16 are selected by Wave IBM for labelling, since the subject has yet to attain any defined posture. The Punch action is similar to throwing a ball involving late cocking, acceleration, and follow through. In follow through, the arm moves across the body in a diagonal manner and as a result the angle $\phi(t)$ of body from the vertical is quite large [92]. Punch action in $\ddot{\beta}$ is labelled using Wave IBM when $\phi(t) > 15$. The range of $\phi(t)$ in Kick IBM is in between the Stand posture (with tolerance for leaning) and the Punch action. The global angle $\theta(t)$ are 170 and 190, respectively, for Left and Right Bend. The Bend IBM criteria is formulated based on human vision and kinesiology. The Smart Search Algorithm (SSA) in Section 3.4.3.4 uses (Eq. 3.14) in labelling SBPs in Wave, Kick, and Bend IBM.

3.4.2 Silhouette feature reduction

The convex hull method [105] is used to determine SBPs which are located at convex points of a contour as shown in Fig. 3.3 (a), where the line surrounding the silhouette is its convex hull and the shaded regions are its convexity defects. The convexity defects yield a number of convex points on contour which are marked as Head (H), Arm (A), Feet (F), etc. using the IBM criteria in Section 3.4.3 and as illustrated in Fig. 3.3. The convex points (x_{cv}, y_{cv}) are normalised with respect to their bounding rectangle to increase the computational speed as follows

$$x_{nr} = \frac{|x_{cv} - u_{br}|}{w_{br}} \ , \ y_{nr} = \frac{|y_{cv} - v_{br}|}{h_{br}} \tag{3.15}$$

within [0,1]. The Euclidean distance between convex points is computed as

$$DT_{cv}(i) = \sqrt{(cx_{cv} - px_{cv})^2 + (cy_{cv} - py_{cv})^2} \tag{3.16}$$

where (cx_{cv}, cy_{cv}) and (px_{cv}, py_{cv}) respectively denote the current and previous convex points, and i is the number of convex points. Convex points are close to each other in a high resolution video frame but further apart in a low resolution one. This is because in high resolution there are more frequent and sharper edges which will results in more convex points. A threshold Th which is proportional to the frame width I_w, frame height I_h and resolution factor Υ are used to remove nearby convex points, where

$$Th = I_w I_h \Upsilon \tag{3.17}$$

and Υ (determined experimentally) is fixed as follows:

$$\Upsilon = \begin{cases} 0.05 & \text{if } I_w, I_h \leq 200 \\ 0.007 & \text{if } I_w, I_h \geq 400 \\ 0.01 & \text{if } 200 < I_w, I_h < 400. \end{cases} \tag{3.18}$$

A convex point (x_{cv}, y_{cv}) is selected for labelling by first checking if $DT_{cv} > Th$, where Th is determined by using (Eq. 3.17) and (Eq. 3.18).

3.4.3 Significant Body Point (SBP) labelling

The best IBM is used to label normalised convex points (x_{nr}, y_{nr}) as SBP using Table 3.3 as follows. The following SBPs are labelled: Head (H), Arm/hand (A),

Knee (K) and Feet (F). Convex points (x_{cv}, y_{cv}) are compared with x_c and y_c based on (Eq. 3.11) to determine Upper body, Lower body, Right side and Left side. The ranges for Sit and Lie have been determined in the MuHAVi data set since it contains the Collapse and Standup activities. Body sides $\delta 1$, $\delta 2$, $\delta 3$ and $\delta 4$ are described in Table 3.2.

3.4.3.1 Stand

In the Stand posture, Stand to Lie, and Lie to Stand, clockwise and anti-clockwise directions, Head and Feet are respectively assigned using

$$\mathrm{H} = \left\{ (x_{nr}, y_{nr}) | y_{nr} < G1 \quad \text{if } \delta 1 \right. \tag{3.19}$$

$$\mathrm{F} = \left\{ (x_{nr}, y_{nr}) | y_{nr} > G5 \quad \text{if } \delta 2. \right. \tag{3.20}$$

Arm in the Stand posture, Stand to Lie, and Lie to Stand for clock and anti-clockwise directions are respectively assigned using

$$\mathrm{A} = \left\{ (x_{nr}, y_{nr}) | G2 < y_{nr} \leq G4 \quad \text{if } \delta 3/\delta 4 \right. \tag{3.21}$$

$$\mathrm{A} = \begin{cases} (x_{nr}, y_{nr}) | y_{nr} > G4 & \text{if } \delta 3/\delta 4 \ \& \ \delta 1/\delta 2 \\ (x_{nr}, y_{nr}) | G2 < y_{nr} \leq G4 & \text{if } \delta 3/\delta 4 \ \& \ \delta 2. \end{cases} \tag{3.22}$$

3.4.3.2 Sit

In the Sit posture, Stand to Lie, and Lie to Stand, clock and anti-clockwise direction, Head and Feet are respectively assigned using

$$\mathrm{H} = \left\{ (x_{nr}, y_{nr}) | y_{nr} < G1 \quad \text{if } \delta 3/\delta 4 \ \& \ \delta 1 \right. \tag{3.23}$$

$$\mathrm{F} = \left\{ (x_{nr}, y_{nr}) | y_{nr} > G5 \quad \text{if } \delta 3/\delta 4 \ \& \ \delta 2. \right. \tag{3.24}$$

The Arm is respectively assigned for Stand to Lie, and Lie to Stand for clockwise and anti-clockwise directions using

$$\mathrm{A} = \left\{ (x_{nr}, y_{nr}) | G1 < y_{nr} \leq G2 \quad \text{if } \delta 3/\delta 4 \ \& \ \delta 2 \right. \tag{3.25}$$

$$\mathrm{A} = \left\{ (x_{nr}, y_{nr}) | y_{nr} \geq G5 \quad \text{if } \delta 3/\delta 4 \ \& \ \delta 2. \right. \tag{3.26}$$

3.4.3.3 Lie

In the Lie posture, Stand to Lie, and Lie to Stand, clockwise and anti-clockwise directions, Head and Feet are respectively assigned using

$$\mathrm{H} = \begin{cases} (x_{nr}, y_{nr}) | x_{nr} < G1 & \text{if } \delta1/\delta3 \ \& \ \delta4 \\ \ \& \ y_{nr} < G1 & \text{if } \delta1/\delta3 \ \& \ \delta4 \\ (x_{nr}, y_{nr}) | x_{nr} < G1 & \text{if } \delta1/\delta3 \ \& \ \delta4 \end{cases} \tag{3.27}$$

$$\mathrm{F} = \left\{ \ (x_{nr}, y_{nr}) | x_{nr} > G5 \quad \text{if } \delta2. \right. \tag{3.28}$$

Head is also assigned using

$$\mathrm{H} = \begin{cases} (x_{nr}, y_{nr}) | x_{nr} \geq G2 \ \& \ y_{nr} \geq G4 & \text{if } \delta1 \\ \text{ or } x_{nr} > G2 \ \& \ y_{nr} < G5 & \text{if } \delta1 \\ \text{ or } x_{nr} \leq G4 \ \& y_{nr} > G4 & \text{if } \delta2. \end{cases} \tag{3.29}$$

For Stand to Lie and Lie to Stand, clockwise and anti-clockwise directions, arm and head are respectively assigned using

$$\mathrm{A} = \left\{ \ (x_{nr}, y_{nr}) | G1 < x_{nr} \leq G2 \quad \text{if } \delta3/\delta4 \right. \tag{3.30}$$

$$\mathrm{H} = \left\{ \ (x_{nr}, y_{nr}) | x_{nr} < 0.5G1 \quad \text{if } \delta1 \ \& \ \delta3/\delta4. \right. \tag{3.31}$$

In Lie to Stand, as the subject is trying to stand, support of arms is used to assist in manoeuvring. (Eq. 3.22) for Lie to Stand is utilized for labelling SBPs as the subject is manoeuvring from Sit to Stand. However, during this manoeuvring when $h_{br} > 1.7w_{br}$, (Eq. 3.21) is used instead of (Eq. 3.22).

3.4.3.4 Smart Search Algorithm (SSA)

In the β activities, i.e., Wave, Kick, and Bend IBMs, SSA is used to label SBPs. Based on (Eq. 3.14) SSA is initiated to locate the convex points in the non anthropometric segment ranges. $\dot{\beta}$ refers to the subject in the Stand posture who has yet to attain the posture of models shown in Fig. 3.4 (a)-(c). It is an indication that the subject is likely to perform Wave. In Fig. 3.4 $H(t-1)$ and $H(t)$ are respectively the location of previous (x^H_{t-1}, y^H_{t-1}) and current (x^H_t, y^H_t) head points, and ϵ is the horizontal distance between them. SSA divides the wave model into four horizontal segments, and as the hand goes near or above the head, the following steps are

defined for labelling convex points as SBPs in the segment range [G1-G4] as shown in Fig. 3.4 (a):

Step 1: Locate the arm in the segment range $G(1,2]$ of shoulder S by dividing the bounding rectangle width w_{br} into three equal vertical sections, and reallocate normalised convex points (x_{nr}, y_{nr}) as arm point A if $x_{nr} < w_{br}/3$ or $x_{nr} > 2w_{br}/3$ or $|y_{nr} - y^H| > 0.7D_{seg}$ represented by the shaded region in Fig. 3.4 (a).

Step 2: Verify no arm point was identified using Step 1. Next, every normalised convex point (x_{nr}, y_{nr}) in the head segment range $G[1]$ of Stand to Lie, clockwise and anti-clockwise directions, is reallocated as A if $\epsilon > 0.7D_{seg}$, where $\epsilon = |x_t^H - x_{t-1}^H|$ as shown in Fig. 3.4 (a).

Step 3: Check if no arm point has been labelled using the above two steps. Find two points in the segment range [G1-G4] that are at maximum distance from the centre and lie to its right and left, respectively, denoted by arrows in Fig. 3.4 (a). These points are then labelled as arm points.

Step 4: If an arm point is labelled using one of the above three criteria then it implies that a wave IBM best represents the activity, hence the head point is reallocated as follows: $x^H = x_c,\ \ y^H = y_c - \tau D_{seg}$, where $\tau = 1, 1.7, 2.5$ respectively for resolution factor $\Upsilon = 0.05, 0.007, 0.1$. This is based on the fact that the centre of mass moves upward when the human arms are above the head.

In $\ddot{\beta}$ based on (Eq. 3.14), for the kick IBM, only Step 1 and 2 of the SSA are invoked. Steps 1 and 2 are used in the segment range of the arm $G(2,4]$ and $G[1]$ to reallocate foot point for right and left Kick as shown in the shaded region of Fig. 3.4 (b). In $\ddot{\beta}$ for Bend IBM, the global angle $\theta(t)$ is near Sit, and the head to feet distance reduces (denoted by dashed arrows) in Fig. 3.4 (c). This model slightly overlaps with the Sit model of Stand to Lie, and Lie to Stand, hence, Sit criteria Stand to Lie in Section 3.4.3.2 is used to label SBPs. Depending upon the global angle the proposed framework automatically switches to Lie to Stand using Fig. 3.13 (b).

3.4.4 2D Stick figure

Researchers mostly use a manual or semi-automated selection of human joints on images to construct a model and trajectories [7,8,74,86]. The information extracted from this is then utilized for applications such as trajectory analysis, activity recognition, sit to stand analysis, etc. The proposed framework can be used for the

animation of the stick figures of a human body formed by joining the SBPs of every video frame. To form a stick figure, first the maximum distance between shoulder point (x^S, y^S) and head point (x^H, y^H) is computed as

$$x^S = \max(x^H - x^S) \ , \ y^S = \max(y^H - y^S) \tag{3.32}$$

for an activity. Noting that a shoulder point is mostly at a constant distance from the head point, (Eq. 3.32) is used to find a shoulder point (x^S, y^S) for all activities. According to human anatomy, the head and feet points are connected to the centre (x_c, y_c) of the silhouette contour and the arm points are connected to the shoulder point (x^S, y^S) as shown in Fig. 3.19.

3.5 Experimental Results

Most methods in Section 3.2 only provide qualitative evaluation. In W4 system [19], [20] for Computer Vision based Human body Segmentation and Posture estimation (CVHSP), [21] for Star skeletonization (STAR), [22] for extremities as posture representation, and the fast detection and modelling of human body parts (FDMHP) method in [38], SBPs are detected but the accuracy of their localization with respect to ground truth coordinates of each SBP is not presented. Thus, it is not possible to compare the accuracy of SBP localization using the proposed framework with these methods. Therefore, qualitative results are presented in Section 3.5.1 for comparison with these methods.

This absence of quantified evaluation in the other reported work makes it necessary to perform ground truth mark-up in order to obtain quantified evaluation in this work. Silhouette contours for all activities of the two data sets are skeletonized using the method in [106]. Manual annotation is performed on the results of the skeletonized silhouette using mouse cursor to obtain ground truth coordinates of SBPs as shown in Fig. 3.17 and Fig. 3.18 for the Weizmann [58] and MuHAVi [73] data sets respectively. Note that the manual annotation of ground truth also involves some guesses of SBPs in cases where these points are not localized by skeletonization or not clearly visible to the human eye. The accuracy of SBP localization is presented in Section 3.5.2 in terms of distance in pixels between the manually annotated (i.e., the ground truth) and detected SBPs.

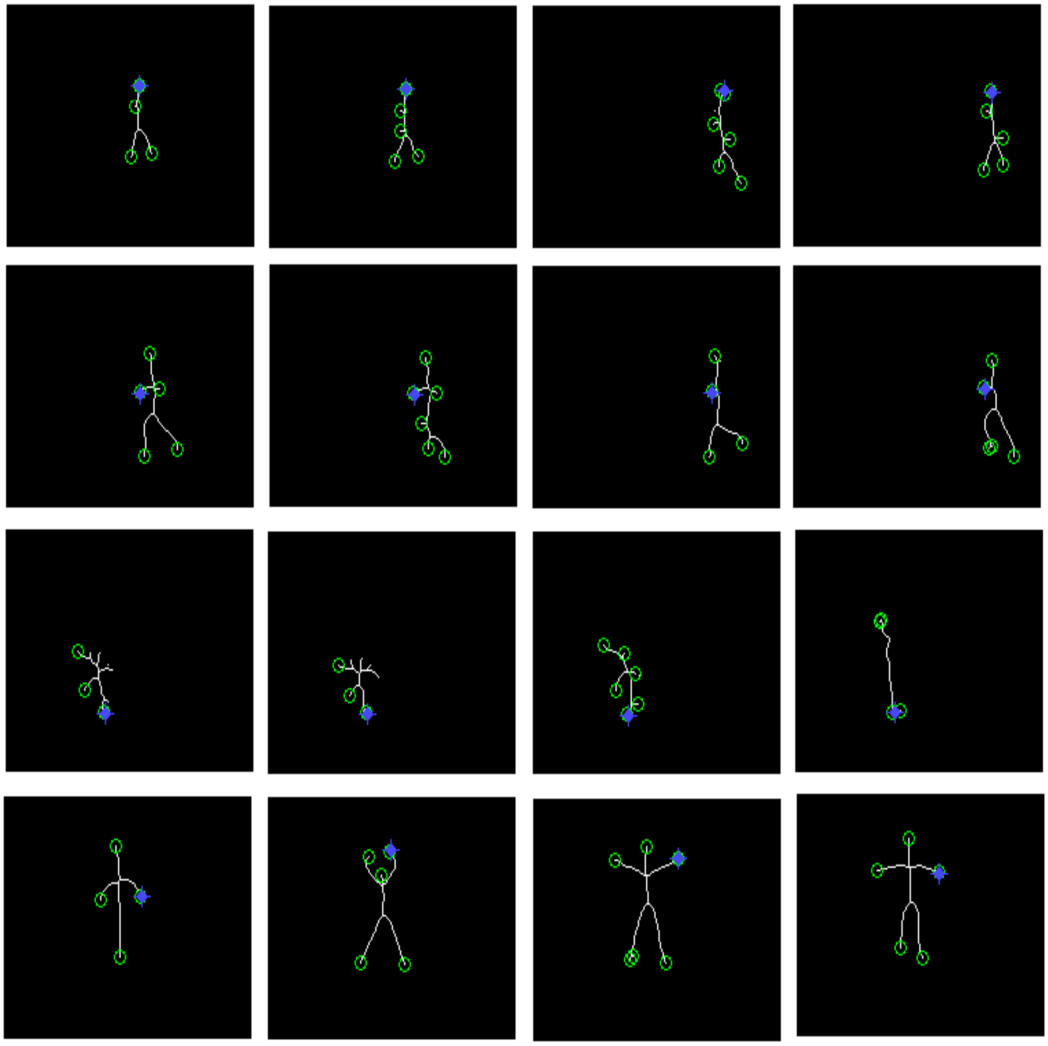


Figure 3.17: Examples of annotated (blue target) SBPs (green circle) on the Weizmann data set. Side, Run, Bend and Jack from top to bottom row.

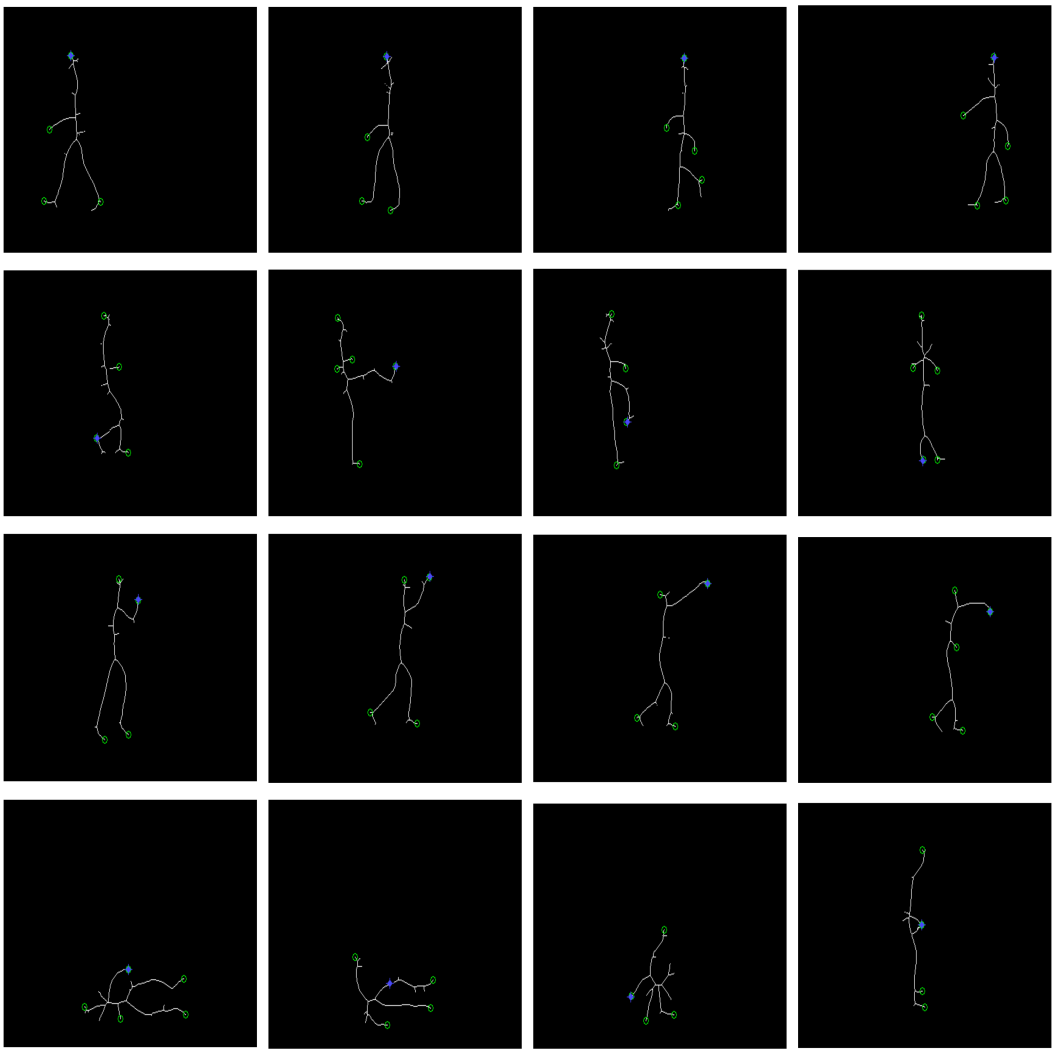


Figure 3.18: Examples of annotated (blue target) SBPs (green circle) on the MuHAVi data set. Walk, Kick, Punch and Standup from top to bottom row.

In [33] for Connectivity based human body modelling (CBHM) only 4 SBPs are evaluated quantitatively and they do not provide their data set for comparison. Also, First Sight method [37] detects body parts and not SBPs. Section 4.5.2.3 contains the quantitative comparison with these methods after the tracking method (Chapter 4) is incorporated in the proposed framework.

3.5.1 Qualitative evaluation

In Fig. 3.19 the Freeman chain code contours of subjects enclosed in the bounding rectangle and the rescaled ellipse, with generated stick figures and labelled SBPs are shown for qualitative evaluation on the activites of Weizmann data set. The left column shows the Walk, Side, Skip, Jump, the middle column shows the Jump-in-place-on-two-legs activities, Run, One Hand Wave, Two Hand Wave and the right column shows the Jack and Bend activities. It can be observed that the proposed SBP framework accurately detects and labels Head (H), Arm (A), Shoulder (S), Knee (K) and Feet (F) on the low resolution videos of the Weizmann data set. It can be seen that the proposed framework based on IBMs is able to robustly label SBPs in all the actions. An initial missed or undetected convex point, results in an incomplete stick figure.

The adaptability and generality of the proposed framework is validated by applying it with the same parameter settings on the MuHAVi data set. Fig. 3.20 shows the labelled SBPs on the high resolution videos of the MuHAVi data set. It can be seen that the proposed framework is capable of detecting SBPs in all the actions. The first row in Fig. 3.20 shows SBPs labelled on the (a)-(b) Walk and (c)-(d) Run actions. The second row shows identified SBPs on the (e)-(f) Punch, (g)-(h) Kick actions. The last two rows show labelled SBPs in Collapse and Standup actions respectively. Fig. 3.19 and Fig. 3.20 show that the proposed framework successfully labels SBPs and is able to generate stick figures in various activities.

The qualitative results on both the data sets show that the proposed framework is capable of detecting SBPs in both low and high resolution videos of 15 activities that involve rapid movements and posture changes. In the reported work [19], [20], [21], [22], [33], [37] and [38] only 2-14 activities have been used for qualitative evaluation on either low or high resolution videos.

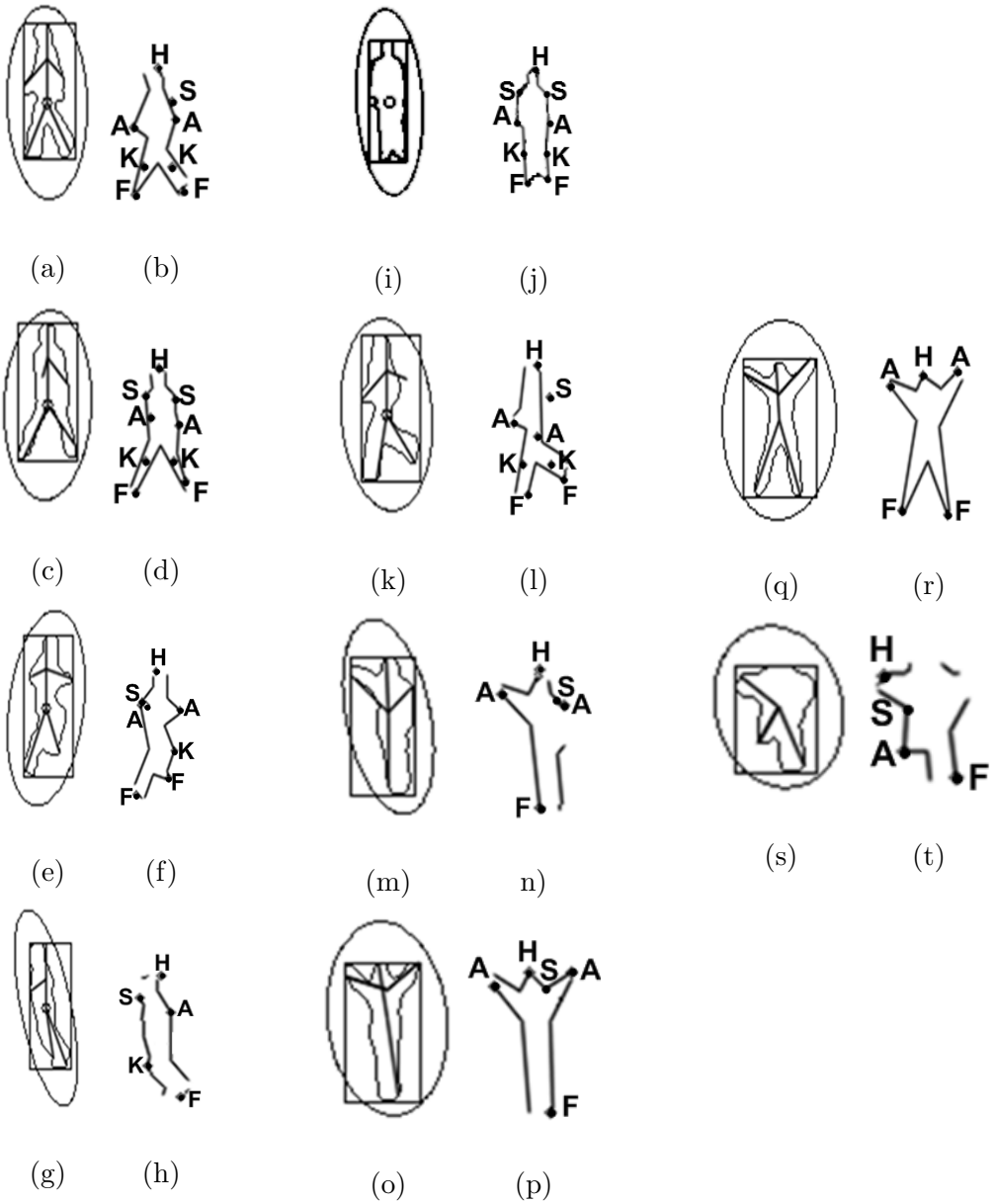


Figure 3.19: Weizmann data set. (a)-(b) Walk, (c)-(d) Side, (e)-(f) Skip, (g)-(h) Jump, (i)-(j) Jump-in-place-on-two-legs, (k)-(l) Run, (m)-(n) One Hand Wave, (o)-(p) Two Hand Wave, (q)-(r) Jack and (s)-(t) Bend respectively (Contour, bounding rectangle, ellipse and stick figure). SBPs labelled as Head (H), Shoulder (S), Arm (A), Knee (K) and Feet (F) in the corresponding activities. Note that S and K are displayed in some cases to show that it is possible to determine more than 5 SBPs using the proposed framework.

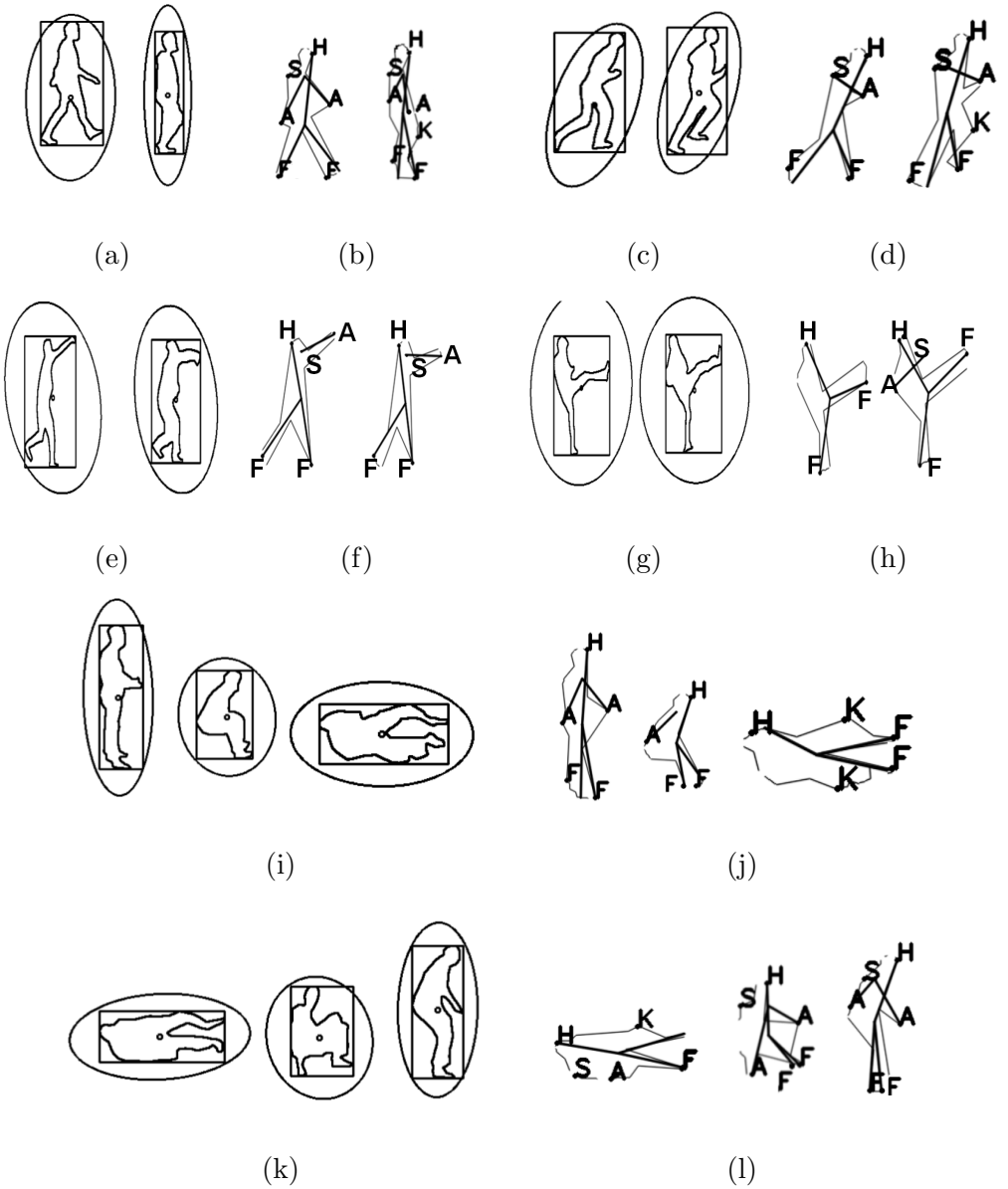


Figure 3.20: MuHAVi data set. SBPs labelled as Head (H), Shoulder (S), Arm (A), Knee (K) and Feet (F) in (a)-(b) Walk, (c)-(d) Run, (e)-(f) Punch, (g)-(h) Kick, (i)-(j) Collapse and (k)-(l) Standup. Note that S and K are displayed in some cases to show that it is possible to determine more than 5 SBPs using the proposed framework.

3.5.2 Quantitative evaluation

3.5.2.1 Accuracy of localization

The location of every SBP obtained using the proposed framework is compared with the ground truth in each frame of the video sequence. The overall accuracy of the proposed framework is defined by the average error in pixels in detecting each SBP, i.e.,

$$Error(x_{avg}, y_{avg}) = \frac{\sum_{n=1}^{N} |G_n(x,y) - L_n(x,y)|}{N} \tag{3.33}$$

where $G_n(x,y)$ and $L_n(x,y)$ are respectively the coordinates of each SBP obtained from the ground truth and the proposed framework, and N is the total number of frames.

The average error in x and y coordinates of each SBP, i.e., Head (x^H, y^H), Front Arm (x^{FA}, y^{FA}), Back Arm (x^{BA}, y^{BA}), Left Foot (x^{LF}, y^{LF}), and Right Foot (x^{RF}, y^{RF}), in various activities (see Table 3.1) performed by all subjects of both data sets is shown in Table 3.4. For Jump-in-place-on-Two-Legs/Pausejump ($\beta 7$), Side ($\alpha 4$), and Walk ($\alpha 1$) of the Weizmann data set (which have less lateral head movement), the x-coordinate head error is less than other activities whereas the y-coordinate head error is similar in all activities. The front and back arm points are occluded more than any other SBPs, hence they have greater errors. A common average error is obtained for the right and left foot because they are joined in Jump ($\alpha 5$), Jump-in-place-on-Two-Legs ($\beta 7$), One Hand Wave ($\beta 9$), and Two Hand Wave ($\beta 10$). The feet have smaller vertical movement than horizontal movement in consecutive frames in all activities, hence, the average y-coordinate error is less than the x-coordinate for both feet. For the MuHAVi data set, the y-coordinate head error is less than the x-coordinate average error in all activities. The errors in the front and back arm points are also greater due to occlusion. The highest average error occurs in Collapse and Standup due to severe self occlusion of front and back arms. The right and left feet have similar average errors. The average Avg of five SBP errors per activity is presented in the last column of Table 3.4. In Table 3.4 and Table 3.5 the best results are shown in bold.

Weizmann and MuHAVi data sets have $180 \times 144 = 25920$ pixels and $720 \times 576 = 414720$ pixels per frame, respectively. An overall average error of 5.02 and 7.8 pixels in location of SBPs on all activities for five SBPs (from average of last column of Table 3.4), respectively, on two diverse data sets show that the proposed

Table 3.4: Average Error in pixels of SBPs w.r.t Ground Truth. Mean Height is 68 and 200 pixels for Weizmann and MuHAVi data set respectively.

Activity	x^H	y^H	x^{FA}	y^{FA}	x^{BA}	y^{BA}	x^{LF}	y^{LF}	x^{RF}	y^{RF}	Average
	Weizmann Data set with prediction										
Walk	2.3	5.5	5.3	7.5	4.8	10.3	4.6	2.4	4.3	2.3	4.93
Run	3.8	5.6	5.3	**3.4**	8.7	8	5	3.7	4	3.4	5.09
Skip	4.3	5.4	7	5.9	8.6	6	5	4.1	3.8	2.1	5.22
Side	1.6	5	6.5	6.3	4.5	7.5	3.8	3.1	4	3.5	4.58
Jump	3.6	5.1	7.3	11	6.1	7.1	5.3	3.6	5.3	3.6	5.8
Pausejump	**1**	4.5	6.5	8.6	3.9	6.5	6.2	2.9	6.2	2.9	4.92
Bend	7.3	6.5	7.2	9.6	5	6.8	4.2	2.5	4.2	2.5	5.58
OneHandWave	9.6	5.4	5.2	6	**2.6**	5.2	6	1.7	6	1.7	4.94
TwoHandWave	5.7	**4**	8.5	8.5	8.6	8.7	6	**1.6**	6	**1.6**	5.92
Jack	5.3	**4**	**3.3**	4.4	2.8	**3.3**	**2.4**	2	**3.2**	2.3	**3.3**
Average/Mean Height	0.06	0.07	0.09	0.1	0.08	0.1	0.07	**0.04**	0.07	**0.04**	
	MuHAVi Data set with prediction										
Walk	11	**3.3**	5.7	7.2	8.5	12.3	8	4.6	8.3	4.9	7.38
Run	9.65	3.8	6.4	6.7	9.2	16.3	8.3	5.2	9.7	6	8.12
Turn	10.2	3.7	5.7	11.9	**5.3**	14.2	7.7	4.4	8	**4.3**	7.54
Standup	9	5.2	32	23.5	11.7	13	12	10.4	11.4	7	13.52
Collapse	8.4	5.5	11.6	11.2	7.7	5.6	9.8	8.4	13.1	8.5	8.98
Kick	10.8	4.9	4.1	5.4	6.5	**5.2**	11.5	9.5	7.2	6.5	7.2
Punch	8.6	4.9	3.6	6.4	7.5	6.4	4.3	3.3	7.4	4.6	5.7
Guard − to − Kick	7.3	5.6	**2.9**	4.9	7.9	5.4	3.8	4.3	**6.2**	8	**5.6**
Guard − to − Punch	**5.5**	5.8	3.3	**3.2**	6.1	10.7	**3.7**	**3.1**	10.3	6.3	5.78
Average/Mean Height	0.04	**0.02**	0.04	0.04	0.04	0.05	0.04	**0.03**	0.04	**0.03**	

framework is accurate and adaptable to data sets of different resolution.

The average error in pixels as a proportion of the mean height of subjects for all the activities of Weizmann and MuHAVi data set are shown in the last rows of Table 3.4. This can be used to have a picture of how much an error, e.g., 5 pixels, means with respect to the size of the human body. For example, the human head is one-eighth the human height, i.e., 0.125. Hence, a 5 pixel error equates to approximately 0.07 that is almost half of the height of the human head. In Table 3.4 the average error as proportion of the mean height is between 0.04 and 0.1 for the Weizmann data set. It can be seen that the average error in pixels of all the five SBPs as a proportion of the mean height of subjects for high resolution MuHAVi data set is consistently lower than Weizmann data set.

3.5.2.2 Accuracy of detected SBPs vs observed

The accuracy of detection is evaluated in terms of precision (PR), recall (RC), and error (ER), i.e.,

$$PR = \frac{\sum_1^q CT}{\sum_1^q DT} \tag{3.34}$$

$$RC = \frac{\sum_1^q CT}{\sum_1^q OB} \tag{3.35}$$

$$ER = \frac{\sum_1^q DT - \sum_1^q CT}{\sum_1^q DT} \tag{3.36}$$

where DT and CT are respectively the number of detected and correctly detected SBPs. OB is the observed SBPs and q is the number of subjects. The number of detected SBPs includes misclassified SBPs which are manually counted by visual inspection on every frame of video sequence. The number of correctly detected SBPs is obtained by deducting misclassified SBPs from the number of detected SBPs.

The detection accuracy of five SBPs is computed by using the proposed framework first with no prediction and then with Particle Filter prediction. This demonstrates the impact of prediction on the performance of the framework. In Table 3.5 for SBP detection with no prediction, observed (OB) SBPs are the manually counted visible SBP only with no guess work involved.

In Table 3.5, for no prediction, smaller recalls are obtained for Run ($\alpha 2$), Skip ($\alpha 3$), Jump ($\alpha 5$), and Two Hand Wave ($\beta 10$) that have abrupt human limb movement as compared to Walk ($\alpha 1$), Side ($\alpha 4$), Jump-in-place-on-Two-Legs ($\beta 7$),

Table 3.5: Precision and Recall of SBP detection with no prediction.

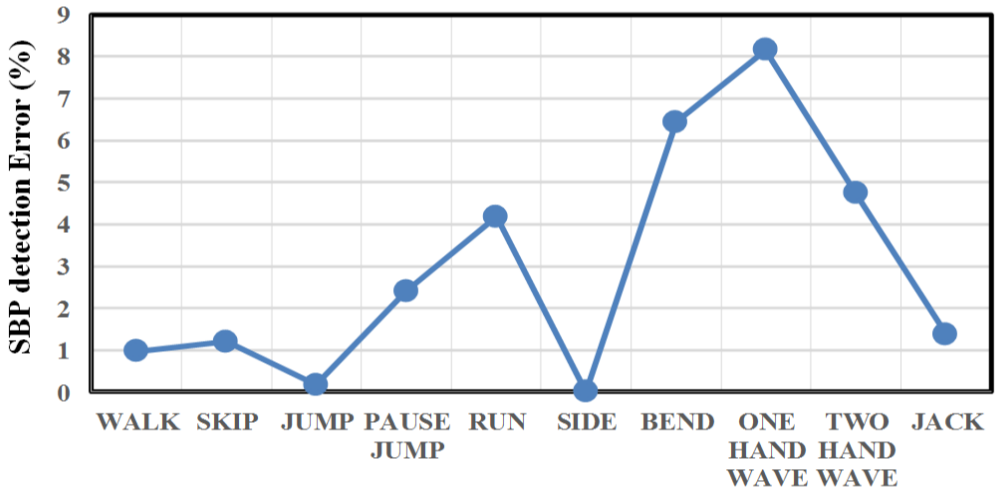
Weizmann Data set					
Activity	CT	OB	DT	RC%	PR%
$Walk^9$	2655	2768	2681	95.9	99
$Skip^9$	1566	1664	1585	94.1	98.8
$Jump^9$	1756	1877	1759	93.5	**99.8**
$PauseJump^9$	2231	2271	2286	**98.2**	97.6
Run^9	1468	1623	1532	90.4	95.8
$Side^9$	1726	1786	1726	96.6	**100**
$Bend^9$	3067	3195	3278	96	93.6
$OneHandWave^9$	3265	3265	3555	**100**	91.8
$TwoHandWave^9$	2875	3120	3018	92.1	95.3
$Jack^9$	3157	3370	3201	93.7	98.6
MuHAVi Data set					
Activity	CT	OB	DT	RC%	PR%
$Walk^4$	1188	1231	1191	96.2	99.8
$Collapse^4$	1131	1306	1152	86.6	98.1
$Standup^4$	1431	1471	1505	97.4	95
$Turn^4$	868	1046	868	83	**100**
Run^4	975	1198	985	81.4	99
$Guard-to-Punch^4$	529	533	529	**99.2**	**100**
$Punch^4$	729	757	739	96.3	98.6
$Guard-to-Kick^4$	503	512	507	**98.2**	99.2
$Kick^4$	828	922	865	89.8	95.7

Bend ($\beta 8$) and One Hand Wave ($\beta 9$). The smallest recall and precision respectively occur in Run ($\alpha 2$) and One Hand Wave ($\beta 9$). The maximum recall and precision, respectively, occur in Side ($\alpha 4$) and One Hand Wave ($\beta 9$). The proposed framework with no prediction obtains an overall average $Avg\%$ recall and precision of 95.3% and 96.5%, respectively, for all activities of the Weizmann data set. On the MuHAVi data set it obtains the smallest recall for Run ($\alpha 2$) but is robust in detecting SBPs in Walk ($\alpha 1$), Standup ($\beta 12$), Punch ($\beta 15$), Guard-to-Kick ($\beta 16$) and Guard-to punch ($\beta 17$). In Turn ($\alpha 6$), Collapse ($\beta 13$), and Kick ($\beta 14$) it is able to produce SBPs with reasonable accuracy. It has the least precision for complex movement such as Standup ($\beta 12$). It achieves an overall average $Avg\%$ recall and precision of 92.01% and 98.4%, respectively, for all activities of the MuHAVi data set.

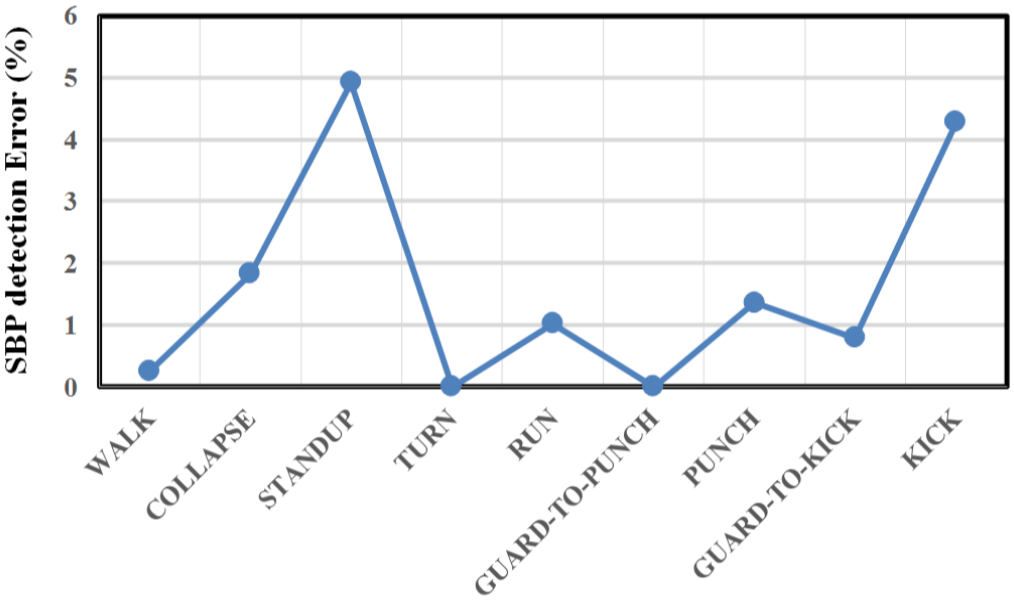
Fig. 3.21 (a) and (b) show the error in percentage % in significant body point labelling on the Weizmann and MuHAVi data sets respectively. In Fig. 3.21 (a) the error in SBP detection is more for Bend ($\beta 8$), One Hand Wave ($\beta 9$) and Two Hand Wave ($\beta 10$). This is because in the Bend ($\beta 8$) the arm goes below the knee and close to feet which might cause missed arm points while in the One Hand Wave ($\beta 9$) and Two Hand Wave ($\beta 10$) the arm goes above the head that creates a convex hull with peaks as arm points and a valley at the head point that is not detected as a convex point. In Fig. 3.21 (b) more error in SBP detection is observed for Collapse ($\beta 13$), Standup ($\beta 12$) and Kick ($\beta 14$). A possible reason for more error might be rapid postural changes that affect the SBP detection in these activities. The average error for all activities of the Weizmann and MuHAVi data sets computed using (Eq. 3.36) are 3.5% and 1.9%, respectively. This shows that the proposed framework robustly labels SBPs in both low and high resolution videos containing several complex activities with rapid limb movement and posture changes.

3.6 Summary

In this chapter, a novel automated marker-less implicit body model-based human significant body points (SBPs) detection and labelling framework is presented. It labels anatomical landmarks (e.g. Head, Hand/Arm, and Feet), which are referred to as significant body points, using six implicit body models innovated from human anthropometry, kinesiology and biomechanics. By considering the human body as an inverted pendulum model, ellipse fitting and contour moments are applied to classify it as being in the Stand, Sit or Lie posture. A convex hull of the silhouette



(a)



(b)

Figure 3.21: SBP detection error in pixels (%) using (Eq. 3.36) on (a) Weizmann data set and (b) MuHAVi data set, with no prediction.

contour is used to determine the locations of SBPs. Stick figures are generated by connecting SBPs. The results demonstrate that the proposed framework robustly locates and labels SBPs in several actions on two low and high resolutions data sets.

Chapter 4

Human Body Part Tracking

4.1 Introduction

In the past decade, marker-less articulated human motion analysis and tracking has been a prime focus of research in the computer vision research community due to its numerous applications. Robust tracking requires dealing with occlusion, variance in illumination, rapid motion, view invariance, structural ambiguity, multiple subjects, etc. Sequential Monte Carlo methods, also known as the Particle Filters (PFs), have been extensively used to address such problems [55]. Monte Carlo methods have applications in many fields of sciences, e.g., medical imaging [107], engineering, finance etc. The human body has high dimensions, i.e., degree of freedom, and human motion is non-linear and non-Gaussian in nature. The ability of the Particle Filter to represent non-Gaussian non-linear assumption and multiple hypothesis makes it suitable for visual tracking.

A Marker-less implicit body model based (IBM) human motion analysis framework that is able to detect and label significant body parts or points (SBP) was presented in Chapter 3. In this Chapter, two methods, i.e., Particle Filter with memory and feedback (PFMF), and Motion Flow (MFL), based prediction are presented to track the 2D image coordinates of SBPs. The standard Particle Filter struggles in prediction when there is no measurement in the image (i.e., in occlusion). The proposed Particle Filter combines the temporal information of the previous observations and estimation with a feedback to predict SBPs in occlusion. The motion flow based method considers the human arm as a pendulum attached to the shoulder joint. The arm is one of the most occluded body parts or points in

various activities. Hence, a prediction method specifically designed to predict arms is useful. MFL considers arm motion like a pendulum swing and defines conjectures to predict SBPs in occlusion.

4.2 Literature review

Real-time detection and tracking of humans from videos require estimation of the subject's states such as location, orientation, size, etc. This is not as simple as it seems to be because of the missed detection, artefacts, and false detection due to clutter [108]. Although researchers have proposed various solutions to human body tracking, a universal human body tracker capable of handling real-time scenarios does not yet exist. This reveals the complexity of the task. Most of the research is focused in developing articulated-model based systems to track the human body in videos. A realistic articulated human body model has at least 25 degree of freedom. Due to the high dimensionality of the human body model and the exponentially increasing computational speed, specialized algorithms such as a Particle Filter is required to perform complete human body tracking in videos [50].

4.2.1 Particle Filter

Estimation is a process by which we infer the value of a quantity of interest, by processing data that is in some way dependent on it. A Particle Filter is composed of two words; particle, and filter. Particles are a set of randomly chosen weighted samples used to approximate a probability density function. A Filter is a procedure that estimates parameters (state) of a system. State estimation is based on probability theory. A Particle Filter has three operational steps, i.e., sample, predict and estimate, as described in Section 2.2.2.

The Particle Filter which is also known as the condensation algorithm was first introduced for visual tracking by Isard and Blake in 1998 [9, 10]. However, it lacks the ability to work in real-time since the number of particles is large in order to account for sudden movements of the object being tracked. Due to a large search space, a large degree of freedom of the human body increases the computational complexity and cost exponentially. Techniques such as partitioned sampling by MacCormick [47], layered sampling by Sullivan [48, 49], and annealed Particle Filtering [50] have been proposed to reduce the search space. The Partitioned sampling is a variation of the Particle Filter that reduces the number of particles required

to perform multiple object tracking. It was applied to the problem of articulated tracking of objects by MacCormick and Isard [47]. The use of partitioned sampling reduces the search space by partitioning it for more efficient Particle Filtering and thus making the problem in hand more tractable. Nevertheless, this method is not extendible for complete human body posture recognition. The layered sampling approach proposed by Sullivan et al. is another variation of the standard Particle Filter. In [48, 49] the number of particles required to describe the posterior density is also reduced. It utilizes the concept of importance sampling to reduce the search space. A better use of a particle set allows the removal of ambiguities arising from human kinematics. This method has been shown experimentally to suffer when the tracking complexity increase above 30 degree of freedom [48, 49]. Partitioned annealed Particle Filtering is an approach proposed by Deutscher to enhance the efficiency of the annealed Particle Filter [50]. It slowly initiates the influence of narrow peaks in the fitness function by utilizing a continuation principle which is based on annealing. The algorithm is able to recover complete articulated human body motion swiftly. This method is more effective in reducing the number of particles required for tracking. It is capable of handling tracking for more than 30 dimensions [50]. In [43], an analytical inference is incorporated into the framework of the Particle Filter to alleviate the computational burden. It is also useful for automatic initialization and recovering from tracking failure. The state parameters describing the human posture are updated using the analytical inference supplied by the body parts detection. This aids in reducing the number of particle required for tracking and the extent of randomness. The modified Particle Filter is much more robust then the standard Particle Filter.

The Particle Filter algorithm suffers from inefficiency in sampling due to degeneracy (in which the weights of the majority particles become small after a few iterations) and impoverishment (samples are too concentrated) [55]. Also, a large number of particles is required to overcome the samples impoverishment problem by populating some areas of the state-space that may be left empty due to prediction of the motion model that tends to cluster the particles in a small area due to the predicted motion. Mean shift is used to trace the local maximum of probability distribution in the direction of gradient and tracks single hypothesis. This makes it incapable of handling occlusions and similar objects in the video scene. Keeping in mind the pros and cons of mean shift and Particle Filter tracker, a novel technique was proposed by Shan et al. [54] which combines mean shift with the Particle Filter

to come up with a Mean Shift Embedded Particle Filter (MSEPF). The particles are herded (grouped) near local modes with large probability by performing mean shift on every particle in the propagation phase of the Particle Filter. This addresses the problem of degeneration. The work of Koichiro et al. and Maggio and Cavallaro also merge mean shift with a Particle Filter [53]. These methods will inevitably concentrate the particles and would give rise to sample impoverishment. The Continuously Adaptive Mean Shift (CamShift) is an enhanced version of the mean shift procedure which was proposed by Bradski et al. in 1998. The concept of the MSEPF was extended by Zhaowen Wang et al. by incorporating the CamShift procedure with a Particle Filter to introduce the CamShift Guided Particle Filter (CAMSGPF) [55]. In the CAMSGPF, sampling efficiency is improved due to optimization of the scale and position of each particle by the CamShift procedure. The inclusion of the CamShift facilitates the use of fewer particles for tracking as compared to the standard Particle Filter. The multiple hypotheses tracking of the Particle Filter facilitates the CamShift to regulate scaling factors adaptively. Furthermore in the CAMSGPF, the CamShift method is modified to increase the efficiency of the algorithm. CAMSGPF is superior to the standard Particle Filter and mean shift based tracker in terms of robustness and efficiency [55]. The CAMSGPF has only been used to track a target in a video sequence enclosed by a rectangular window. The ability of this approach to efficiently track a complete human body is yet to be explored.

Several researchers have integrated colour information with the framework of the Particle Filter to perform robust tracking in complex scenarios. In [52], the standard Particle Filter has been enhanced to initialize and track multiple objects with the same colour. It utilizes the principle of an adaptive colour based Particle Filter. The adaptive colour based Particle Filter method is capable of efficiently handling variations in target dynamics and shape in complicated backgrounds but fails to track multiple objects with same colour. This limitation has been removed in [43] by integrating colour histograms as target object features in the framework of the Particle Filter. It also keeps a record of the number of targets present in the video sequence. The tracking mechanism of the smart camera architecture in [29] uses colour distributions in hue, saturation, and value for robust tracking. Particle Filters are used to track the region of interest, while a distinct colour-based Particle Filter is assigned to each new object. The approach is an automated distributed video surveillance system for tracking and activity recognition with major processing

embedded in each smart camera node. The information is processed in the sensor and only the results are transmitted. In order to evaluate the performance of the proposed system, a complete prototype system comprising of four smart cameras and one server PC were installed within a home for the elderly in Germany. If the occurrence of a person falling is detected, the person's location is marked with a red warning icon on the visualization node and broadcasted as a text message to a particular phone by means of the alarm handler [29]. Colour information is not reliable in scenes with varying illumination.

The 2D models proposed in literature for tracking are constrained to particular types of motions which are linear and restricted to a pre-set view point [42]. The articulated tracking in [42] is performed by tracking each limb with a dynamic Markov network and then refining the positions by adding constraints among various sub-parts with mean field Monte Carlo method. A novel method that utilizes a set of Particle Filters has been proposed to track the human body parts. It uses the Kalman and Particle Filters to perform articulated tracking of low human body parts. A 2D articulated model constrained by human biomechanics has been used for reducing the complexity of tracking. The 2D articulated model introduced is as robust as a 3D model in tracking the lower body motion. Tracking is accomplished by identifying the static foot during motion and storing its trajectory. Subsequently, human body parts are tracked by means of the proposed 2D articulated model using a set of Particle Filters. The constrained biomechanical articulated 2D model of human motion facilitates the analysis of 3D motion patterns. Due to this reason it is capable of handling variance in orientation, depth, and camera viewpoint. The scheme utilizes a set of Particle Filters to fit the proposed 2D articulated human model on every frame of the input video sequence. The tracking process of the articulated model is refined by instantiating two Particle Filters in parallel to the initial Particle Filter. This aids in addressing the degradation and potential divergence issues that arise in tracking while using a single Particle Filter [42].

In [51], a Particle Filter based tracker is proposed that adapts and balances uncertainty in its static and dynamic components of its state space model for visual tracking. A histogram based approach is utilized to describe the target. In [109], a Particle Filter for joint detection and tracking is proposed which uses a single particle to describe the number of objects in the scene and their surrounding boxes. This method refines the detections of colour objects instead of tracking them, thus, it is a time varying estimator rather than a tracker. The state density estimate is used

to perform tracking by marking corresponding states over time. This technique is dependent upon constructing an appearance model by segmenting the targets in the test sequence manually. The method in [109] is modified in [108] by incorporating an update of the measurement model using foreground detections with a background model, and labelling the tracks from the state density estimate. The efficiency of the proposed approach is enhanced by using threshold estimate.

A gravity optimised Particle Filter method was proposed in [56] that considers particles as masses and uses the Newton's law of universal gravitation to heard them locally. At each stage the new set of particles are replicated at the location nearer to where the particles are supposed to move. It has been shown to be successful for tracking fingers of human hand but with the finger performing a linear up and down motion. It does not present any procedure to address the sample impoverishment problem created due to concentrating the particles. Also, recent investigation of sampling laws for Particle Filter algorithms lead to the development of a new class termed as 'twisted' Particle Filters [110] that are validated with asymptotic analysis. Its ability to track in occlusion and on real data is not known.

The continuous human movement recognition framework in [41] uses forward smoothing Particle Filters with an optimized search space for tracking. This framework comprises tracking and recognition modules with a feedback from recognition, to a tracking module to optimize computation of the Particle Filter. If a subject wears loose clothes, this framework fails to recognize correct movements. The range of joint angles is restricted by limiting the degree of freedom related with each joint, thus fine movements are not modelled. Particle Filters used for tracking of thirty-two degree of freedom are computationally bulky. The method in [111] stores all the past estimations and observations in a memory module. It combines the standard Particle Filter with memory module to handle occlusion. It follows the standard Particle Filter when there is no occlusion and uses memory module when there is occlusion to perform robust tracking. It requires significant memory for storage and might produce incorrect prediction. For example, the past might be up movements and the most recent might be down movement. If all the past movements are taken into account then it will produce incorrect prediction. Therefore, a memory based strategy that involve the most recent information is explored in this work for robust tracking.

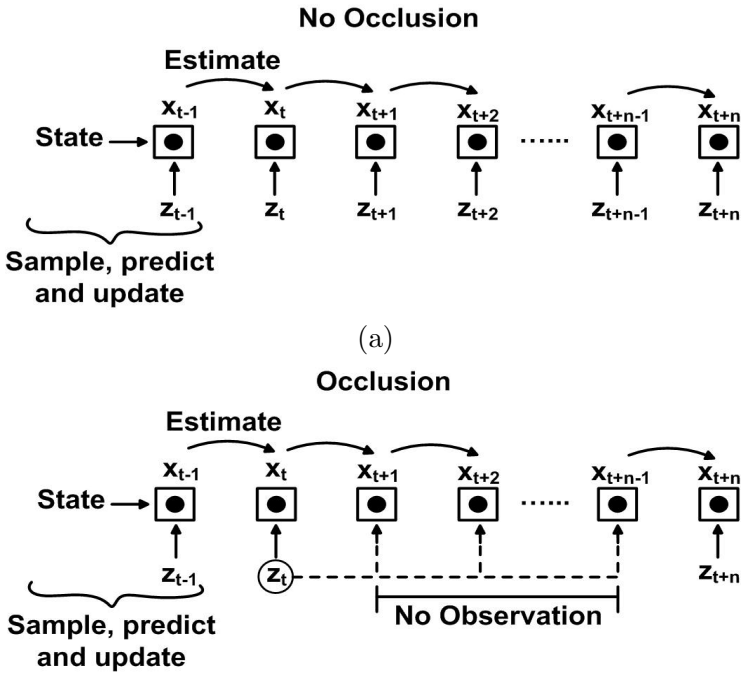


Figure 4.1: Concept of the Particle Filter for state prediction (a) No occlusion; and (b) Occlusion.

4.3 Foundation of proposed methods

4.3.1 Concept of proposed Particle Filter tracking

Let the state vector x_t describe the tracked object parameters and the vector z_t denote all the observations $z_1, ..., z_t$ up to time t. Baye's estimator or rule, can be used to estimate the current state x_t given all the data available up to and including z_t as

$$p(x_t|z_t) = \frac{p(z_t|x_t)p(x_t|z_{t-1})}{p(z_t|z_{t-1})}. \tag{4.1}$$

Fig. 4.1 shows the conceptualization of standard Particle Filter behaviour with and without occlusion. When there is no occlusion, the particle weights are updated with respect to the observation z_{t-1} known from the last frame to estimate the state vector x_t in the next frame. In occlusion, the last known observation z_t is used by the general Particle Filter to estimate the state vector, i.e., $x_{t+1}, x_{t+2}, x_{t+n-1}$ for all the upcoming frames till an observation z_{t+n} becomes available.

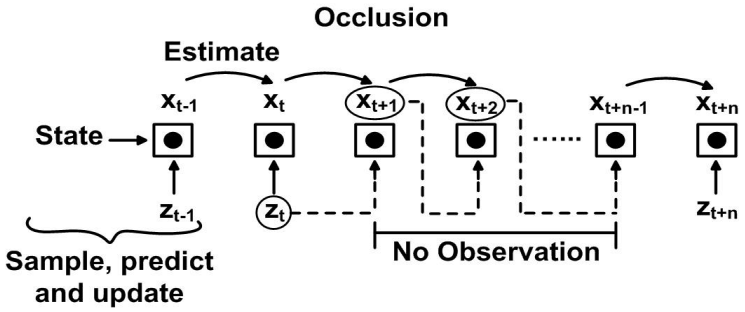


Figure 4.2: Concept of proposed Particle Filter for state prediction in occlusion.

If the occlusion is for a small number of frames, then the state predicted using the last observation is quite close to the ground truth. However, if the occlusion continues for significant number of frames, then the predicted state diverges from the ground truth. A Particle Filter adjusts the weights of the particles based on the most current observation to predict the next state. Hence, the lack of current observation is a clear reason for error in estimation of the state for frames at time $t+1, t+2, t+n-1$. This can be seen using the qualitative results in Section 4.5 on SBP tracking.

In stochastic dynamics a somewhat general assumption is made for the probabilistic framework that the object dynamics form a temporal Markov chain so that

$$p(x_t|X_{t-1}) = p(x_t|x_{t-1}). \tag{4.2}$$

This means that the new state is conditioned directly only on the immediately preceding state independent of the earlier history.

In Fig. 4.2 a new Particle Filter strategy or concept is illustrated to estimate the state in occlusion. During occlusion the last known observation z_t is only used to estimate the state x_{t+1} at time $t+1$. This state x_{t+1} is used as an observation to generate the next subsequent state x_{t+2}. Similarly, the state x_{t+2} is used as an observation to next state x_{t+n-1} and so on until an observation z_{t+n} is obtained. This strategy works because the most recent state is used as an observation to generate the next state in all time frames during occlusion. The proposed Particle Filter algorithm is described in Algorithm. 4.3.1 (see Section 4.4.1 for details).

Algorithm 4.3.1: PROPOSED PARTICLE FILTER ALGORITHM(x, z, s, π)

Construct a new weighted particle set $S = \{(s_t^{(n)}, \pi_t^{(n)})\}_{n=1}^N$ for time t from the old weighted particle set $S = \{(s_{t-1}^{(n)}, \pi_{t-1}^{(n)})\}_{n=1}^N$ at time $t-1$.
Select N particles from the set $S = \{(s_{t-1}^{(n)}, \pi_{t-1}^{(n)})\}_{n=1}^N$ to give $S = \{(s_{t-1}^{'(n)}, 1/N)\}_{n=1}^N$.
Predict each particle using the dynamic model $p(x_t|x_{t-1}) = s_{t-1}^{'(n)}$ to give $\{(s_{t-1}^{'(n)}, 1/N)\}_{n=1}^N$.
No Occlusion:
Measure and weight the particles as $\pi_t^{(n)} \propto p(z_t|x_t = s_t^{'(n)})$ to give $S = \{(s_t^{(n)}, \pi_t^{(n)})\}_{n=1}^N$. Normalize $\pi_t^{(n)}$ so that $\sum_n^N \pi_t^{(n)} = 1$.
Estimate the tracking result for time t as $E[x_t] = \sum_{n=1}^N \pi_t^{(n)} s_t^{(n)}$.
Occlusion:
For **non-consecutive** occlusion, use the last known measurement $S = \{(s_t^{(n)}, \pi_t^{(n)})\}_{n=1}^N$ to estimate the tracking for next time step.
For **consecutive** occlusion, use the last estimation $E[x_t] = \sum_{n=1}^N \pi_t^{(n)} s_t^{(n)}$ as measurement $S = \{(s_t^{(n)}, \pi_t^{(n)})\}_{n=1}^N$ for estimation in next time step.

4.3.2 Concept of Motion Flow (MFL) tracking

The direction of the instantaneous angular velocity (which is measured over an extremely small time interval [90]) is the basis for motion flow prediction. Consider the human arm as a pendulum attached at the shoulder joint producing curvilinear motion (incurring an angular displacement) as shown in Fig. 4.3. As the pendulum (arm) swings from its equilibrium position (vertical) to its maximum displacement, the magnitude and direction of angular velocity vector change. Two geometric constraints are proposed for predicting arm location based on pendulum motion. For an extremely small time interval in consecutive time frames:

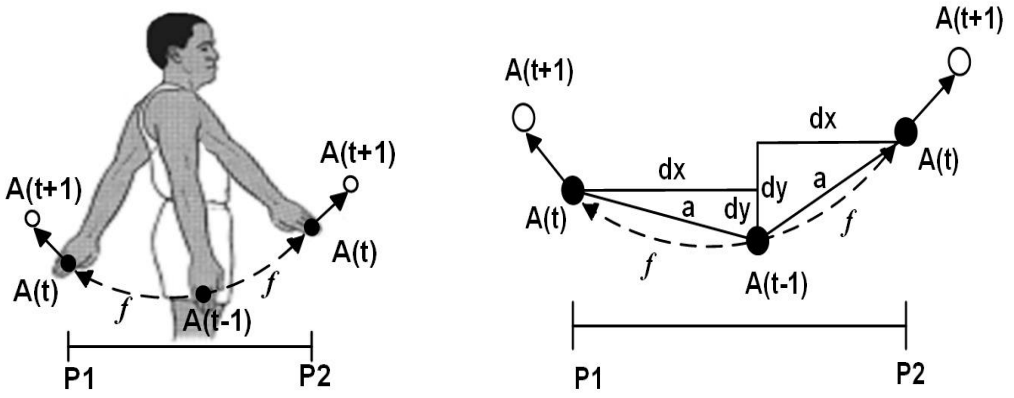
Conjecture 1:

The direction of the instantaneous angular velocity must be the same until the arm reaches its maximum displacement.

Conjecture 2:

A large instantaneous angular displacement shows that the arm has passed its maximum displacement.

Based on first conjecture the point to be predicted $A(t+1)$ should be close to the last arm point and continue in the direction of the previous two arm points, i.e.,



(a) (b)

Figure 4.3: Motion flow based arm prediction A using previous arm Ap and current arm Ac during occlusion (see Section 4.3.2).

follows the swing of arm for cyclic activities, as shown in Fig. 4.3 (a). The second conjecture leads to identify the change in direction of arm swing.

Consider the arm motion as a pendulum swing which draws a small dotted curve f in each frame as shown in Fig. 4.3 (b). Denote (x^A_{t-1}, y^A_{t-1}) and (x^A_t, y^A_t), respectively, as coordinates of labelled arm points in the previous and current frames. For every frame, the linear displacement between the current and previous arm points is

$$dx = x^A_t - x^A_{t-1} \ , \ dy = y^A_t - y^A_{t-1}. \tag{4.3}$$

The length L of the entire curve f (i.e., angular displacement) traced by the arm movement on the interval [P1-P2] can be approximated as a summation of all the line segments of the entire piecewise linear curve. The a^{th} line segment is the hypotenuse of a triangle with base dx and height dy, and has length

$$L_a = \sqrt{(x^A_t - x^A_{t-1})^2 + (y^A_t - y^A_{t-1})^2}. \tag{4.4}$$

By the Mean Value Theorem, there exists $x^* \in [x^A_{t-1}, x^A_t]$ such that

$$\frac{y^A_t - y^A_{t-1}}{x^A_t - x^A_{t-1}} = f^{'}(x^*). \tag{4.5}$$

$$y^A_t - y^A_{t-1} = f^{'}(x^*)dx \tag{4.6}$$

Substituting (Eq. 4.6) in (Eq. 4.4) gives

$$L_a = \sqrt{1 + [f'(x^*)]^2}dx. \tag{4.7}$$

Finally, the length of the entire polygon path with k subintervals is

$$\sum_{a=1}^{k} L_a = \sum_{a=1}^{k} \sqrt{1 + [f'(x^*)]^2}dx \tag{4.8}$$

which has the form of Riemann sum, i.e.,

$$L = \lim_{\Lambda \to 0} \sum_{a=1}^{k} \sqrt{1 + [f'(x^*)]^2}dx = \int_a^k \sqrt{1 + [f'(x)]^2}dx. \tag{4.9}$$

Increasing the number of subintervals or line segments of a piecewise linear curve such that $\Lambda = \max(dx) \to 0$ in (Eq. 4.9) proves the approximation that the length of polygon line segments is equal to the length of the curve, i.e., $\sum_{a=1}^{k} L_a \to L$. This mathematical proof and above-mentioned conjectures lead to the proposed motion flow based prediction (see Section 4.4.2) of arm points.

4.4 Overview of proposed SBP tracking

Depending on the user's choice, the proposed Particle Filter with memory and feedback (PFMF) or the motion flow based prediction is used for tracking SBPs in occlusion. The ability of PFMF to track any SBPs without any prior information of activity make it the default choice for SBPs prediction. The arm is the most occluded SBP, hence, motion flow method is designed to track arm SBP in cyclic activities.

4.4.1 Particle Filter with memory and feedback for SBP prediction

Particle Filter for visual tracking requires updating a confidence interval by calculating probability with respect to the newly available information, i.e., observation (measurement) to predict the state vector at next time step. This confidence decreases when there is no measurement available, i.e., the target being tracked is occluded (no target measurement in the image). It becomes lower as the number of frames without an observation increases once the target state is lost or occluded. The

standard Particle Filter will fail under this circumstance. A typical way of avoiding this problem is to restart the tracking algorithm, however, it might not be the best solution. Several researchers have proposed solutions to deal with occlusions without restarting the tracking algorithm [112], [111]. In [112], an adaptive Particle Filter is presented that uses a Rayleigh probability distribution during occlusion, while the memory-based Particle Filter in [111] combines the standard Particle Filter with a memory strategy to handle occlusions.

The proposed Particle Filter has two tracking (operation) modes, i.e., no occlusion and occlusion as shown in Fig. 4.4. In the no occlusion mode, the proposed Particle Filter behaves similar to the standard Particle Filter when the SBP is not occluded or missed by the SBP labelling framework (see Section 3.4.3). In the occlusion mode, the proposed Particle Filter uses a memory based feedback scheme when the SBP is occluded or missed by the SBP labelling framework. The proposed Particle Filter is able to track and predict SBPs in the presence or absence of occlusion, or missed SBPs, as shown in Fig. 4.4.

Given the current observation location, i.e, $P(t-1) = (x_{cv}, y_{cv})$, of a SBP at time step $t-1$, the Particle Filter predicts the location $P(t) = (x'_{cv}, y'_{cv})$ of a SBP at time step t. The state vector $X_{t-1} = [P^T \ V^T]$, where P is the position of a convex point at time $t-1$ and $V = P(t-2) - P(t-1)$. A constant-acceleration dynamic model X_t is used to update the state vector, where

$$X_t = MX_{t-1} = M[P^T \ V^T] \tag{4.10}$$

$$\text{M}= \begin{bmatrix} 1 & 0 & dt & 0 \\ 0 & 1 & 0 & dt \\ 0 & 0 & dt & 0 \\ 0 & 0 & 0 & dt \end{bmatrix} \tag{4.11}$$

where dt is the time lapse between two frames. The confidence interval is updated using the new weights, i.e.,

$$\pi_t^{(n)} = \exp[-0.05\sqrt{\sum_{n=1}^{N}(x_{cv} - s_t^{(n)})^2 + (y_{cv} - s_t^{(n)})^2}]. \tag{4.12}$$

where $(x_{cv} - s_t^{(n)})$ and $(y_{cv} - s_t^{(n)})$ represent the distance of N particles $s_t^{(n)}$ from the observations x_{cv} and y_{cv} respectively.

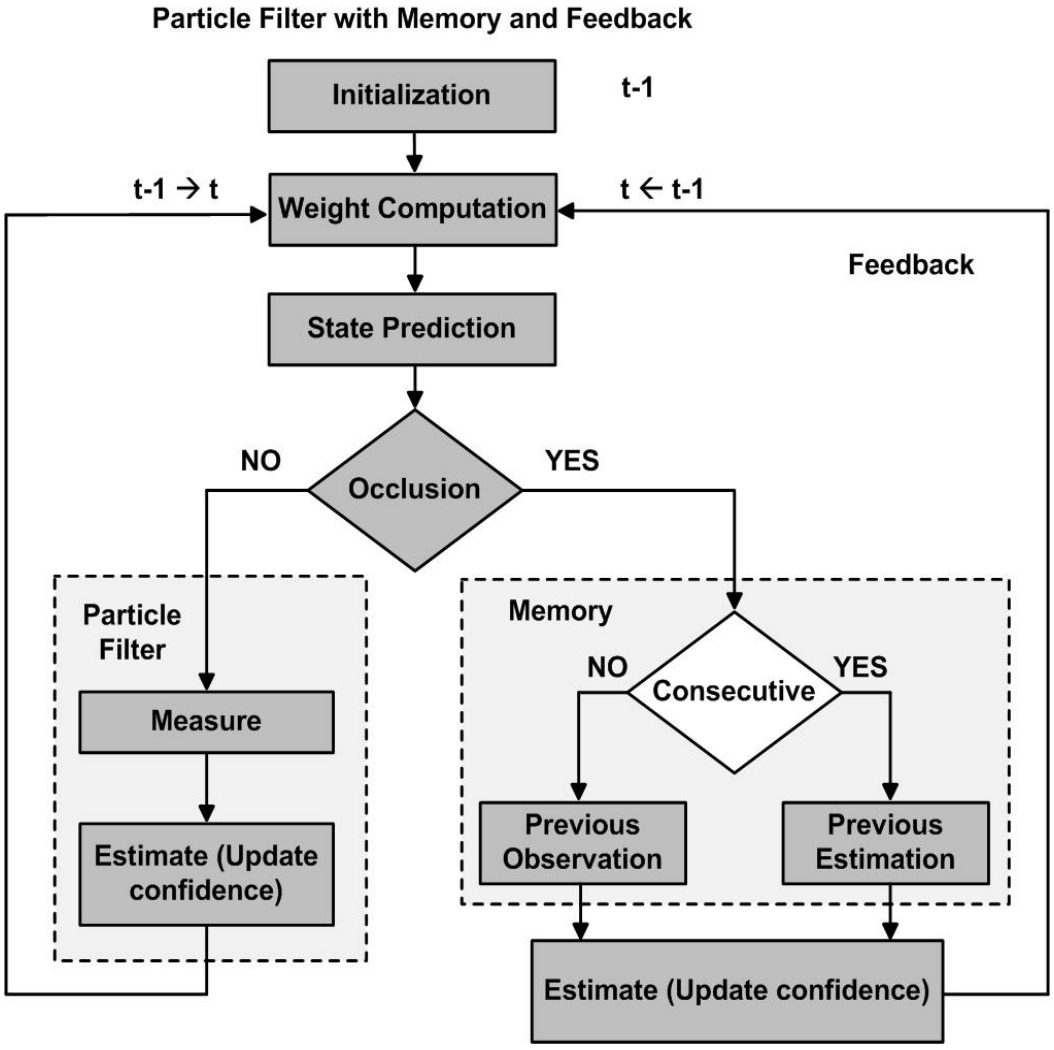


Figure 4.4: Block diagram describing the tracking (operation) modes, i.e., no occlusion and occlusion of the proposed Particle Filter with memory and feedback.

For each SBP, a Particle Filter with 100 particles is instantiated for optimum accuracy of prediction with particles ≥ 30 producing good results. During occlusion, the Particle Filter is initialized with the last known observation to predict the next SBP (x'_{cv}, y'_{cv}). This is achieved by keeping the temporal information of every previous measurement and observation. In the event of occlusion in consecutive frames, the predicted values in the first frame $P(t)$ and V are fed back as observations to initialize the Particle Filter for the subsequent frames.

The proposed Particle Filter with memory and feedback differs from the memory-based Particle Filter in [111] in the following aspects:

1. The memory consists of both past estimations and observations (measurements) while in memory-based Particle Filter only the past estimations are stored.

2. Only the two most recent past estimations and observations, i.e., at time $t-2$ and $t-1$ are stored in the memory, and are used to predict the state at time t, while in the memory-based Particle Filter past estimations include a complete history till the current time t.

3. The occlusion condition is established based on the input from SBPs' labelling framework in Section 3.4.3, instead of a predefined threshold.

4. A combination of memory and feedback is used to predict the state vector in occlusion.

4.4.2 Motion flow for SBP prediction

Motion flow employs the direction of linear displacement, prior knowledge of the activity, temporal information of a SBP, and geometry of the human body to define criteria for locating, labelling and tracking SBPs, i.e., arm points (x^A, y^A) during occlusion, as detailed in Table 4.1. If the displacement dx between current arm x_t^A and previous arm x_{t-1}^A point is greater than a threshold $\zeta = D_{seg}/6 = 5$ (where D_{seg}=30, see Section 3.4.1.4), it suggests that the maximum displacement is reached and direction of the arm swing arm has changed. Only dx is used because the horizontal displacement of arm (pendulum) from equilibrium position to maximum displacement is intuitively more than vertical displacement. The direction of the front arm movement is constrained based on the previously labelled front arm points. The criteria in Table 4.1 are used to predict front and back arm points during Walk,

Table 4.1: Parameters and their value for Motion flow based arm prediction.

Activity	$\|dx\|$	x_t^A	y_t^A	x^A	y^A
$Walk$	$-, < \zeta$	$\leq x_{t-1}^A$	$\geq y_{t-1}^A$	$x_t^A \mp dx$	$y_t^A + dy/0.4\zeta$
$Walk$	$> \zeta$	$-$	$-$	$x_t^A - 0.4\zeta$	$y_t^A + dy/0.4\zeta$
Run	$< \zeta$	$\leq x_{t-1}^A$	$\geq y_{t-1}^A$	$x_t^A \mp dx$	$y_t^A + dy/0.4\zeta$
Run	$-, \geq \zeta$	$-$	$-$	$x_t^A \mp 0.8\zeta$	$y_t^A + dy/0.4\zeta$
$Skip$	$\leq \zeta$	$\leq x_{t-1}^A$	$-$	$x_t^A \mp dx/0.4\zeta$	y_t^A
$Skip$	$-$	$-$	$-$	$x^H \pm 1.4\zeta$	$y^H + 4\zeta$
$Side$	$< \zeta$	$\leq x_{t-1}^A$	$-$	$x_t^A \mp dx$	y_t^A
$Side$	$> \zeta$	$-$	$-$	$x_t^A \mp dx/\zeta$	y_t^A
$Jump$	$< \zeta$	$\leq x_{t-1}^A$	$-$	$x_t^A \mp dx$	y_t^A
$Jump$	$> \zeta$	$-$	$-$	$x_t^A \mp dx/\zeta$	y_t^A
$PauseJump$	$< \zeta$	$-$	$\leq y_{t-1}^A$	x_t^A	$y_t^A + dy$
$PauseJump$	$> \zeta$	$-$	$-$	x_t^A	y_t^A

Side, Jump-in-place-on-Two-Legs , Jump Left to Right, Run Right to Left, and Skip on the Weizmann data set.

In Table 4.1, x^H and y^H, and x^A and y^A, respectively denote the coordinates of the head and predicted arm points. The upper polarity is used for Right to Left, and the lower polarity is used for Left to Right. Front arm and Back arm are distinguished respectively on Right side and Left side based on (Eq. 3.11). For all actions the arm point is predicted at the centre (x_c,y_c) when no conditions are satisfied or when more than three points have been predicted consecutively. In the first row of Walk, Side, Skip, Pause Jump , and Run in Table 4.1, the relational operator and polarity of criteria for current arm (x_t^A, y_t^A) and predicted arm (x^A, y^A) are respectively reversed for front and back arm prediction in Right to Left and Left to Right. The second row of these actions is used to predict back points when they are not predicted by the first row. For Walk, dx is not used for front arm point prediction (which is denoted by a dash) but is used to predict back arm point only. For Jump, front arm point is predicted at centre (x_c, y_c) in occlusion, while the back arm point is predicted using the two rows of jump. However, if $dx > 2\zeta$ pixels then

back arm point is predicted at the centre.

4.5 Experimental Results

4.5.1 Qualitative Evaluation

The first evaluation is performed to establish whether the proposed Particle Filter with memory and feedback performs better than the standard Particle Filter. The Fig. 4.5 visually presents the arm significant body points predicted using the standard Particle Filter on the cyclic activities of the Weizmann data set. It can be seen that the standard Particle Filter is unable to predict arm point (in blue circle) at the accurate positions. The large distance of the predicted arm point from the human body also suggests a decrease in the confidence of the standard Particle Filter due to lack of observation. In contrast, it can been observed from the results in Fig. 4.6 that the proposed Particle Filter with memory and feedback is more accurate than the standard Particle Filter algorithm. This qualitative evaluation is sufficient to prove the superior performance of the proposed Particle Filter with memory and feedback. The performance of standard Particle Filter is poor for tracking in occlusion and does not require quantitative comparison.

In the second evaluation, the reliability of predicted significant body point using the Particle Filter with memory and feedback is compared with the motion flow method. Fig. 4.6 and Fig. 4.7 show the detailed results of significant body point labelling and tracking using the Particle Filter with memory and feedback instantiated only for arm points in all ten activities of the Weizmann data set. Significant body points labelling and tracking using motion flow based prediction for arm points is shown in Fig. 4.8 and Fig. 4.9. The predicted arm significant body points are shown with a light green circle. It can be observed from these results that the predicted arm point using the Particle Filter with memory and feedback and Motion flow are accurately identified and tracked. In Fig. 4.7 and Fig. 4.9, the overlapping bold head (H) and arm (A) point are the reallocated significant body points using the Smart Search Algorithm in Section 3.4.3.4. Section 4.5.2 contains the complete quantitative results on both Weizmann and MuHAVi dataset.

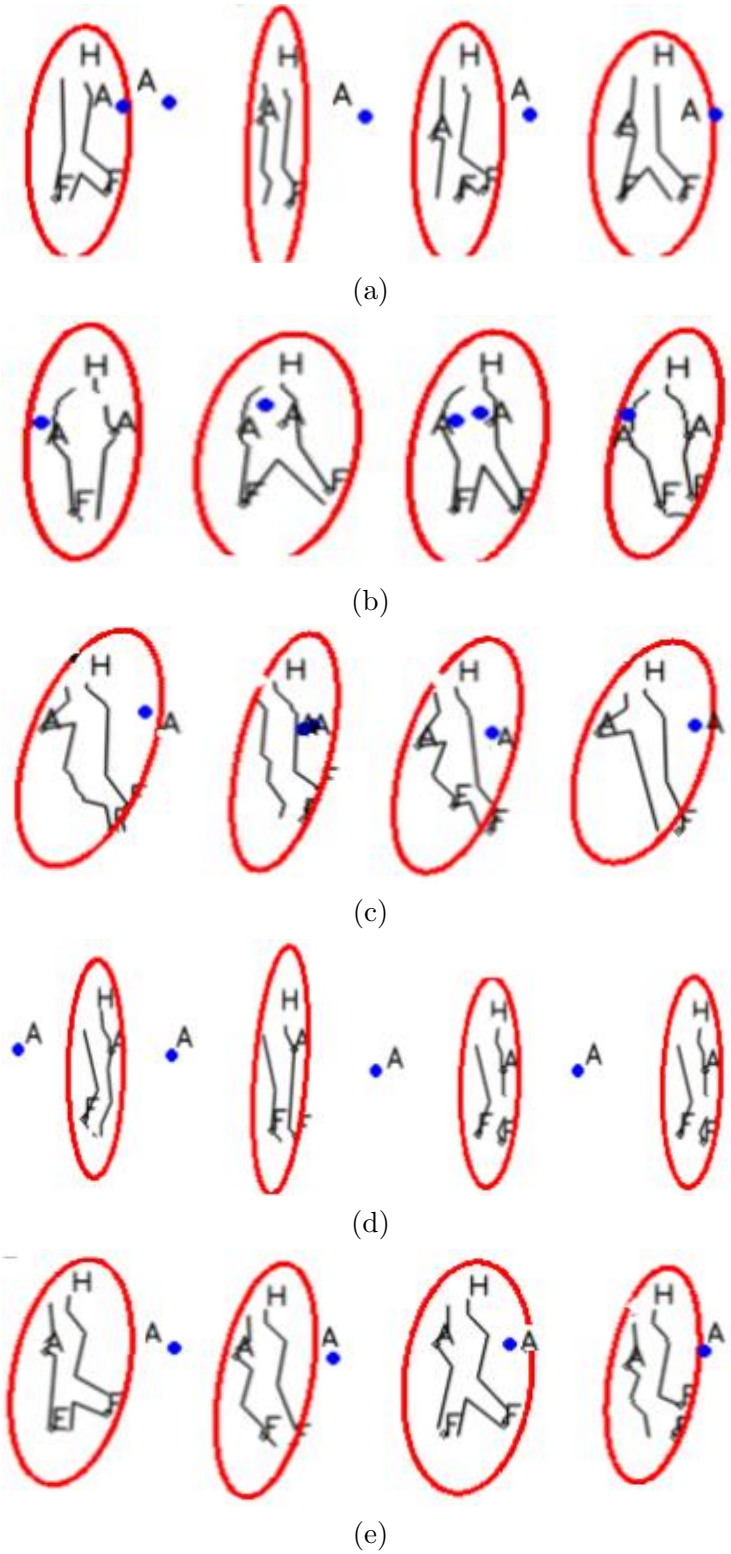


Figure 4.5: Arm SBP predicted using the standard Particle Filter. The predicted arm is shown in blue circle for (a) Walk, (b) Side, (c) Skip, (d) Jump and (e) Run activities.

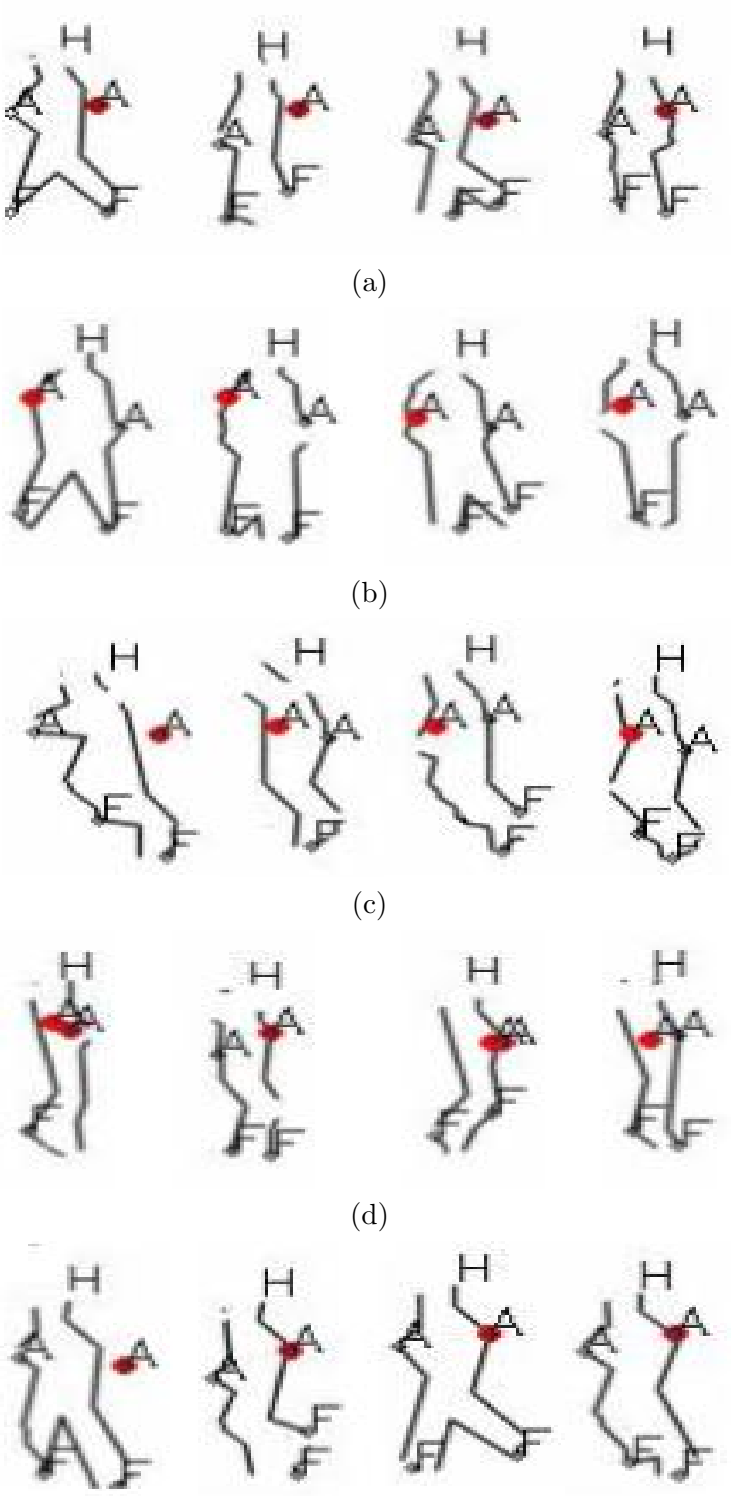


Figure 4.6: Arm SBP tracking using the Particle Filter with memory and feedback shown in red circle (a) Walk, (b) Side, (c) Skip, (d) Jump and (e) Run.

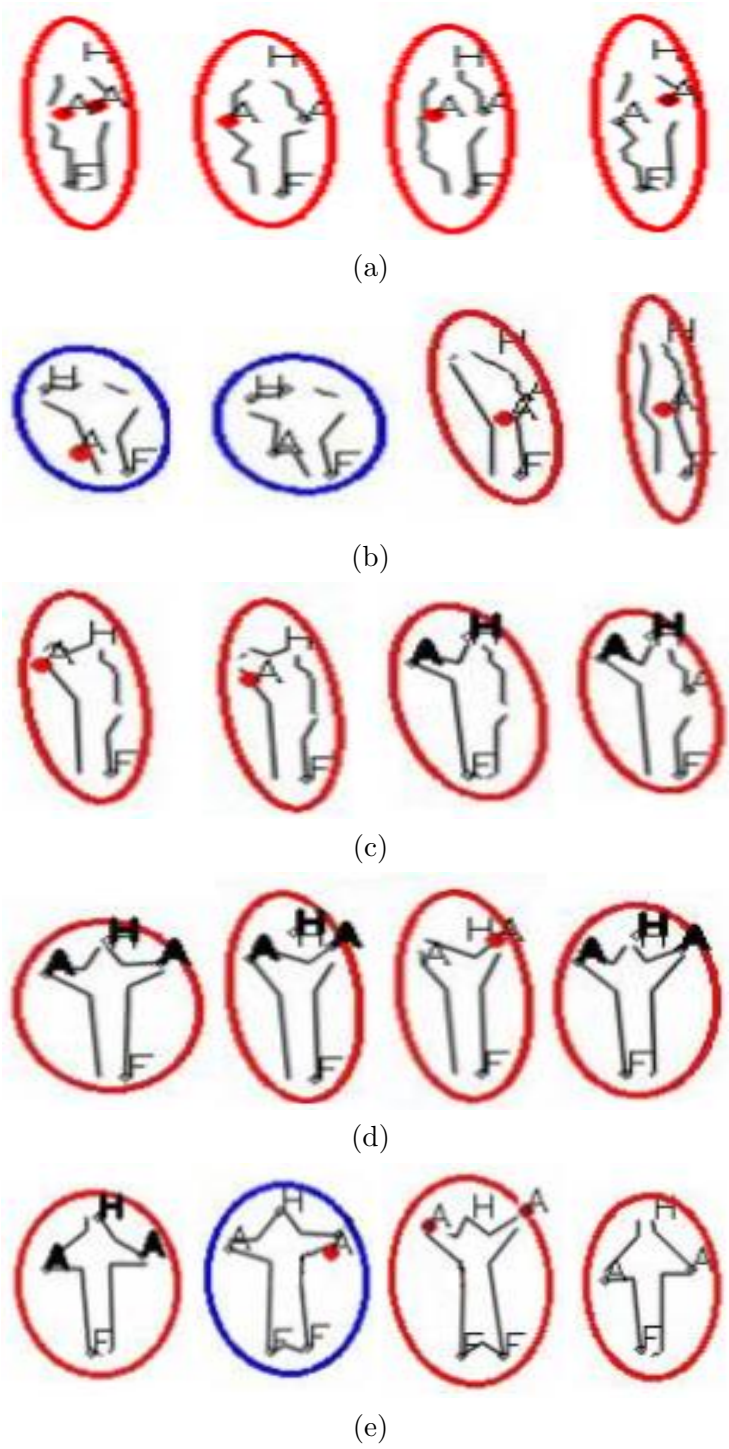


Figure 4.7: SBP tracking using the Particle Filter with memory and feedback shown in red circle (a) Jump-in-place-on-Two-Legs, (b) Bend, (c) One hand wave, (d) Two hand wave and (e) Jack. The reallocated Head (H) and Arm (A) points using Smart Search Algorithm in Section 3.4.3.4 are superimposed in black bold.

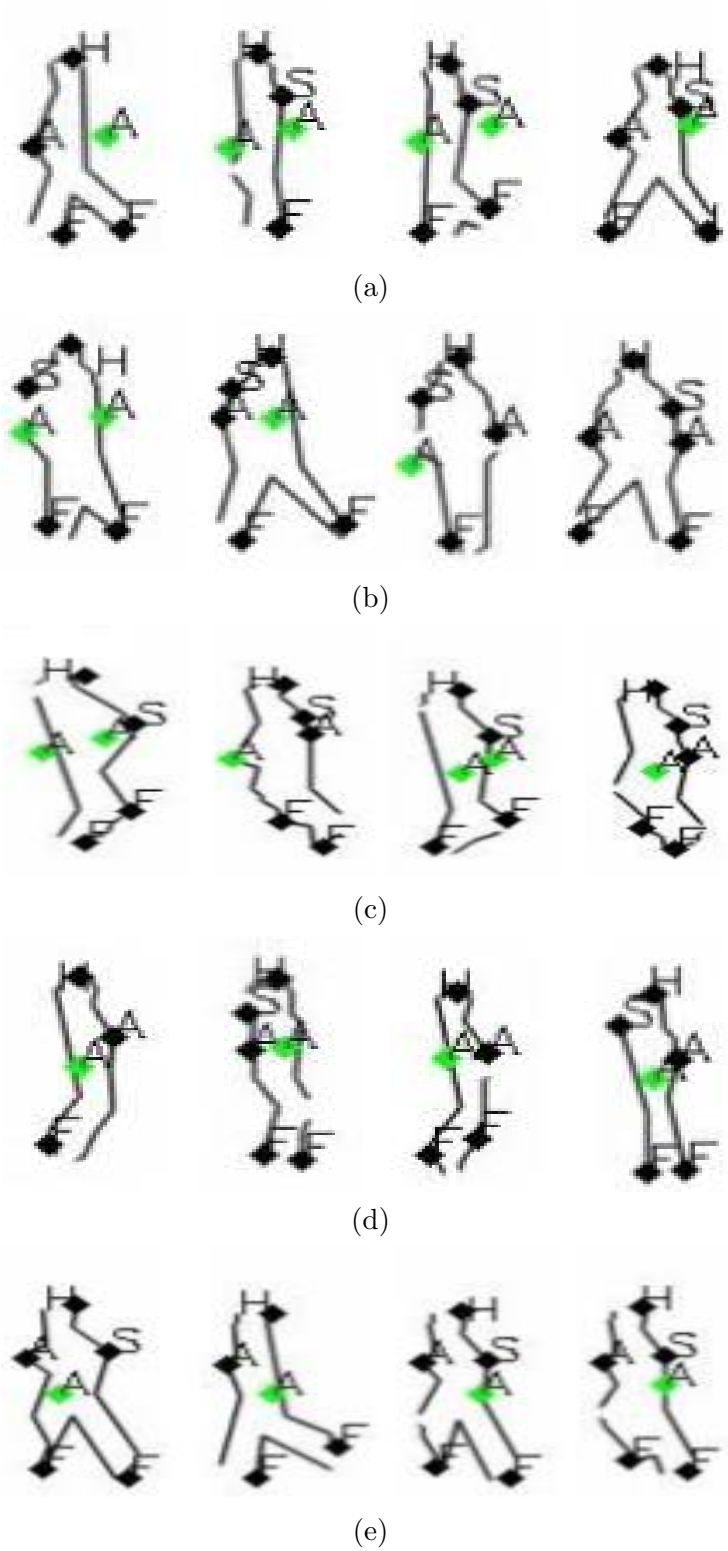


Figure 4.8: SBP tracking using the motion flow prediction shown in green circle (a) Walk, (b) Side, (c) Skip, (d) Jump and (e) Run.

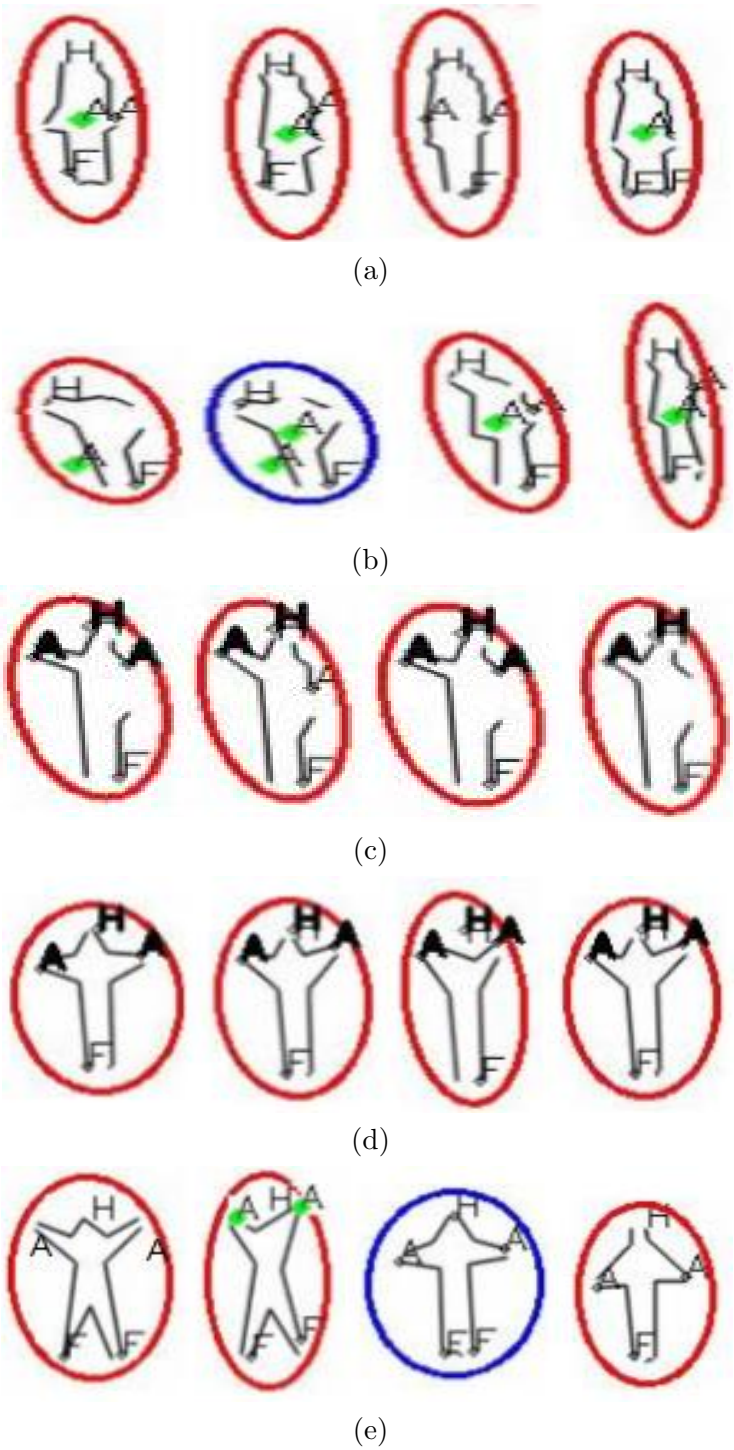


Figure 4.9: SBP tracking using the motion flow prediction shown in green circle (a) Jump-in-place-on-Two-Legs, (b) Bend, (c) One hand wave, (d) Two hand wave and (e) Jack. The reallocated Head (H) and Arm (A) points using Smart Search Algorithm in Section 3.4.3.4 are superimposed in black bold.

Table 4.2: Particle Filter with memory and feedback (denoted by p), and Motion flow (denoted by m) prediction error in pixels unit. Mean height is 68 and 200 pixels for Weizmann and MuHAVi data set respectively.

Activity	x_p^{FA}	y_p^{FA}	x_m^{FA}	y_m^{FA}	x_p^{BA}	y_p^{BA}	x_m^{BA}	y_m^{BA}
Walk	7.7	12.9	**4.2**	**3.3**	9.23	19.4	**3.4**	6.4
Run	7.5	8.1	8.3	**3.3**	9.9	15.4	6.8	8.4
Skip	8.5	9.4	4.8	6.3	13	9.2	4.1	**5.7**
Side	5.4	**8**	6.1	5	3.5	11	5	6.6
Jump	8.2	14.2	4.1	6.2	6.9	**8.5**	5	6.5
PauseJump	**4.4**	12.2	7	6.1	**2.9**	10	4.5	6
Average	6.9	10.8	**5.8**	**5**	7.1	12.2	**4.8**	**6.6**
Average/Mean Height	0.1	0.15	**0.08**	**0.07**	0.1	0.18	**0.07**	**0.09**

4.5.2 Quantitative Evaluation

The best results in the tables in this section are represented in bold.

4.5.2.1 Localization accuracy of predicted arm SBP

The Particle Filter with memory and feedback, and the motion flow method are compared for arm prediction on only cyclic activities of both data sets because they are the most occluded SBP. It is vital to verify the accuracy of location of predicted arm SBP versus the ground truth. Table 4.2 shows the error in the location using Particle Filter and motion flow in occlusion, where the average location error of predicted SBP in pixel units is

$$ErrorPred(x_{avg}, y_{avg}) = \frac{\sum_{n=1}^{N} |G_n(x,y) - Pred_n(x,y)|}{N} \tag{4.13}$$

and $Pred_n(x,y)$ are the predicted SBP coordinates.

The Particle Filter with memory and feedback, and motion flow method are compared for the arm prediction in cyclic activities (see Table 4.2), i.e., Walk ($\alpha 1$), Run ($\alpha 2$), Skip ($\alpha 3$), Side ($\alpha 4$), Jump ($\alpha 5$) and Jump-in-place-on-Two-Legs/Pause Jump ($\beta 7$) of both data sets because it is the most occluded SBP. Table 4.2 shows that the Particle Filter and motion flow accurately predict arm point, i.e., close to

ground truth location. The y-coordinate error of the front and back arm points using motion flow prediction are consistently smaller than those obtained using Particle Filter. The x-coordinate error is also smaller in most activities. Hence, the motion flow outperforms Particle Filter which is demonstrated by smaller average errors in all activities in Table 4.2. However, the lack of necessity for prior information makes the Particle Filter a better choice for prediction. Results on Walk ($\alpha 1$) and Run ($\alpha 2$) activities of both data sets are collectively shown in Table 4.2.

The average error in pixels as a proportion of the mean height of subjects for all the activities is shown in the last row of Table 4.2. This can be used to have a picture of how much an error, e.g., 5 pixels, means with respect to the size of the human body. For example, the human head is one-eighth the human height, i.e., 0.125. Hence, a 5 pixel error equates to 0.07 that is almost half of the height of the human head.

4.5.2.2 Accuracy of detected SBPs with prediction vs observed

In case e.g., occlusion, when the SBP detection method in Chapter 3 cannot identify a convex point to be labelled as SBP then the tracking method is used to predict the position of a SBP. This can help SBP detection by using prediction from the tracking method. Hence, improving the number of correctly detected SBPs. Table 4.3 presents the accuracy of SBP detection with prediction and compares it with no prediction (see Chapter 3). For SBP detection with prediction in Table 4.3, observed (OB) SBPs is the manually counted visible SBP with guessed SBPs. (Eq. 3.36) is used to compute PR, RC and ER. In Table 4.3, for prediction, an overall 2.5% and 2.4% percentage increase in recall and precision, respectively, are obtained in cyclic actions of the Weizmann data set using the proposed Particle Filter with memory and feedback prediction. Specifically, the highest percentage increase of 7.3% in recall is achieved in Run ($\alpha 2$), which has the smallest recall with no prediction. For the MuHAVi data set, the Particle Filter with memory and feedback prediction is only used for Walk ($\alpha 1$) and Run ($\alpha 2$) since they are cyclic actions. A percentage increase of 10.7% in recall is attained in Run ($\alpha 2$). There is a decrease in precision for both Walk ($\alpha 1$) and Run ($\alpha 2$), which suggests an increase in misclassified arm SBPs. However, more importantly the Particle Filter with memory and feedback prediction enhances the recall in all cyclic actions of both data sets. The proposed framework with prediction obtains an overall average % recall and precision of 97.7% and 98.8%, respectively, for all activities of the Weizmann data set. It achieves an

Table 4.3: Precision and Recall of five SBPs detection of proposed framework.

	Weizmann Data set									
	No prediction			Prediction			No prediction		Prediction	
Activity	CT	OB	DT	CT	OB	DT	RC%	PR%	RC%	PR%
Walk	2655	2768	2681	3134	3195	3160	95.9	99	**98.1**	99.2
Run	1468	1623	1532	1828	1885	1892	90.4	95.8	**97**	96.6
Skip	1566	1664	1585	2108	2170	2127	94.1	98.8	**97.1**	99.1
Side	1726	1786	1726	2183	2220	2183	96.6	**100**	98.3	**100**
Jump	1756	1877	1759	2220	2290	2223	93.5	99.8	**97**	99.9
PauseJump	2231	2271	2286	2654	2690	2709	98.2	97.6	98.7	98
Bend	3067	3195	3278	-	-	-	96	93.6	-	-
OneHandWave	3265	3265	3555	-	-	-	**100**	91.8	-	-
TwoHandWave	2875	3120	3018	-	-	-	92.1	95.3	-	-
Jack	3157	3370	3201	-	-	-	93.7	98.6	-	-
Average %	-	-	-	-	-	-	95.3	96.5	**97.7**	**98.8**
	MuHAVi Data set									
Walk	1188	1231	1191	1326	1351	1502	96.2	99.8	**98.1**	88
Run	975	1198	985	1080	1198	1160	81.4	99	**90.1**	93.1
Turn	868	1046	868	-	-	-	83	**100**	-	-
Standup	1431	1471	1505	-	-	-	97.4	95	-	-
Collapse	1131	1306	1152	-	-	-	86.6	98.1	-	-
Kick	828	922	865	-	-	-	89.8	95.7	-	-
Punch	729	757	739	-	-	-	96.3	98.6	-	-
$Guard-to-Kick$	503	512	507	-	-	-	98.2	99.2	-	-
$Guard-to-Punch$	529	533	529	-	-	-	**99.2**	**100**	-	-
Average %	-	-	-	-	-	-	92.01	98.4	**94.2**	95.7

overall average % recall and precision of 94.2% and 95.7%, respectively, with prediction for all activities of the MuHAVi data set.

4.5.2.3 Comparative evaluation of SBP labelling and tracking

The performance of the proposed SBP Labelling (Section 3.4.3) and Particle Filter tracking framework is compared with state of the art approaches, i.e., First Sight (FS) [37] and CBHM [33], with a similar extent of occlusion and type of activity, respectively. The accuracy of First Sight to detect five body parts, i.e., Head, Arms, and Feet, is evaluated in terms of the parts observed by the human eye. Five SBPs identified by the proposed framework correspond to the five body parts detected by First Sight. The activities used by First Sight differ in terms of no, mild and severe self occlusion. In the data sets for this chapter, Walk (α1), Run (α2), Side (α4), Turn (α6), Jump-in-place-on-Two-Legs/Pause Jump (β7), Punch (β15), Guard-to-Kick (β16), and Guard-to-Punch (β17) have mild self occlusion, whereas Skip (α3), Jump (α5), Bend (β8), One hand wave (β9), Two hand wave (β10), Standup (β12), and Collapse (β13) have severe self occlusion. Table 4.4 shows the performances of the proposed framework and First Sight (as reported in [37]) on activities with mild and severe occlusion on all subjects of the Weizmann and MuHAVi data sets. In Table 4.4, results on Walk (α1) and Run (α2) activity of both data sets are presented collectively. The average % of the five SBPs error computed using (Eq. 3.36) is clearly much less than First Sight.

Due to unavailability of the data set used by CBHM, Table 4.4 compares the average precision and recall of the proposed framework in detecting four SBPs (i.e., hands and feet) in similar activities with those of CBHM as reported in [33]. It shows that the proposed framework obtains better recall and precision than CBHM in Run (α2), Jump (α5) and Collapse (β13). It also achieves a slightly better recall for Walk (α1). The recall obtained for Standup (β12) is close to this approach, thus, overall the proposed framework performs better than CBHM.

4.5.2.4 Comparative evaluation of Stick figure generation

The consistency of the stick figures generated from the SBPs detected by the proposed SBP labelling and tracking framework is compared with those generated using skeletonized (SKEL) [106] and Computer Vision based Human body Segmentation and Posture estimation (CVHSP) or Star skeletonization (STAR) [20, 21] by evaluating the total number of correctly detected five SBPs in video sequence of various

Table 4.4: SBP detection: Proposed vs CBHM and FS.

		4 SBPs Accuracy				5 SBPs Error		
Classification		CBHM [33]		Proposed		Proposed		FS [37]
Occlusion	Activity	RC%	PR%	RC%	PR%	ER%	Average%	Average%
Mild	*Walk*	95.2	100	**97.4**	99.2	0.6		
Mild	*Run*	76.8	90.8	**97**	**97**	2.59		
Mild	*Side*	-	-	98.1	**100**	**0**		
Mild	*Turn*	-	-	80.2	**100**	**0**		
Mild	*PauseJump*	-	-	98.3	97.5	2.4		
Mild	*Kick*	-	-	87.2	94.5	4.2		
Mild	*Punch*	-	-	95.5	98.3	1.35		
Mild	*Guard − to − Kick*	-	-	97.8	99	0.79		
Mild	*Guard − to − Punch*	-	-	**99.1**	**100**	**0**	**1.33**	15
Severe	*Jump*	88.5	70.4	**97**	**99.8**	**0.17**		
Severe	*Standup*	99.7	82.6	95.9	**94.4**	4.91		
Severe	*Collapse*	83.3	83	**85.7**	**97.6**	1.82		
Severe	*Bend*	-	-	97.6	92.2	6.43		
Severe	*OneHandWave*	-	-	**100**	89.6	8.15		
Severe	*TwoHandWave*	-	-	91	94	4.73		
Severe	*Jack*	-	-	92.1	98.3	1.37		
Severe	*Skip*	-	-	94.8	97.1	1.19	**3.59**	21

activities. CVHSP and STAR use the same distance curve method to locate convex points which serve as SBPs. The distance curve method in [20,21] is implemented to compare its SBP detection accuracy with the proposed framework. Table 4.5 shows that the proposed framework with prediction consistently obtains more SBPs than SKEL and CVHSP or STAR (denoted only by CVHSP in the table) across all activites except Two hand wave (β10) and Jack (β11) of Weizmann data set. The total number of SBPs detected by the proposed framework (8425) across all activities of MuHAVi data set is more than SKEL (8170) and CVHSP/STAR (7591), hence, it is more consistent in generating stick figures of various activities. It obtains SBPs consistently more than SKEL and CVHSP in most activities and competes well in the remaining activities.

Table 4.6 summarises the various components of the most related methods. It shows the ability of the methods to tackle various activities with respect to the number of cues, criteria, and pose estimation. It compares the tracking and occlusion handling capability of each method. It also shows whether the methods have provided quantitative analysis for justifying their robustness and whether they generate stick figures. It can be seen that the proposed framework deals with the more number of activities, has tracking and occlusion handling ability, determines the posture of the human body and generates automated Stick Figures, and provides quantitative evaluation of the SBP detection and tracking.

4.5.2.5 Computational complexity

The proposed framework runs in real time due to its computational simplicity. The computational time of the proposed framework implemented in Microsoft Visual Studio 2010 Express Edition environment with OpenCV 2.4.6 on an Intel (R) Core (TM) i7 processor working at 2.93 GHz with 4 GB RAM running Windows 7 operating system is measured using the computer system clock. The proposed framework labels SBPs in 0.031 seconds per image frame on the Weizmann data set at 20-30 frame per second. It labels SBPs in 0.071 seconds per image frame on the MuHAVi data set.

The convex hull is computed using the Sklansky's algorithm [105] which has a computational complexity of $O(N)$, where N in the number of convex points. The contour moments algorithm is based on the Green theorem [100] which has a computational complexity of $O(L)$, where L is the length of the boundary of the object. The performance of the Particle Filter enhances with the increase in

Table 4.5: SBP detection: Proposed vs SKEL and CVHSP.

Weizmann Data set			
Activity	SKEL [106]	CVHSP [20]	PROPOSED
Walk	2768	2379	**3134**
Run	1623	1323	**1828**
Skip	1664	1398	**2108**
Side	1626	1347	**2183**
Jump	1455	1244	**2220**
PauseJump	2271	1210	**2654**
Bend	2669	1609	**3067**
OneHandWave	2667	1782	**3265**
TwoHandWave	2987	2064	2875
Jack	3299	2835	3157
MuHAVi Data set			
Walk	1239	1209	**1326**
Run	1005	899	**1080**
Turn	901	778	868
Standup	1464	1394	1431
Collapse	1189	958	1131
Kick	770	711	**828**
Punch	672	695	**729**
$Guard-to-Kick$	467	404	**503**
$Guard-to-Punch$	463	543	**529**

Table 4.6: Proposed approach versus Related approaches.

Method	STAR [106]	FDMHP [38]	CBHM [33]	CVHSP [20]	Proposed
No. of Cues	2	3	4	6	4
Criteria	-	-	Heuristic	Heuristic	Anthropometry Kinesiology Biomechanics Human vision
Pose Estimation	No	No	No	Yes	Yes
Tracking	No	Yes	Yes	No	Yes
Occlusion	No	No	No	Partial	Full
No. of Activities	2	5	6	14	15
Quantitative result	No	No	Yes	No	Yes
Stick Figure	Yes	Yes	Yes	No	Yes

number of particles. It is formally $O(N \log N)$, however, it can be made $O(N)$ with minor modifications to the sampling procedure. In the proposed framework, the Particle Filter is initialized with 100 particles with a state vector constituting of four parameters. As a result its computational speed can be considered to be real time. This is similar to [10] where a 6-12 degree of freedom model with 100 particles run in real time.

4.6 Summary

In this chapter two methods for SBP tracking are presented, i.e., Particle Filter with memory and feedback, and motion flow. The former method does not require any knowledge of activity and performs better than the standard Particle Filter. The latter method is more accurate, however, requires prior knowledge of activity. The proposed Particle Filter with memory and feedback is combined with the SBP labelling framework which improves SBP identification during occlusion or missed SBP. The tracking method increases the SBP detection accuracy. Comparative results demonstrate better SBP detection performance versus state of the art approaches. In future, the proposed Particle Filter with memory and feedback can be extended to predict SBPs in more activities.

Chapter 5

Activity Recognition

5.1 Introduction

Human activity recognition is important due to potential applications in video surveillance, assisted living, animation etc [113] [114]. In general, a standard activity recognition framework consists of the feature extraction, feature selection (dimension reduction) and pattern classification. The feature extraction can be broadly categorized into the holistic (shape or optical flow) [12–14, 115], local feature (descriptors of local regions) [15–17, 116] and model-based (prior model) or model-free (no prior model) approaches. Techniques such as Principal component analysis (PCA) [117] or Linear Discriminant Analysis (LDA) [118] are commonly used to select the most prominent features. Decision tree (DT) [13] or Support Vector Machines (SVMs) [114] are used for efficient classification.

The current state-of-the-art human activity recognition method varies with respect to application scenario as each method has been designed and verified for data sets containing different challenges such as similar activities, industrial environment, illumination variation, varying clothing, complex backgrounds, multiple actors, person-to-person interaction, human object interaction, multiple views etc. (see [72] for details on datasets). Also, it has been noted in literature [119] that human activity recognition methods have different performances on different data sets. The apparent reason for this lies in the feature extraction approach, i.e., holistic, local feature and model-based/model-free, and the different characteristics of the activities in the data sets [119]. The local features approach that extract the neighbourhood information of the regions or interest points focus more on the lo-

cal motion than on the figure shape. Hence, it is suitable for activities with more intra-class dissimilarity in the shape of figures. In contrast, the holistic and model-based/model-free approach are focused on figure shape characteristics which makes them suitable for activities with more inter-class similarity in the local motion, i.e., similar activities such as Walk, Run etc.

Recognizing similar activities still remains a challenge (see Section 5.2). The local feature and holistic approaches are computationally expensive and require intensive training while the model-based/model-free approach is efficient but less accurate. Therefore, the robust and efficient implicit body model based approach for significant body point (SBP) detection described in Chapter 3 and Chapter 4 [120] is used for feature extraction. In this context, the work in [21] that extracts the leg frequency and torso inclination is extended to determine two more features, i.e., the leg power and torso power. Also, the SBP detection method is augmented to extract features (similar to [115]) that extract variations in the movement of different body parts at different directions, i.e., up, down, right, and left, during an activity. As in [115] PCA or LDA is not used as we extract less than 15 features. These features are used to create two feature descriptors.

For efficient classification, mostly researchers use off-the-shelve classifier such as SVM and DT but with a trade-off of performance, e.g., SVM struggles due to the lack of generalized information, i.e., each test activity is compared with the training activity of one subject [115]. On the other hand DT imposes hard constraint that leads to separation problems when the number of categories increases or when categories are similar, i.e., a lack of clear separation boundary [18]. To achieve high accuracy while being fast the Relaxed Hierarchy (RH) method in [18] uses relaxed constraint, i.e., postpone decisions on confusing classes, to tackle the increased number of categories but still remains prone to accurately discerning similar categories. The Hierarchical Strategy (HS) method in [121] uses the RH and group together easily confused classes to improve the classification performance. RH and HS have only been applied to the spatial domain. Hierarchical methods [122, 123] are also used at lower levels for feature-wise classification. Note, however, similar to [18] this work focuses on building high-level class hierarchies and look into the problem of class-wise partitioning.

In order to recognize similar human activities efficiently and accurately, we propose a hierarchical relaxed partitioning system (HRPS) (see Section 5.3 for details). This is a system that classifies and organizes activities in a hierarchical

manner according to their type, i.e., pure activities (easily separable) and impure activities (easily confused). Subsequently, it applies relaxed partitioning to all the easily confused activities by postponing the decisions on them until the last level of the hierarchy, where they are labelled by using a novel majority voting scheme (MVS). As opposed to a conventional multi-class classifier as in [121] that can distinguish between only two similar activities, i.e., two classes overlap simultaneously, the proposed MVS is able to discern between three or more similar activities, i.e., three classes overlap concurrently. Thus, making the HRPS more robust and suitable for identifying activities in real world scenarios.

The major contributions of this work are as follows: (a) extending [21], Chapter 3 and Chapter 4 to built two feature descriptors and (b) implementing HRPS with the majority voting scheme to recognize similar activities.

This Chapter, is organized as follows. Section 5.2 reviews related methods. Section 5.3 and Section 5.4 present the foundation of HRPS and its application to activity recognition, respectively. Experiments are shown in Section 5.5.

5.2 Literature review

Several human activity recognition methods, e.g., [13, 15, 16, 22, 23, 119, 124–126] verified on the benchmark data sets (see [72] for data sets) struggle in correctly classifying similar activities of the Weizmann data set. The methods [13,14,115,116] that are able to correctly classify similar activities of the Weizmann data set are either computationally expensive or require intensive training or need to learn a large set of features. These methods require tuning of parameters with respect to the data set. Therefore, they require extensive re-training for new activities. Some methods [14,15,23] require more number of frames (approximately 100 to 200 frames) for training, thus duplicate or up-sample the training data.

5.2.1 Holistic and local feature approaches

In [13], a binary prototype tree based on shape and motion feature is learned, and a lookup table is used to match actions. Both shape and motion cues are required to recognise similar activities accurately. In [14], the clusters of motion curves from the optical flow of probe video sequences are matched with the clusters of training video sequences. In [115] the optical flow and random sample consensus methods are used to localize the subject. Next, it extracts a feature vector that contain variations

in the movement of different body parts at different directions during an activity. Euclidean distance or SVM is used with the feature vector for action recognition. In [116] the locality preserving projection method (that learns a projection onto a low dimensional space while optimally preserving the neighbourhood structure) is supervised to recognize similar activities by not ignoring the local information of the data. These methods are either computationally expensive or require intensive training or tuning of multiple parameter on a data set.

In [15], the kinematic features from the optical flow extracted from videos are converted into kinematic modes using principal component analysis. These kinematic modes are then used in a bag of kinematic mode representation with a nearest neighbour classifier for human action recognition. It has high computational cost, requires intensive training and confuses similar activities. In [16], videos are represented as word $\times$ time tables and the extracted temporal patterns are used with supervised time-sensitive topic models for action recognition. It also confuses similar activities.

5.2.2 Model-free and model-based approaches

A star is a shape that is formed by connecting the centre of mass of a human silhouette contour to the extreme boundary points. The method in [21] creates a one-star by using a local maximum on the distance curve of the human contour to locate the SBPs which are at the extremities. It uses two motion features, i.e., leg frequencies and torso angles, to recognize only the Walk and Run activities. A two star method [34] extends [21] by adding the highest contour point as the second star. It uses a 5D feature descriptor with a hidden Markov model (HMM) to detect the fence climbing activity. The method in [22] extends [34] by using the medial axis [106] to generate the junction points from which variable star models are constructed. It is compared with [21] and [34] on the fence climbing activity, and evaluated on the Weizmann data set. In [20], multiple cues such as the skin colour, principal and minor axes of the human body, the relative distances between convex points, convex point curvature, etc., are used to enhance the method in [21] for the task of posture estimation. It does not provide quantitative results, and uses a non-standard and non-publicly available data set. Thus, it requires extensive further work to validate and apply it to activity recognition. The method in [23] assumes that SBPs are given and uses the chaotic invariant for activity recognition on the Weizmann data set. It uses the trajectories of SBPs to reconstruct a phase

space, and applies the properties of this phase space such as the Lyapunov exponent, correlation integral and dimension, to construct a feature vector, for activity recognition. The above-described distance curve based methods are sensitive to the silhouette contour, occlusion, resolution, etc., which affects their accuracy for activity recognition. The method in [22] and [23] confuse similar activities while only two features of the method in [21] are not sufficient for recognizing more than two similar activities.

The method in [33] uses the Poisson equation to obtain the torso, and negative minimum curvature to locate the SBPs. An 8D feature descriptor from the articulated model is used with the HMM to recognize six activities. In [38], the dominant points along the convex hull of a silhouette contour are used with the body ratio, appearance, etc., to fit a predefined model. It is extended in [71] for activity recognition. These methods are evaluated on non-standard and publically unavailable data sets. The method in [71] confuses similar activities. The method in [19] uses the convex hull to identify the SBPs. However, it is designed to be used for surveillance purposes. In Chapter 3 implicit body models are used with the convex hull of a human contour to label SBPs. It tracks the SBPs by using a variant of the Particle Filter described in Chapter 4. This method works in real-time by fitting the knowledge from the implicit body models. It outperforms most of the cutting edge methods that use the distance curve method. Thus, we are motivated to extend and apply it for activity recognition.

5.3 Foundation of proposed method

A DT learns from a data and features the best class separation based on an optimization criteria. Let $p(m|t)$ denote the fraction of samples belonging to a class m at a given node t. Then, for M number of classes, $Entropy = -\sum_{m=0}^{M-1} p(m|t) \log_2 p(m|t)$, can be used as an optimization criteria to determine the best split at each node by measuring the class distribution before and after the split. Techniques such as pruning that optimizes tree depth (leafiness) by merging leaves on the same tree branch can then be used to avoid over-fitting. Random Forest (RDF) is an ensemble learning method that generates many DT classifiers and aggregate their result to avoid over-fitting issue of DT and improve classification performance [127]. Methods like DT and RDF assume that at each node the feature-space can be partitioned into disjoint subspaces, however as mentioned in [18] this does not hold when there are

similar classes or when there are large number of classes. In this case finding a feature-space partitioning that reflects the class-set partitioning is difficult as observed in [18]. Therefore, similar to [18, 121] the goal of this work is to establish a class hierarchy and then train a classifier such as simple binary classifier at each node of the class hierarchy to perform efficient and accurate classification. This allows us to define different set of rules for classifying different types of activities. This is important as different feature sets are useful for discerning different types of activities [128].

In this context, a class hierarchy is created and at each node a binary decision rule is learned that ignores easily confused categories. At the bottom node of the hierarchy a MVS is used to perform decisions on easily confused categories. Let us demonstrate the concept of creating a HRPS using a simple example with three overlapping classes that represent similar categories as shown in Fig. 5.1(a). It can be seen from Fig. 5.1(a) that it is not possible to clearly distinguish between only two overlapping classes by using the RH method as it assumes that only two classes overlap simultaneously. This is because now the overlap is among three classes concurrently, i,e., the overlap between the two classes A and B also contain some overlap with the third class C. Similar phenomena occurs for B and C, and A and C classes. In addition, a combined overlap occurs, i.e, $A \cap B \cap C \neq \emptyset$. Hence, the RH method is not capable of tackling the multiple overlaps class separation problem.

The proposed HRPS method addresses this deficiency in the RH method by splitting the set of classes $K = A' \cup B' \cup C' \cup X$, where $X = \{X_{AB} \cup X_{BC} \cup X_{AC}\}$ and $X_{AB} = A \cap B - A \cap B \cap C$, $X_{BC} = B \cap C - A \cap B \cap C$, $X_{AC} = A \cap C - A \cap B \cap C$ and $X_{ABC} = A \cap B \cap C$. X contains samples from two or more overlapping classes. First, at each level of the hierarchy the clearly separable samples of each class are partitioned into the A' or B' or C' as shown in Fig. 5.1(b)-(d).

$$A' = A - X_{AB} - X_{AC} - X_{ABC} \tag{5.1}$$

$$B' = B - X_{AB} - X_{BC} - X_{ABC} \tag{5.2}$$

$$C' = C - X_{AC} - X_{BC} - X_{ABC}. \tag{5.3}$$

Next, the overlapping samples of each class as shown in Fig. 5.1(e) are partitioned into A or B or C via a majority voting scheme (see Section 5.4.2). The class hierarchy structure for HRPS method is shown in Fig. 5.1(f). Note that at each level one class is partitioned from the remaining group of easily confused classes [113] [121].

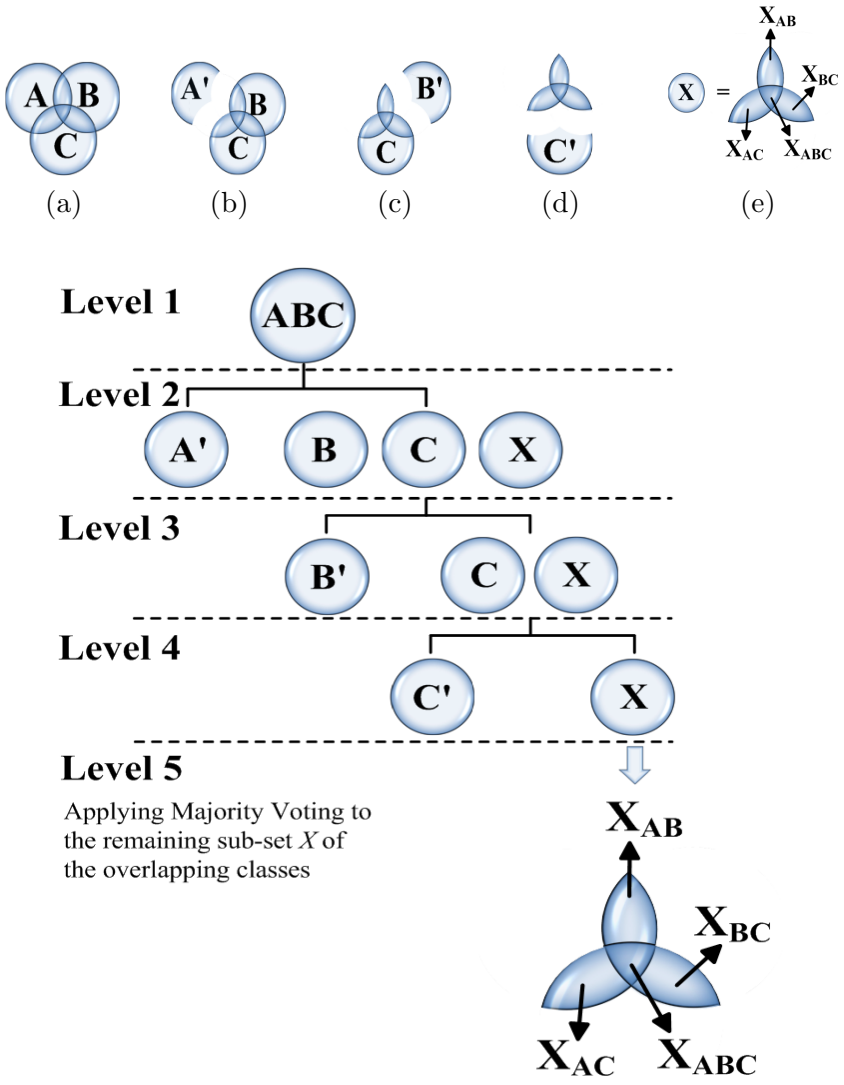


Figure 5.1: (a) Example of three classes to illustrate multiple overlaps class separation problem, (b)-(e) Hierarchical relaxed partitioning system: (b), (c) and (d) Partition non-overlapping samples from class A, B and C respectively, (e) Remaining overlapping samples of all the three classes discerned using the majority voting scheme (see Section 5.4.2 for details), and (f) the corresponding class hierarchy structure.

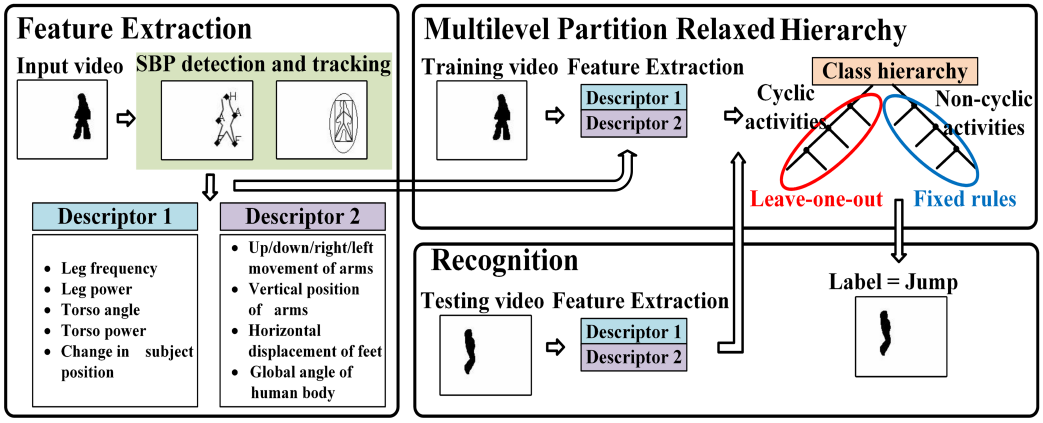


Figure 5.2: The main components and work flow of the proposed human activity recognition.

5.4 HRPS for Activity Recognition

We present HRPS for the Weizmann data set [58] containing multiple similar activities such as Walk, Run, Side, Skip, etc. that are easily confused by the activity recognition methods in the literature. HRPS for the Multi-camera Human Action Video (MuHAVi) data set [73] containing similar activities e.g., walk , run, turn, etc., is also described in order to establish its generality, i.e., adaptability to work on a different data set. The work flow of the proposed activity recognition is shown in Fig. 5.2.

5.4.1 Feature extraction

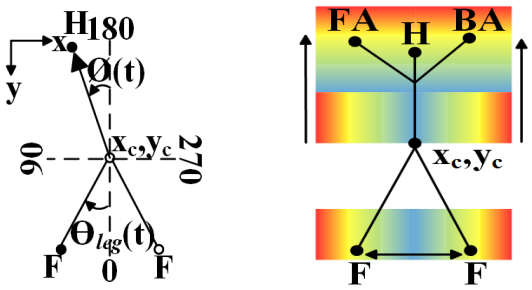
Distinguishing between the cyclic and non-cyclic activities is vital for activity recognition [129]. Thus, we augment our earlier work in Chapter 3 and Chapter 4 to build two feature descriptors $D_{i,\ i=1,2}$. The 2D stick figure shown in Fig. 5.3 (a) is used to describe

$$D_1 = [V_1\ \ V_2\ \ V_3\ \ V_4\ \ V_5] \tag{5.4}$$

for cyclic activities, while the 2D stick figure shown in Fig. 5.3 (b) is utilized to build

$$D_2 = [V_6\ \ V_7\ \ V_8\ \ V_9\ \ V_{10}\ \ V_{11}\ \ V_{12}\ \ V_{13}] \tag{5.5}$$

for non-cyclic activities. The $V_{i,\ i=1,2,...12}$ represents the feature elements of the descriptors. In Fig. 5.3, the SBPs are labelled as the Head (H), Front Arm (FA),



(a) (b)

Figure 5.3: Feature extraction. (a) 2D stick figure analysis for cyclic activities and (b) The upper and lower body analysis based on the arm and feet movement. The SBPs labelled as Head (H), Front Arm (FA), Back Arm (BA) and Feet (F).

Back Arm (BA) and Feet (F). Each SBP abbreviation can be considered as a vector which has a 2D position, e.g, $FA = (x^{FA}, y^{FA})$, $F = (x^F, y^F)$. Here, the superscripts denote the abbreviations of SBP.

The 2D stick figure motion analysis method in [21] uses two motion based features, i.e., the leg power and torso inclination angle, to discern between the Walk and Run activities. This method is suitable for only classifying the cyclic activities with less inter-class similarity, i.e., the activities are not similar to each other. Therefore, we propose two more features, i.e., the torso angle and torso power, to strengthen the method in [21]. Given the global angle from contour moments $V_6 = \theta(t)$ at time t, centre (x_c, y_c), and SBPs from chapter 3 [120], we extend the method in [21] to acquire D_1 which contains four motion based features, i.e., the leg cyclic frequency (V_1) and leg power (V_2), and the torso inclination angle $V_3 = \phi(t) = |90 - \theta(t)|$ and torso power V_4 for the cyclic activities. The foot point $x^F > x_c$ is used for computing

$$\theta_{leg}(t) = \tan^{-1}(\frac{x^F - x_c}{y^F - y_c}). \tag{5.6}$$

Note that this choice does not guarantee the same leg is used for analysis. However, the cyclic nature of the activities makes it unnecessary to detect the same leg in every frame of the video sequence because the cyclic nature is discernible from the motion of this SBP [21].

The computed torso angle $V_3 = \phi(t)$ and leg angle $\theta_{leg}(t)$ are converted into

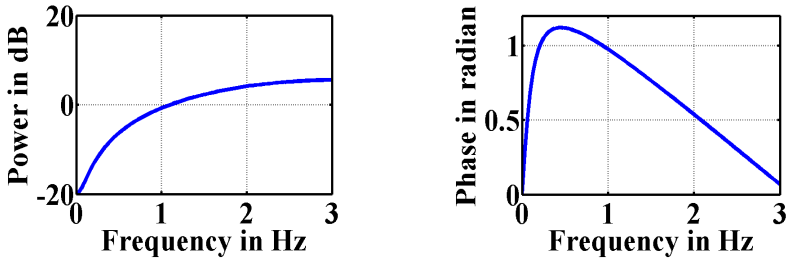


Figure 5.4: High pass filter. (a) magnitude-frequency response and (b) phase-frequency response.

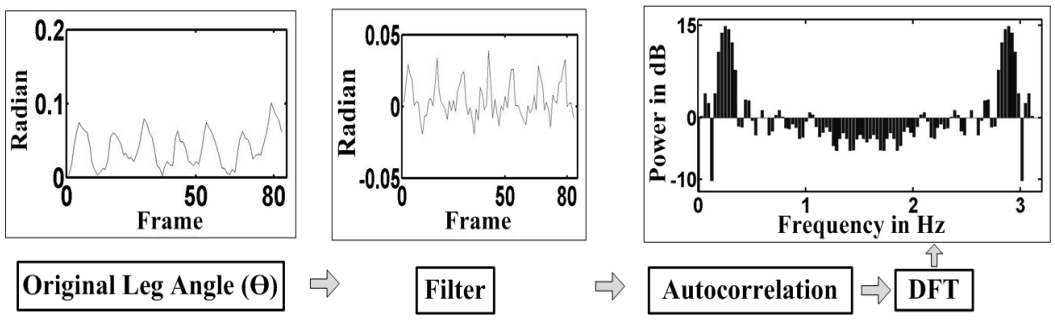


Figure 5.5: Process of acquiring D_1 feature descriptor for the cyclic activities.

radians. A highpass digital filter $Y(e^{jw})$ is applied to $\theta_{leg}(t)$.

$$Y(e^{jw}) = b(1) - b(2)e^{-jw} \tag{5.7}$$

Here, $b(1) = 1, b(2) = -0.9$ as in [21]. The magnitude-frequency response and phase-frequency response of this filter are shown in Fig. 5.4. The filtered leg angles $\theta_{leg}(t)$ are then autocorrelated in order to emphasise the major cyclic components as shown in Fig. 5.5 middle column. The discrete Fourier transform (DFT) is applied to the autocorrelated leg angles to quantify the leg frequency V_1 and magnitude expressed as leg power V_2 in decibels [21] as shown in Fig. 5.5 right column. It shows that the for Walk most freuqncies are in the 1-2Hz range with low power. In this work the high pass digital filter $Y(e^{jw})$ is also applied to the torso angle V_3 (in radians) in order to remove the low frequency components in contrast to [21] where this filter is only applied to the leg angle $\theta_{leg}(t)$. Next, the autocorrelation and DFT steps in Fig. 5.5 are performed on the filtered torso angle to compute a

new feature, i.e., the torso magnitude expressed as torso power V_4 in decibels. This extension allows us to extract more distinct characteristics from the leg and torso angle features because the feature descriptor D_1 contains four motion based features as compare to two features used in [21]. Most of the similar cyclic activities can be easily distinguished due to different cyclic leg frequency and leg power, torso angle and torso power. The change in direction of movement or position of subject is incorporated as

$$V_5 = min(x_c^{t+1} - x_c^t) \tag{5.8}$$

$\forall\ t \in 1, N-1$, where N is the total number of frames, min gives the minimum value. A positive and negative value of V_5 respectively indicate whether subject moved in the same direction or changed direction (turn around) of movement during an activity.

The feature descriptor D_2 characterises the upper body (torso and arms) and lower body (legs) movements as a proportion of the mean height μ_h at different directions during an activity as shown in Fig. 5.3 (b) for the non-cyclic activities. The inter-frame displacement (movement) of the front and back arms are described as

$$V_7 = max(|x_{t+1}^{FA} - x_t^{FA}|)/\mu_h, \quad V_8 = max(|y_{t+1}^{FA} - y_t^{FA}|)/\mu_h \tag{5.9}$$

$$V_9 = max(|x_{t+1}^{BA} - x_t^{BA}|)/\mu_h, \quad V_{10} = max(|y_{t+1}^{BA} - y_t^{BA}|)/\mu_h \tag{5.10}$$

$\forall\ t \in 1, N-1$, max gives the maximum value. The features V_7, V_8, V_9, and V_{10} do not contain information with respect to the actual positioning of the front and back arm SBPs, i,e., where the arm displacement is being taken place. This information is represented as

$$V_{11} = min(y_t^{FA}), \quad V_{12} = min(y_t^{BA}),\ \forall\ t \in 1, N \tag{5.11}$$

which uses the vertical position of the front and back arms to represent their maximum height (as the minimum y location of the front and back arms). The variation in the lower body movement due to the leg can be represented by computing the maximum inter-frame horizontal displacement between the two feet as

$$V_{13} = max(|x_{t+1}^{F} - x_t^{F}|)/\mu_h,\ \forall\ t \in 1, N-1. \tag{5.12}$$

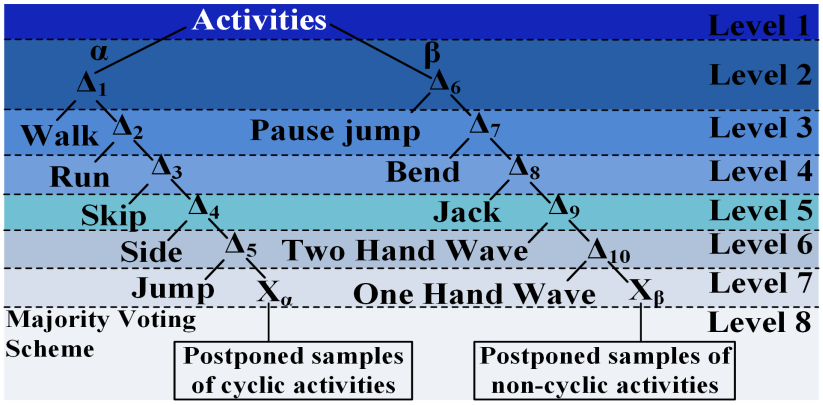


Figure 5.6: Hierarchical relaxed partitioning system for the Weizmann data set. $\Delta_{i,\ i=1,2,..10}$ are the decision rules, and X_α and X_α are the unassigned impure cyclic and non-cyclic activities, respectively, with significant multiple overlaps.

5.4.2 Classification: HRPS for Weizmann data set

The Weizmann data set contain ten activities, i.e., the Walk ($\alpha 1$), Run ($\alpha 2$), Skip ($\alpha 3$), Side ($\alpha 4$), Jump ($\alpha 5$), Jump-in-place-on-two-legs or Pause Jump ($\beta 7$), Bend ($\beta 8$), One Hand Wave ($\beta 9$), Two Hand Wave ($\beta 10$) and Jack ($\beta 11$). In [130], a binary decision tree splits the activities into still and moving categories at the root node in order to obtain better classification. Therefore, an expert knowledge motivated from [130] is added at the root node level 1 to automatically split the above-mentioned ten activities in two groups, i.e., significant translation (α) and no significant translation (β) by using

$$\begin{aligned} \alpha &= 0.25I_w > x_c \text{ or } x_c > 0.75I_w \\ \beta &= 0.25I_w < x_c \text{ or } x_c < 0.75I_w \end{aligned} \tag{5.13}$$

as shown in level 2 of Fig. 5.6. I_w and I_h are the frame width and frame height, respectively. Thus, most cyclic activities, i.e., the Walk ($\alpha 1$), Run ($\alpha 2$), Skip ($\alpha 3$), Side ($\alpha 4$) and Jump ($\alpha 5$), which have significant translation of the subject and repetitive nature are grouped together under α. The activities, i.e., the Pause Jump ($\beta 7$), Bend ($\beta 8$), One Hand Wave ($\beta 9$), Two Hand Wave ($\beta 10$) and Jack ($\beta 11$), which have no significant translation of the subject are grouped under β. A HRPS with 8 levels is created with decision rules $\Delta_{i,\ i=1,2,...10}$ as shown in Fig. 5.6. The decision rules $\Delta_{i,\ i=1,2,...6}$ for cyclic activities are learned by using Algorithm. 5.4.1

Algorithm 5.4.1: PARTITION LEARNING ALGORITHM(D_1)

Input: Training sequences $S_1, ..., S_M$
Corresponding labels $y_1, ..., y_M$
Feature descriptor $D_1 = [V_1 \ V_2 \ V_3 \ V_4 \ V_5]$
Output: Decision rules $\Delta_{i,\ i=1,2,...5}$

Step 1: For each activity, determine the mean μ_j and standard deviation σ_j of feature elements $V_{j,\ j=1,...,5}$ from K training subjects/samples as

$$\mu_j = \sum_{k=1}^{K} V_j^k / K \quad , \ \sigma_j = \sqrt{1/K \sum_{k=1}^{K} (V_j^k - \mu_j)^2}.$$

Step 2: Learn decision rules as one standard deviation on either side of the mean

$$\Delta_{i,\ i=1,2,...5} = \mu_j - \sigma_j < V_j < \mu_j + \sigma_j.$$

Step 3: Update decision rules by using a variable adjustment κ to separate clearly separable samples, i.e., pure samples, of one activity from the samples of all the remaining activities

$$\Delta_{i,\ i=1,2,...5} = \mu_j - \sigma_j + \kappa < V_j < \mu_j + \sigma_j + \kappa$$

Step 4: Accumulate impure samples of an activity that are closer to the samples of all the remaining activities in X_α.

on the training data set that contains the activities performed by eight subjects. The last subject is used as the testing data set in a leave-one-person-out cross validation approach to determine the performance of the HRPS for cyclic activities. The Algorithm. 5.4.1 postpone decisions on those samples of an activity that are closer to the samples of all the remaining activities by updating the decision rules $\Delta_{i,\ i=1,2,...5}$ by using variable adjustment κ. In Chapter 3, SBPs were accurately detected by using implicit body models (IBMs) that are based on the human kinesiology and anthropometric studies, and observed human body characteristics. This inspired us to define decision rules $\Delta_{i,\ i=6,7,...10}$ that are fixed based on the human kinesiology (torso flexion or extension V_6) [90] and anthropometric studies (upper body motion V_7, V_8, V_9, V_{10} and leg motion V_{13}) [6], and individual arm location V_{11} and V_{12}), observed human body characteristics and experimental cues for non-cyclic activities. The Pause Jump ($\beta 7$) is a cyclic activity with no significant translation but has repetitive nature. Thus, it is first separated using V_6 from the non-cyclic activities, i.e., Bend ($\beta 8$), One Hand Wave ($\beta 9$), Two Hand Wave ($\beta 10$), Jack ($\beta 11$). This knowledge will assure an increase in the accuracy and reliability of the activity

classification.

$$\Delta_6 = \begin{cases} \beta 7 & \text{if } |90 - V_6| < 9 \\ \Delta 7 & \text{Otherwise.} \end{cases} \tag{5.14}$$

A full flexion of the vertebra in the Bend ($\beta 8$) activity causes a large increase in the torso angle [90]. Based on the experimental observation in Section 5.5.1 most training subjects have a torso angle variation greater than 9 degrees, thus,

$$\Delta_7 = \begin{cases} \beta 8 & \text{if } |90 - V_6| > 9 \\ \Delta 8 & \text{Otherwise.} \end{cases} \tag{5.15}$$

The Jack ($\beta 11$) activity which involves a large upper body and lower body movement is determined based on large arm and feet displacement by using

$$\Delta_8 = \begin{cases} \beta 11 & \text{if } V_7 \text{ or } V_8 > 15/\mu_h \text{ and } V_9 \text{ or } V_{10} > 15/\mu_h \\ & \text{and } V_{13} > 20/\mu_h \\ \Delta 9 & \text{Otherwise.} \end{cases} \tag{5.16}$$

where $\mu_h = 68$ pixels for the Weizmann data set. The human head is one-eighth the human height, i.e., 0.125. Hence, a 15 pixel movement equates to $15/68 = 0.22$ that is almost twice of the height of the human head.

The individual arm motion in the Two Hand Wave ($\beta 10$) and One Hand Wave ($\beta 9$) activities is discerned using the location information. In the Two Hand Wave ($\beta 10$) activity there will be significant movement of both arms while in the One Hand Wave ($\beta 9$) activity there will be significant movement of only one arm. Therefore, the Two Hand Wave ($\beta 10$) and One Hand Wave ($\beta 9$) activities are described below:

$$\Delta_9 = \begin{cases} \beta 10 & \text{if } V_{13} < 20/\mu_h \text{ and } V_8 \geq 5/\mu_h \text{ and } V_{10} \geq 5/\mu_h \\ & \text{and } V_{11} \leq 55 \text{ and } V_{12} < 50 \\ \Delta 10 & \text{Otherwise.} \end{cases} \tag{5.17}$$

$$\Delta_{10} = \begin{cases} \beta 9 & \text{if } V_{13} < 20/\mu_h \text{ and } V_8 \text{ or } V_{10} \leq 8/\mu_h \\ & \text{and } V_{11} \leq 55 \text{ and } V_{12} > 50 \\ X_\beta & \text{Otherwise.} \end{cases} \tag{5.18}$$

5.4.2.1 Majority Voting Scheme (MVS)

The unassigned impure activities X_α and X_β at the second last level of the HRPS (see Fig. 5.6) are given a label by using a novel majority voting scheme in Fig. 5.7.

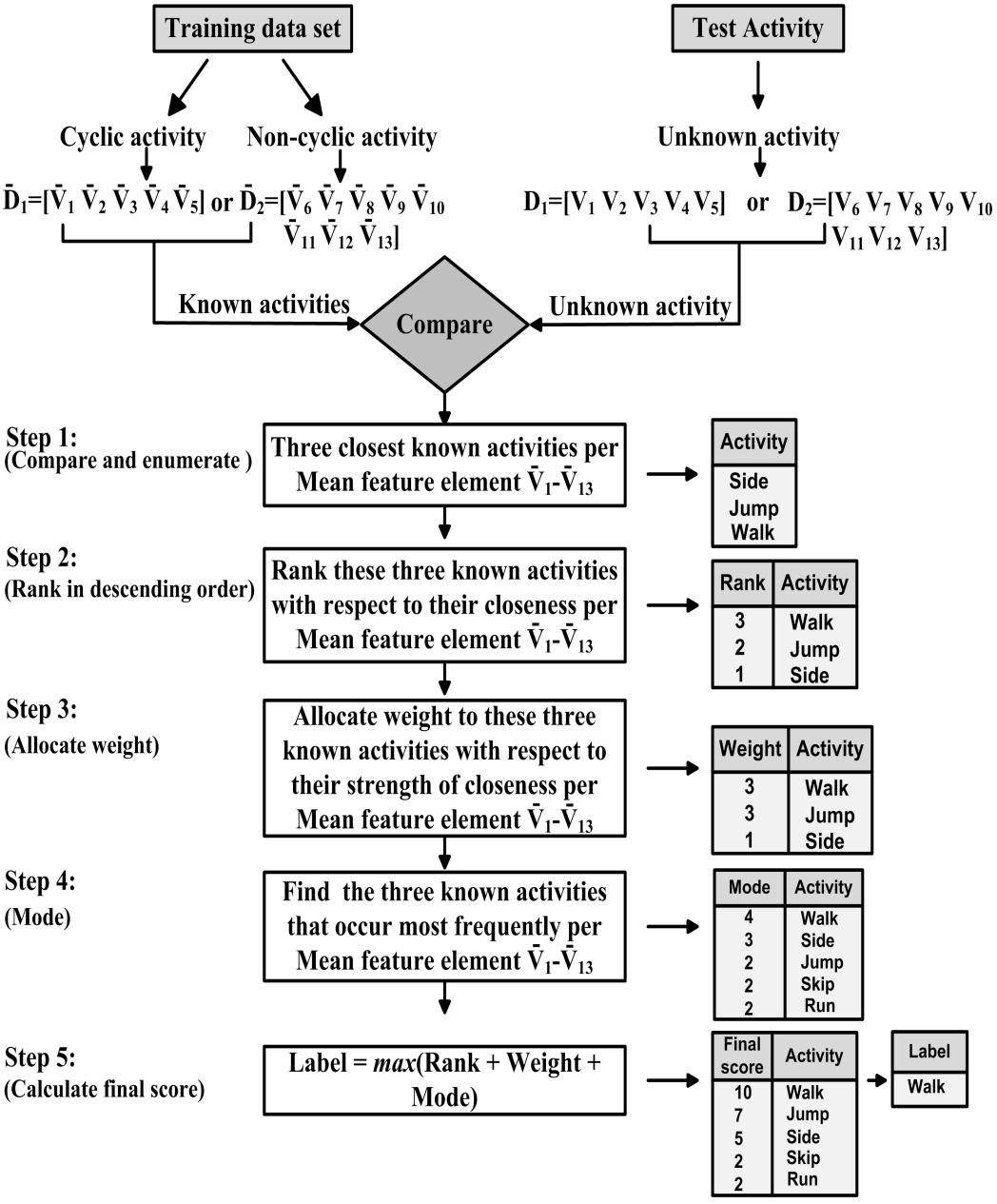


Figure 5.7: Proposed majority voting scheme for the unassigned impure activities X_α and X_β using the mean $\bar{D}_{i,\ i=1,2}$.

This scheme is an integral part of the HRPS and is designed to cater for the increase complexity of multiple overlaps in the feature space of two or more activities. The key idea of this scheme is to accumulate votes based on the rank, assigned weight and frequency (mode) value in order to deduce more accurate decisions at the bottom level of the HRPS.

Given the mean feature descriptors, i.e., $\bar{D}_1 = [\bar{V}_1 \ \bar{V}_2 \ \bar{V}_3 \ \bar{V}_4 \ \bar{V}_5]$ and $\bar{D}_2 = [\bar{V}_5 \ \bar{V}_6 \ \bar{V}_7 \ \bar{V}_8 \ \bar{V}_9 \ \bar{V}_{10} \ \bar{V}_{11} \ \bar{V}_{12}]$, of the known activities of training data set, the goal is to label an unknown impure activity (which contain significant overlaps in the feature space) by extracting the feature descriptors, i.e., $D_1 = [V_1 \ V_2 \ V_3 \ V_4 \ V_5]$ and $D_2 = [V_6 \ V_7 \ V_8 \ V_9 \ V_{10} \ V_{11} \ V_{12} \ V_{13}]$, in order to calculate the rank, weight and mode as shown in Fig. 5.7. D_1 and D_2 are used for cyclic and non-cyclic activities, respectively. $V_1 - V_{13}$ represent each feature element of the feature descriptors. The label for the unknown impure activity is determined as follows.

- **Step 1:** Compare each feature element of the feature descriptor, i.e., D_1 or D_2, of one unknown impure activity with the respective mean feature elements of the feature descriptor, i.e., $\bar{D}_1$ or $\bar{D}_2$, for each of the known activities in order to enumerate three closest known activities per mean feature element.

- **Step 2:** Assign a score (rank) $\nu = 3, 2, 1$ to the three activities enumerated in Step 1 based on their closeness to each of the mean feature elements of $\bar{D}_1$ or $\bar{D}_2$. Next, arrange them in the descending order of their ranks.

- **Step 3:** Allocate a weight $\omega = 3, 2, 1$ to the three ranked activities in Step 2 based on their strength of closeness to the mean feature elements of $\bar{D}_1$ or $\bar{D}_2$.

- **Step 4:** Find the three known activities that occur most frequently (i.e., mode ϖ) per mean feature element of $\bar{D}_1$ or $\bar{D}_2$.

- **Step 5:** Calculate the final score to find the label of the unknown activity. The known activity of the training data set whose rank, weight, and mode yield the maximum score with respect to the unknown activity is assigned as the label for the unknown activity, i.e., Label=max$(\varpi + \nu + \omega)$.

5.4.3 Classification: HRPS for the MuHAVi data set

The robustness of the proposed HRPS method is further validated by applying it with the same feature descriptors $D_{i,\ i=1,2}$ on the MuHAVi dataset [73]. The

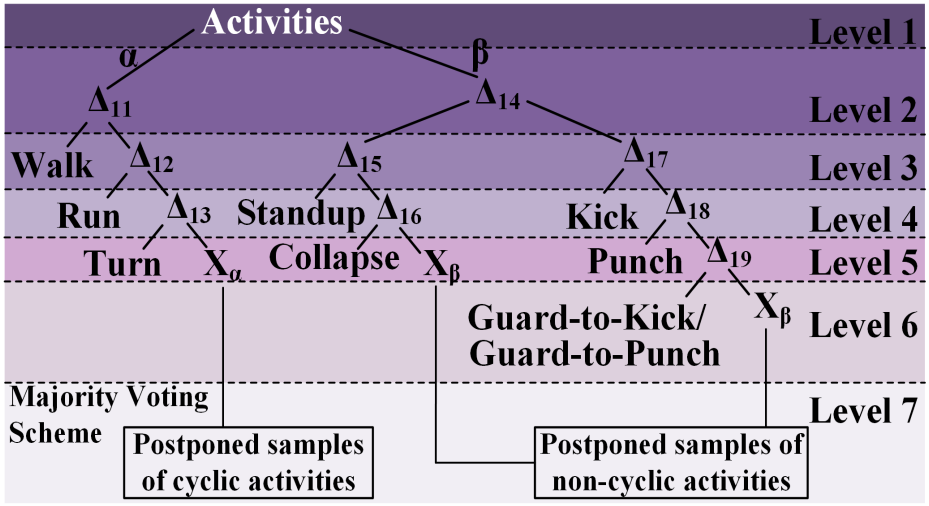


Figure 5.8: Hierarchical relaxed partitioning system for the MuHAVi data set. $\Delta_{i,\ i=11,12,..19}$ are the decision rules, and X_α and X_β are the unassigned impure cyclic and non-cyclic activities, respectively, with significant multiple overlaps.

MuHAVi data set contain eight activities, i.e., the Walk ($\alpha 1$), Run ($\alpha 2$), Turn ($\alpha 6$), Standup ($\beta 12$), Collapse ($\beta 13$), Kick ($\beta 14$), Punch ($\beta 15$) and Guard-to-kick or Guard-to-punch ($\beta 16/\beta 17$). As in Section 5.4.2 the root node is split into α and β activities by using (Eq. 5.13). A HRPS with 7 levels is created with decision rules $\Delta_{i,\ i=11,...,19}$ as shown in Fig. 5.8. Algorithm. 5.4.1 is used on the 7 training samples of the MuHAVi data set to learn the decision rules $\Delta_{i,\ i=11,12,13}$ for the Walk ($\alpha 1$), Run ($\alpha 2$) and Turn ($\alpha 6$) cyclic activities respectively. The last sample is used as the testing data in a leave-one-out procedure to determine the performance of the HRPS.

Similar to Section 5.4.2 we define decision rules $\Delta_{i,\ i=14,...,19}$ that are fixed based on the human kinesiology [90], anthropometry [6] and body characteristics for non-cyclic activities. Let the reference global angle $V_6 = \theta(t)$ in Stand posture be 90^o. Then, based on biomechanical analysis [92] of human spine the maximum flexion of torso is 60^o, i.e., $(90-60 = 30$ or $90+60 = 150)$, which causes a significant change in posture. Thus,

$$\Delta_{14} = \begin{cases} \Delta 15 & \text{if } 30 \geq V_6 \geq 150 \\ \Delta 17 & \text{Otherwise} \end{cases} \tag{5.19}$$

is used to determine whether a transition occurred $\forall\ t \in 1, N$ frames of the activity video. The transition Δ_{15} includes Standup ($\beta 12$) and Collapse ($\beta 13$) activities which contain significant change in posture while the non-transition Δ_{16} contain Kick ($\beta 14$), Punch ($\beta 15$) and Guard-to-kick or Guard-to-punch ($\beta 16/\beta 17$) which do not have significant change in posture. The decision rules for the Standup ($\beta 12$) and Collapse ($\beta 13$), i.e., Δ_{15} and Δ_{16}, respectively are defined as

$$\Delta_{15} = \begin{cases} \beta 12 & \text{if } 30 \geq V_6 \geq 150, \text{ at } t = 1 \\ & \text{and } 65 \leq V_6 \leq 125, \forall\ t \in 2, N \\ \Delta 16 & \text{Otherwise} \end{cases} \tag{5.20}$$

$$\Delta_{16} = \begin{cases} \beta 13 & \text{if } 65 \leq V_6 \leq 125, \text{ at } t = 1 \\ & \text{and } 30 \geq V_6 \geq 150, \forall\ t \in 2, N \\ X_\beta & \text{Otherwise} \end{cases} \tag{5.21}$$

The range $125 - 65 = 60^o$ [92] is selected as it corresponds to the flexion and extension range of human body while maintaining a somewhat Stand posture. We are motivated from Chapter 3 to borrow the definition of the Kick and Punch IBM as decision rules for the Kick ($\beta 14$) and Punch ($\beta 15$) activities. Hence,

$$\Delta_{17} = \begin{cases} \beta 14 & \text{if } 2 \leq 90 - V_6 \leq 15 \\ \Delta 18 & \text{Otherwise.} \end{cases} \tag{5.22}$$

$$\Delta_{18} = \begin{cases} \beta 15 & \text{if } 90 - V_6 > 15 \\ \Delta 19 & \text{Otherwise.} \end{cases} \tag{5.23}$$

Note that in Punch ($\beta 15$), the arm moves across the body in a diagonal manner and as a result the angle of body from the vertical is quite large. The Guard-to-punch and Guard-to-kick are considered as one class because both primarily have a guard activity with minimal movement of the arms and legs. In Guard-to-kick or Guard-to-punch ($\beta 16/\beta 17$), the human remains in Stand posture with least angle of body from the vertical. Hence,

$$\Delta_{19} = \begin{cases} \beta 16/\beta 17 & \text{if } 90 - V_6 < 2 \\ X_\beta & \text{Otherwise.} \end{cases} \tag{5.24}$$

The unassigned impure activities X_α and X_β are given a label by using the MVS (see Section 5.4.2.1).

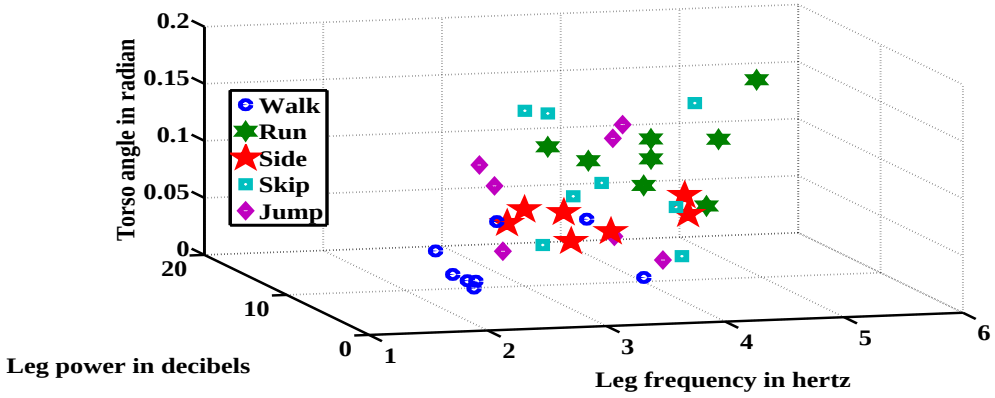
5.5 Experimental results

The Weizmann dataset [58] comprises ninety low-resolution 180 × 144 video sequences of nine subjects performing ten daily activities. The MuHAVi dataset [73] comprises eight high resolution 720 × 576 primitive activity classes of two actors with two samples with two different views (camera 3 and camera 4), i.e., total eight samples, per activity. We use a standard leave-one-out cross validation method.

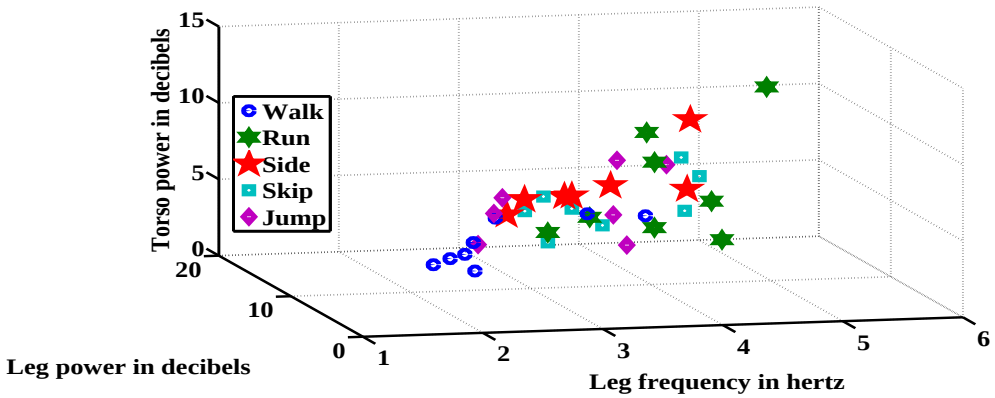
5.5.1 Feature extraction evaluation

The 3D scatter plots of the selected features are shown in Fig. 5.9 and Fig. 5.10 to visualize the distribution of the activities of the input data set. It can be seen from Fig. 5.9 (a) that the Walk activity has the least leg frequency (most blue circles between 2-3 Hz) and the Run activity has the maximum leg frequency (green pentagons lie between 4-6 Hz onwards). Similarly, it can be seen in Fig. 5.9 (b) that the torso power of the Walk activity is much less than the remaining cyclic activities. In Fig. 5.9 (c) it can be seen that the torso angle of most of the Run (green pentagons), Jump (purple diamonds) and Skip (light blue square) activities is more than the Walk (blue circles) and Side (red stars) activity. It can be observed from Fig. 5.9 (c) that the Walk activity has the least torso angle (blue circles between 0-0.05 radian) while the torso angle for the Side (red stars) activity is concentrated between 0.05-0.1 radian.

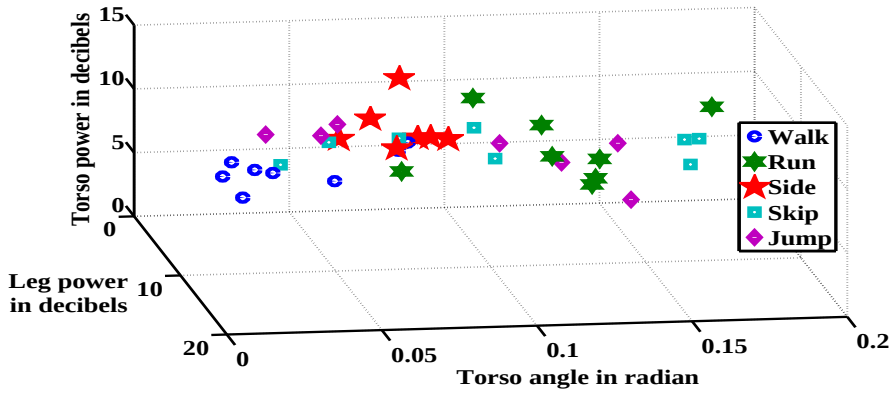
The Fig. 5.10 (a) shows the 3D scatter plots of the selected features for the Bend, Jack, One Hand Wave and Two Hand Wave activities of the Weizmann data set. It can be seen that the Jack activity has the maximum displacement of the feet as a proportion of the mean height of subject. Also, it can be seen that in the Two Hand Wave (light blue square) activity both front and back arm have minimum position in pixels, and is well separate from the One Hand Wave (red star) activity. The Fig. 5.10 (b) shows the 3D scatter plots of a selected feature for the Guard-to-Punch or Guard-to-Kick, Kick and Punch activities of the MuHAVi data set. It can be seen that the Guard-to-Punch or Guard-to-Kick has the least variation in the angle of body from the vertical and the Punch has the maximum angle of body from the vertical. The angle of body from the vertical for the Kick activity lies in between the Guard-to-Punch or Guard-to-Kick and Punch activity.



(a)

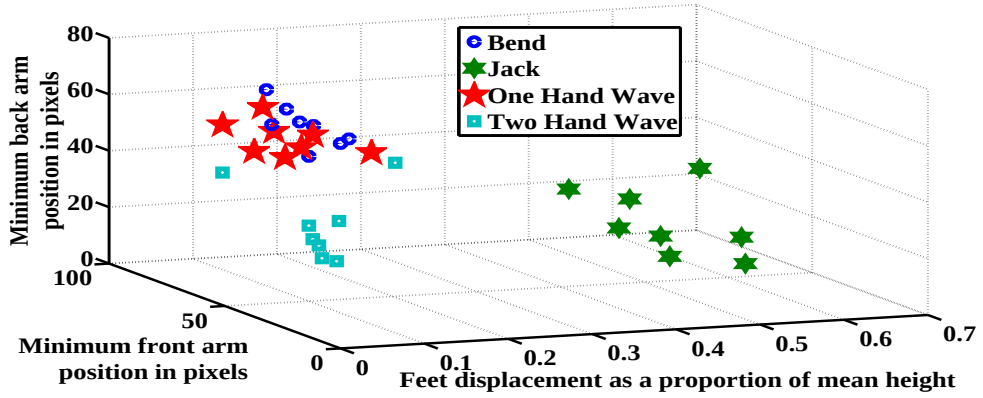


(b)

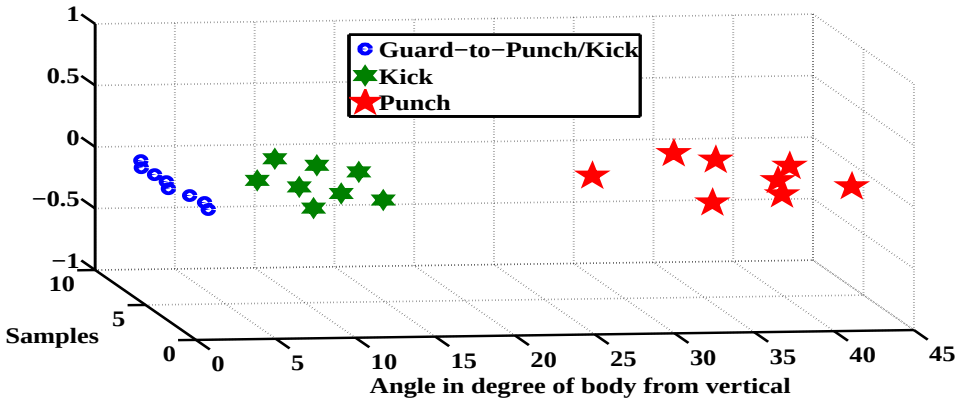


(c)

Figure 5.9: 3D scatter plots of the selected features that show the distribution of the cyclic activities for the input Weizmann data set.



(a)



(b)

Figure 5.10: 3D scatter plots of the selected features that show the distribution of the activities for the input Weizmann and MuHAVi data sets.

In Fig. 5.11, we illustrate the ability of some of the features from $D_{i,\ i=1,2}$ to discern various human activities of the Weizmann and MuHAVi data sets. The error bars show 95% confidence intervals on selected features with two standard deviation as an error metric. Although the leg frequency, i.e., V_1, of the Walk ($\alpha 1$) and Run ($\alpha 2$) activity is dissimilar based on speed of the leg movement but anomalies like some subjects walking faster causes misclassification. However, it can be seen from Fig. 5.11 (a) that the torso angle $V_3 = \phi(t)$ provides a good separation to discern the Walk ($\alpha 1$) and Run ($\alpha 2$) activities. Similarly, the newly introduced torso

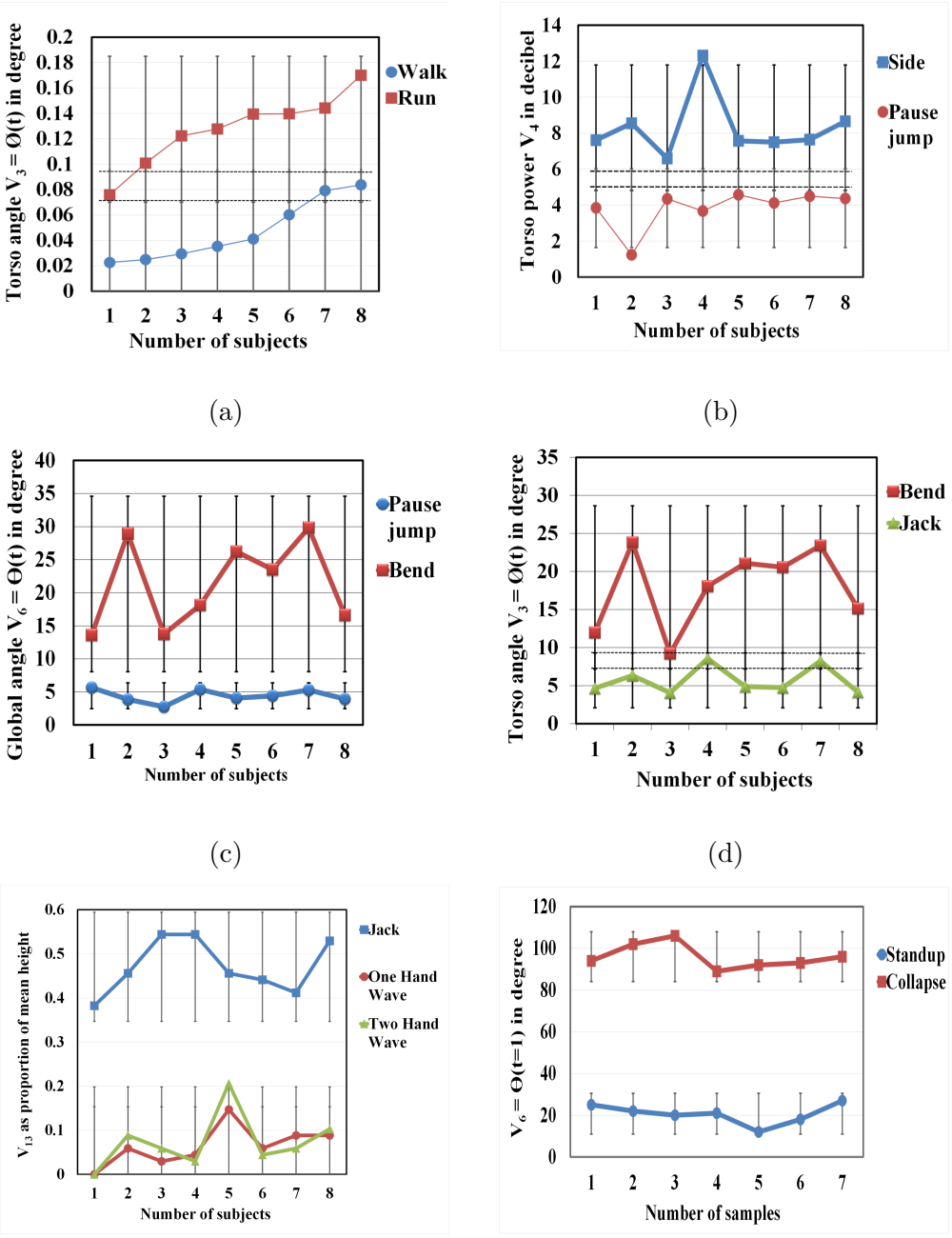
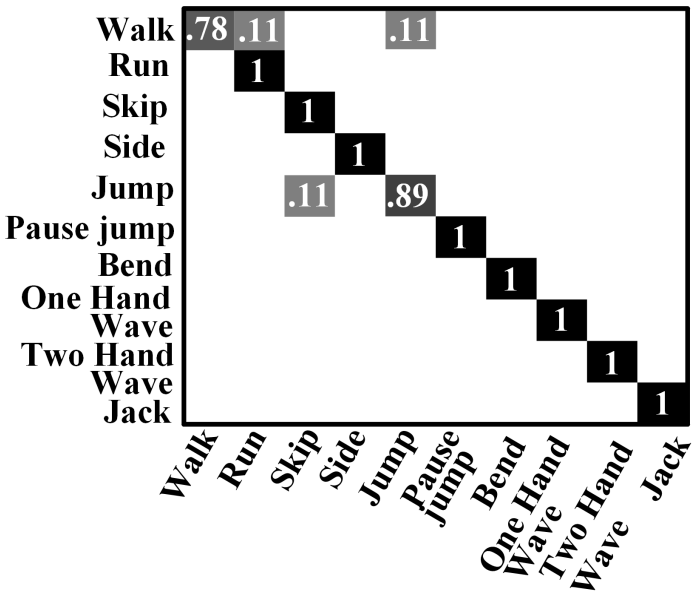


Figure 5.11: Significance of the extracted features for discerning activities. Error bars show 95% confidence intervals on selected features with two standard deviation as an error metric. (a)-(e) Weizmann data set and (f) MuHAVi data set.

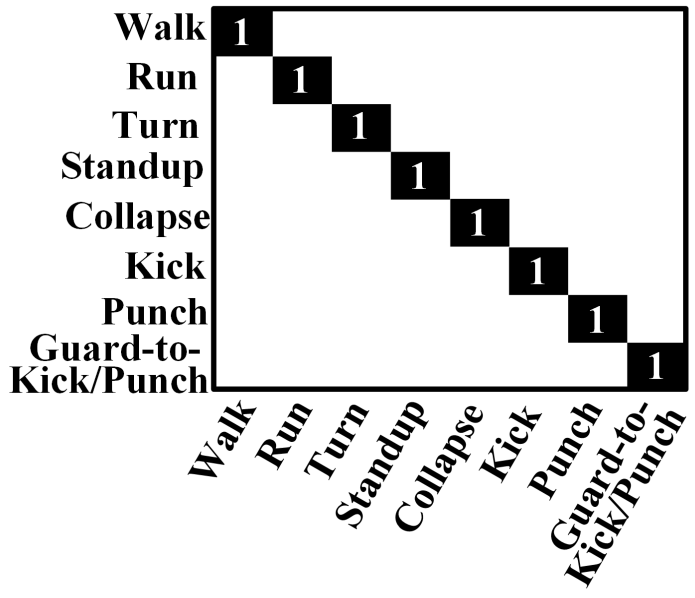
power feature V_4 provides a reasonable distinction between the Side ($\alpha 4$) and Pause Jump ($\beta 7$) activities as shown in Fig. 5.11 (b). In Fig. 5.11 (c), the global angle $V_6 = \theta(t)$ provides clear separation between the Pause Jump ($\beta 7$) and Bend ($\beta 8$) activity while in Fig. 5.11 (d) the torso angle $V_3 = \phi(t)$ provides sufficient discerning ability between the Bend ($\beta 8$) and Jack ($\beta 11$) activity. It can be observed from Fig. 5.11 (e) that the distance between the legs, i.e., V_{13}, gives a very good separation among the Jack ($\beta 11$), One Hand Wave ($\beta 9$) and Two Hand Wave ($\beta 10$) activities. Finally, in Fig. 5.11 (f) the global angle $V_6 = \theta(t = 1)$ easily discern the Standup ($\beta 12$) and Collapse ($\beta 12 = 3$) activities. Thus, the $D_{i,\ i=1,2}$ acquires meaningful information. However, there is slight overlap in the confidence intervals of some of the features, e.g., Fig. 5.11 (a), (b) and (d). This illustrate the importance of using HRPS to postpone decisions on such samples that lie closer to the samples of another activity. Also, for these samples the MVS is better suited because it takes into account multiple criteria based on the average values of all the feature elements obtained from the training data set to assign a label to an unknown activity. As stated in [115] the average features provide more generalized information about the movement pattern of body during an activity.

5.5.2 Classification evaluation

The confusion table for the HRPS method on the Weizmann and MuHAVi data set are shown in Fig. 5.12 (a) and (b) respectively. We obtained a mean classification accuracy of 96.7% for ten activities of the Weizmann data set (see Table 5.1 and details below for significance in comparison to other methods). It shows that our method robustly recognises activities that have significant multiple overlaps in the feature space. In particular, our method recognises four activities, i.e., Run ($\alpha 2$), Side ($\alpha 4$), Jump ($\alpha 5$) and Pause Jump ($\beta 7$), out of the six cyclic activities with a mean classification accuracy of 100%. This proves that our method robustly discerns similar cyclic activities. It obtains a mean classification accuracy of 94.5% for all the six cyclic activities, i.e, Walk ($\alpha 1$), Run ($\alpha 2$), Side ($\alpha 4$), Jump ($\alpha 5$), Skip ($\alpha 3$) and Pause Jump ($\beta 7$). The decomposition of the Walk ($\alpha 1$) into the Run ($\alpha 2$) and Jump ($\alpha 5$) activities is reasonable due to similar motion. Also, the Skip ($\alpha 3$) and Jump ($\alpha 5$) activities are similar in the way the subject bounces across the video. The non-cyclic activities, i.e., Bend ($\beta 8$), Jack ($\beta 11$), Two Hand Wave ($\beta 10$) and One Hand Wave ($\beta 9$) are robustly classified with a mean classification accuracy of 100%. This proves that the decision rules based on human kinesiology and body

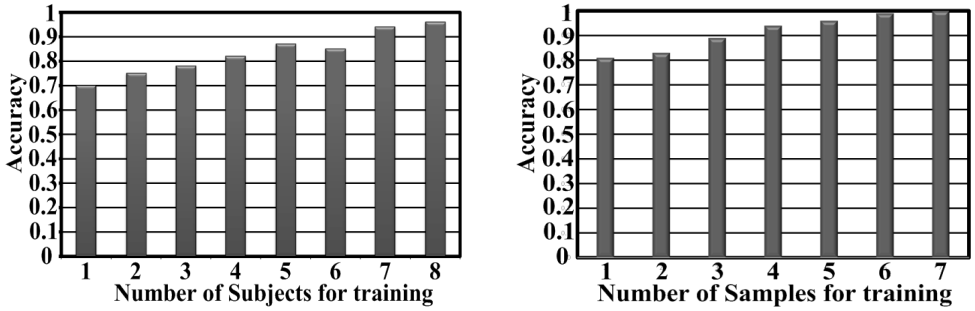


(a)



(b)

Figure 5.12: Confusion table. (a) Weizmann data set and (b) MuHAVi data set.



(a) (b)

Figure 5.13: Classification performance. (a) Weizmann data set and (b) MuHAVi data set.

characteristics work well. We obtained a mean classification accuracy of 100% for eight activities of the MuHAVi data set as shown in Fig. 5.12 (b). The results demonstrate that the proposed HRPS method can robustly distinguish various activities in two different (low and high) resolution data sets. It also show that our method perform well under different views, i.e., camera 3 and camera 4, for the MuHAVi data set. A high accuracy on the Standup (β12), Collapse (β13), Kick (β14), Punch (β15) and Guard-to-kick or Guard-to-punch (β16/β17) activities demonstrate the importance of decision rules based on human kinesiology and body characteristics.

Fig. 5.13 (a) shows classification performance with respect to training subjects of the Weizmann data set. It can be seen that the classification accuracy of the proposed method is about 70% with only one training subject. However, as the number of training subjects increase the classification accuracy also improves. The classification accuracy becomes slightly stable when the number of training subjects is four, five and six. The best performance is achieved with eight training subjects. The classification performance with respect to training samples of the MuHAVi data set is shown in Fig. 5.13 (b). It can be seen that the classification performance increases steadily till it reaches 100% with seven samples used for training.

Table 5.1 compares the HRPS with relevant state-of-the-art methods (see Section 5.2) for activity recognition on the Weizmann data set. It shows that the our method outperforms the methods in [15], [16], [22], [23] in terms of accuracy. Saad et al. [23] only deals with nine activities. The method in [14], [15], [16], [115] and [116] are not real-time since they require intensive training for learning. Zhuolin, et al. [13] required both shape and motion features to achieve 100% accuracy. On

Table 5.1: Comparison on the Weizmann data set.

Method	Accuracy%	Real-time	Intensive training	Year
Michalis, et al. [14]	100	**No**	**Yes**	2014
Marlon, et al. [126]	96.7	Yes	No	2014
Mahbub, et al. [115]	100	**No**	No	2014
Ma, et al. [116]	100	**No**	**Yes**	2013
Romain, et al. [16]	82.79	**No**	**Yes**	2013
Zhuolin, et al. [13]	100	Yes	**Yes**	2012
Saad, et al. [15]	95.75	**No**	**Yes**	2010
Elden, et al. [22]	93.6	Yes	No	2009
Saad, et al. [23]	92.6	-	No	2007
Our method	**96.7**	**Yes**	**No**	2014

a similar basis, i.e., using motion features, they obtain 88.89% accuracy while our method obtains 96.7%. Their method is reported to be fast but requires intensive training and uses optical flow which is usually computationally expensive. Hence, these methods are not suitable for real-world applications. In contrast, our method operates in real-time, avoid intensive training, and it is simple to implement and extend for new activity categories (i.e., for each new category new features can be added to the HRPS). This makes it more suitable for real world applications. The model-free method in [21] recognizes only two activities, i.e., the Walk and Run with 97% accuracy. On similar activities, i.e., Walk ($\alpha 1$), Run ($\alpha 2$), and Jump ($\alpha 5$), the method in [33] has mean classification accuracy of 82.4% while we obtain 92.7% mean classification accuracy. The method in [131] although real-time and non-intensive but achieves only 90.32% on the Weizmann data set. In Table 5.2, our HRPS method is compared with recent methods on the MuHAVi data set. Our method achieved better recognition rate than most of the methods and works in real-time with no intensive training. On both data sets our method is comparable to the method in [126].

On Intel (R) Core (TM) i7 2.93 GHz with 4 GB RAM and Windows 7, the feature extraction in OpenCV 2.4.6 takes 0.031 and 0.071 seconds per image frame on the Weizmann and MuHAVi data sets respectively. The classification in MatLab takes 0.183 seconds for all activities. Marlon, et al. [126] method takes 4.85 and 2859.29 seconds for feature extraction on the Weizmann and MuHAVi data sets

Table 5.2: Comparison on the MuHAVi data set.

Method	Accuracy%	Real-time	Intensive training	Year
Alexandros, et al. [132]	100	Yes	No	2014
Marlon, et al. [126]	100	Yes	No	2014
Alexandros, et al. [131]	97.1	Yes	No	2013
Abdalrahman, et al. [133]	98.5	**No**	No	2011
Sanchit, et al. [73]	97.8	Yes	No	2010
Martinez, et al. [134]	98.4	**No**	**Yes**	2009
Our method	**100**	**Yes**	**No**	2014

respectively. This demonstrates that the HRPS method works in real-time.

5.6 Summary

We proposed a hierarchical relaxed partitioning system to efficiently and robustly recognize activities. Our method first discerns the pure activities from the impure activities, and then tackles the multiple overlaps problem of the impure activities via an innovative majority voting scheme. The results proved that our method not only accurately discerns similar activities, but also obtains real-time recognition on two (low and high) resolution data sets, i.e., Weizmann and MuHAVi respectively. It also performs well under two different views of the MuHAVi data set. These attributes make our method more suitable for real-world applications in comparison to the state-of-the-art methods.

Chapter 6

Conclusions and Future Work

This Chapter concludes the research work carried out in this thesis, presents the significance of the proposed methods and their applications, and the limitations of the proposed methods and future work.

The goal of this thesis is to propose a complete and fully automated system that uses marker-less approach to detect, label, and track human body parts for human activity recognition. The marker-less approach provides an accurate and cost effective solution, which can easily be extended for various applications such as surveillance, assisted living, animation, etc., as compared to the marker based approach, which is very expensive and requires user cooperation, specialized environment and hardware, calibration and set up time per every new scenario, etc. The marker-less model-based approach is explored, in particular, due to its high accuracy over the model-free approach, which is more efficient but lacks good accuracy. In this context, the first step was to propose a novel marker-less model-based method that robustly and efficiently detects and labels human body parts. This method is geared towards human activities observed from the profile view, rather than the front view, which is a more challenging task due to the limited visible surface area of the human body from profile, and self-occlusion of body parts. The next step in the design of a complete system that can perform activity recognition system was to propose robust methods for tracking human body parts during occlusion. Finally, due to the fact that human activity recognition is one of the most active research areas in computer vision and has numerous applications in threat or anomaly detection, incident occurrence, behaviour analysis, etc., the third step in the design of this system is to integrate the methods used for detection and tracking of human

body parts towards human activity recognition.

In Chapter 3, a novel marker-less model-based method is proposed which fits the knowledge from the six implicit body models to detect and label human body parts, rather than explicitly fitting the predefined body models. This is a novel concept which utilizes domain knowledge to detect human body parts, and thus avoids the computationally complex model fitting procedure. The six novel implicit body models have been constructed based on human anthropometry, kinesiology, and human vision inspired studies. This makes them applicable to humans with different anthropometric proportions. The first three implicit body models are designed to detect human body parts in activities in which the human anthropometric body proportions and part positioning are somewhat maintained, e.g., the Head is above the Shoulder, the Arms are above the Knee and below the Head, etc. The remaining three implicit body models are created to detect human body parts in activities in which anthropometric body proportions and part positioning are not maintained, i.e., the Arms go above the Head (e.g., Two hand wave), feet go above the Knee (e.g., Kick). The marker-less model-based human body part detection and labelling is achieved by considering the human body as an inverted pendulum model and then applying ellipse fitting and contour moments procedure to classify it as being in Stand, Sit, or Lie posture. Next, a convex hull method is used on the silhouette contour to determine the extreme locations which are the possible significant body points or parts, i.e., Head, Arm, and Feet. Finally, the significant body points of the human body are labelled by using the six implicit body models. The significant body points are connected to the centre of the human contour to generate realistic 2D stick figures. The proposed method is rigorously evaluated on two different data sets, i.e., Weizmann and MuHAVi, of low (180×144) and high (720×576) resolution, respectively. The qualitative and quantitative results show that the proposed method accurately and reliably detects and labels human body parts in various activities. In addition, the proposed method works in real-time and does not require manual initialization.

In Chapter 4, two novel methods are proposed for human body part tracking during occlusion. The standard Particle Filter struggles to track significant body points when there is no measurement in the image (i.e., in occlusion). Thus, the first proposed method, i.e., the Particle Filter with memory and feedback, combines the temporal information of the previous observation and estimation with a feedback to predict significant body points in occlusion. The proposed method has two op-

eration modes, i.e., no occlusion and occlusion. It behaves like a standard Particle Filter when no occlusion occurs, while it uses memory and feedback when there is occlusion. This method is based on the concept of the temporal Markov chain, i.e., the new state is conditioned directly on the immediately preceding state independent of the previous history. Therefore, the last known measurement in the memory is used to predict in occlusion at first frame, and next this prediction is fed back as an observation for the subsequent occluded frames. This method does not require any prior information about the activity being performed. The human arm is the most occluded body part due rapid motion and self-occlusion in activities observed from the profile view. Thus, the second proposed method, i.e., motion flow, considers the human arm as a pendulum attached to the shoulder joint and defines conjectures to predict the arm during occlusion. The Particle Filter with memory and feedback method is used as default with the above-described significant body part detection method while the motion flow method can be used as per a user's choice. The proposed method is rigorously evaluated on the two above-mentioned low and high resolution data sets. The qualitative and quantitative results show that the proposed tracking methods robustly tracks human body parts during occlusion. The quantitative results also demonstrate that the proposed Particle Filter with memory and feedback enhances the performance of significant body point detection.

In Chapter 5, a novel method, i.e, hierarchical relaxed partitioning system , was proposed for human activity recognition with particular emphasis on multiple overlaps class separation problem in the spatio-temporal domain. The feature space for very similar activities contains significant multiple overlaps which poses great difficulty to accurately classify these activities. The holistic and local feature approaches tackle this problem by intensive training, and extracting computationally complex shape and optical flow features. Thus, an efficient and robust hierarchical relaxed partitioning system was proposed. This method is based on the concept of relaxed hierarchy and hierarchical strategy. The input to the hierarchical relaxed partitioning system are two feature descriptors which are extracted from the 2D stick figure generated using the above-described significant body point detection and tracking method. These two feature descriptors are used to discern the cyclic and non-cyclic activities. The hierarchical relaxed partitioning system employs these two feature descriptors to first discerns the pure (no overlaps occurs) and impure (multiple overlaps occurs) actions, then tackles the multiple overlaps problem of the impure actions via a novel majority voting scheme. The majority voting scheme is

designed to tackle the complex multiple overlaps in the feature space. It uses the two feature descriptors to compare the rank, weight and frequency of known activities with the unknown activity. The unknown activity is given the label of the known activity which has the highest accumulated score of rank, weight and frequency. The proposed hierarchical relaxed partitioning system is evaluated on the challenging low resolution Weizmann data set which contain several very similar activities. It is further verified on high resolution MuHAVi data set to establish its generality. The results show that the proposed method acquires valuable features and robustly discern very similar activities while being comparable to holistic and local feature approaches. The advantage of the proposed method lies in the real-time speed, ease of implementation and extension, and non-intensive training.

In summary, a marker-less implicit body model-based significant body point detection and labelling method is strengthened with a tracking method for robust detection, labelling and tracking of the significant body points or parts. This method is utilized to build feature descriptors which forms an input to the hierarchical relaxed partitioning system for robust and efficient activity recognition.

The human body part detection, labelling and tracking methods developed in this thesis can be employed for various applications such as surveillance, assisted living, behaviour analysis, anomaly detection, activity monitoring, realistic human model generation, human-computer interaction, human-robot interaction, etc. The major advantages of the proposed methods are good accuracy, reliability, high speed, ease of applicability, ease of extension, non-intensive training, and capability to work both on low and high resolution videos. This makes the proposed system more suitable for real world applications.

One of the limitations of the proposed human body part detection and tracking method that it may produce inaccurate prediction when the convex hull does not locate body parts in the first few frames of the video sequence. However, it recovers quickly after the first few frames. A limitation of the proposed activity recognition system is that it only uses motion based features. This, however, can be tackled in future by integrating both shape and motion features to enhance the performance of human activity recognition. In addition, the human body part labelling and tracking can be extended to work for activities observed from multiple views. This is possible because anthropometric constraints have already been used in literature for matching the identified human body parts in the Stand posture with the same posture observed from a different view. It can be an interesting

and powerful enhancement of the proposed method which will boost its applicability to further scenarios. Another future direction is to apply the proposed human body part labelling method on depth images of human activities. Also, the skin colour information can be added to detect face and hands which can increase the performance of hand and head detection. Nevertheless, the proposed method is, as its stands, a real-time universal fully automated and complete method able to detect, label, and track human significant body points for robust and reliable human activity recognition.

Appendix A

Publications

1. F. Azhar and T. Tjahjadi, "Significant Body Point Labelling and Tracking", *IEEE Trans. on Cybern.*, vol. 44, no. 9, pp. 1673-1685, Sep 2014.

2. F. Azhar and C-T. Li, "Multilevel Partition Relaxed Hierarchy for Activity Recognition", *IEEE Trans. on Cybern.*, 2015, (to be submitted).

Significant Body Point Labeling and Tracking

Faisal Azhar, *Student Member, IEEE* and Tardi Tjahjadi, *Senior Member, IEEE*

***Abstract*—In this paper, a method is presented to label and track anatomical landmarks (e.g., head, hand/arm, feet), which are referred to as significant body points (SBPs), using implicit body models. By considering the human body as an inverted pendulum model, ellipse fitting and contour moments are applied to classify it as being in Stand, Sit, or Lie posture. A convex hull of the silhouette contour is used to determine the locations of SBPs. The particle filter or a motion flow-based method is used to predict SBPs in occlusion. Stick figures of various activities are generated by connecting the SBPs. The qualitative and quantitative evaluation show that the proposed method robustly labels and tracks SBPs in various activities of two different (low and high) resolution data sets.**

***Index Terms*—Anthropometry, convex points, implicit body model, significant body points, stick figure.**

I. Introduction

THE marker-less approach to human motion analysis uses video-based methods to detect and track positions of significant body points (SBPs) located at the convex points, i.e., the local maxima, of the silhouette contour. Applications include tracking, stick figure generation, animation for cartoons, and virtual reality, imitation of human action by robots and action recognition for assisted living, surveillance, etc., [1], [2]. The approach offers advantages, e.g., cost effectiveness, no requirement of particular attire and ease of application [3], [4]. The approach can broadly be classified into model-based and model-free approaches. The model-based approach employs a prior model. The model-free approach estimates the motion of regions that enclose relevant anatomical landmarks without prior information about the subject's shape [2]. The former requires fitting, manual annotation, and predefined models which are time consuming while the latter tend to be less accurate.

This paper presents a marker-less method, which uses implicit body models (IBMs), that does not require manual annotation of SBPs, a training phase (learning a classifier), or fitness procedure. IBMs provide anthropometric, geometric, and human vision-inspired constraints for labeling SBPs in activities observed from a profile view and performed by subjects of differing anthropometric proportions. The human body is considered as an inverted pendulum model and ellipse fitting is used to compute the global angle to classify Stand, Sit, and Lie postures. The contour moments are used to find the angle between the principal and vertical axis to provide cues for selecting best IBM. The convex hull [5] of the contour is utilized to determine the locations of SBPs across time. The particle filter method is used to predict SBPs during occlusion, and is compared with the motion flow-based tracker for cyclic activities. Realistic stick figures are generated from the labeled SBPs. The versatility of the proposed method is demonstrated in a number of challenging activities on low and high resolution video data sets.

The paper is organized as follows. Section II presents related methods. The methodology and the proposed framework are presented in Section III and Section IV, respectively. Section V discusses the experimental results, and Section VI concludes the paper.

Manuscript received April 15, 2013; revised September 27, 2013; accepted January 17, 2014. Date of publication February 24, 2014; date of current version August 14, 2014. This work was supported in part by the Warwick Post Graduate Research Scholarship and in part by the Warwick Engineering Bursary. This paper was recommended by Associate Editor D. Goldgof.

The authors are with the School of Engineering, University of Warwick, Coventry CV4 7AL, U.K. (e-mail: faisal.azhar@warwick.ac.uk; t.tjahjadi@warwick.ac.uk; faisal108988@gmail.com).

Digital Object Identifier 10.1109/TCYB.2014.2303993

II. Related Work

The body segmentation and posture estimation method in [1] is model-free and locates convex points on the contour at the local maxima of the distance curve of the silhouette contour pixels. The principal and minor axes of the human body, their relation with the silhouette contour, relative distance between convex points, and convex point curvature are used as rules to label convex points as SBPs. This method uses head point to determine the location of feet, however, an inaccurate head point localization may lead to inaccurate feet point. It also ignores the knee point and does not present quantitative evaluation of labeled SBPs. The Star skeletonization method [6] is also model free and recognizes walk and run from the frequency of leg and torso angles during motion. It does not label local maxima as SBPs.

A model-based modified star skeleton method [7] produces stick figures from monocular video sequences and is extended in connectivity-based human body modeling (CBHM) [8] by using a modified solution of the Poisson equation to obtain torso size and angle. It uses the negative minimum curvature to locate the head, and the nearest neighbor tracking to find the hand and feet. The local maximum method used in [1] and [6]–[8] to identify extremities within the distance curve is sensitive to silhouette contour and these extremities are not always identified due to self occlusion. Furthermore, a smooth distance curve and self occlusion may result in missed local maxima. The method in [9] selects dominant points along the convex hull on a silhouette contour and utilizes prior knowledge of body-ratio within the head, and the upper body and lower body segments to identify SBPs. The body parts are connected to a predefined skeleton model via its center to adapt it to the subject's posture. However, the criteria for labeling convex points as SBPs are not clearly presented in

2168-2267 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

[9]. This method is extended in [10] for activity analysis and 3-D scene reconstruction.

The First Sight (FS) [11] produces stick body parts of a subject performing complex gymnastic movements by matching a prestored labeled body model with an outline of a current image of the subject. The method in [12] generates an elaborate stick figure by a manual selection of anatomical landmarks, body ratios, ratio pruning, and an initial stick figure.

The W4 system [13] classifies a posture into Stand, Sit, Crawl, or Lie, then classifies the postures into front/back, and left-side, and right-side perspectives using vertical and horizontal projection histograms of its silhouette. SBPs are identified using the vertices of convex and concave hulls on the silhouette contour. A topological model is projected onto the contour to label SBPs. The quantitative accuracy of the labeled SBPs is not presented. This system is computationally expensive. In [14], discrete fourier transform (DFT) is applied to the vertical and horizontal histograms of the silhouette. A neural fuzzy network is then used to infer postures from magnitudes of significant DFT coefficients and length-width body ratio. SBPs are not labeled in [14].

In [15], a 2-D model combined with particle filter is used to detect the torso, and color information is used to detect the hands. A posture is recognized by the nearest mean classifier. However, initial camera calibration and use of 500 particles to track only torso and hand limit its application in real time. The method in [16] uses heuristic rules with contour analysis to locate SBPs, and employs color information and particle filter for robust feature tracking. It has only been applied to subjects in Stand. The segmentation of a silhouette contour length into portions is inadequate for activities such as walk, crawl, and bend due to variations in contour lengths. The use of a particle filter with 1000 particles also decreases the speed of computation.

In [17], a part appearance map and an anthropometry-based spatial constraint graph cut are used to locate scope of body parts such as torso, head, arms, and legs. In [18], human body is segmented into parts, and pose is estimated using a combination of joint pixel-wise and part-wise formulation. Each pixel is assigned to an articulated model using a histogram of gradients. This model is segmented into body parts using a given set of joint positions. However, the locations of body parts are not evaluated in these methods.

The pose estimation framework in [19] uses a two layered random forest classifier to localize joints. The first layer classifies the body parts, and the second incorporates the body parts and their joint locations to estimate the pose. In [20], articulated body parts are detected by first finding the torso and then performing a fitness procedure to locate the remaining body parts. It is computationally expensive with no occlusion handling ability.

The recent introduction of the low-cost depth camera has motivated researchers to utilize depth images. In [21], the 3-D pose is estimated from a single depth image. The human body is divided into a set of parts and a random forest is employed to compute the probability of each pixel belonging to each part. The 3-D joint locations are then independently estimated from these probabilities. A similar method in [22] is applied to video images from multiple views. Random forest is used to assign every pixel a probability of being either a body part or background. The results are then back-projected to a 3-D volume. Corresponding mirror symmetric body parts across views are then found by using a latent variable, and a part-based model is used to find the 3-D pose. In [23], a local shape context descriptor is computed from edges obtained from depth images to create a template descriptor of each body part category, i.e., head, hand, and foot. A multivariate Gaussian model is employed on the template descriptor to compute the probability of each category. A greedy algorithm then finds the best match to identify the body parts. The use of multiview and depth images are not within the scope of this paper.

III. Methodology

Human body proportion has been widely studied with applications in engineering, ergonomics, and computer vision [24]. By using the 5th–95th percentile values of body proportion, 90 percent of the world population can be covered [25], [26]. Anthropometry has only been used for Stand postures in a semi-automated manner, since its application in complex actions is not an easy task [27], [28]. Anthropometric transformations do not conform to any known laws, it is thus not possible to formally define invariant properties. A functional definition of anthropometric transforms is presented combining anthropometric, geometric, kinesiology, and human vision (heuristic) inspired constraints, to provide six IBMs for robust labeling and tracking of SBPs. The six IBMs cover most actions, activities and range of motion performed by human from a profile view (see Section V).

In this paper, SBPs are labeled as head (H), shoulder (S), arm (A), knee (K), and feet (F). The abbreviations encapsulate the x-coordinate and y-coordinate of a SBP. The lowercase x and y are, respectively, the x-coordinate and y-coordinate locations of a point. The specific x and y coordinates of an SBP are represented by adding SBP prefixes such as Hx, Hy Ax, Ay etc. The current and previous locations of a point are denoted by lowercase c and p, respectively, e.g., cx, px, Acx, Apx. Subscript refers to a specific entity, e.g., x_c, x_{cv}, and x_{nr} represent the x coordinate of a center, convex point, and normalized convex point, respectively.

A. Implicit Body Models (IBMs)

Several anthropometric studies reveal that in Stand posture the head length is approximately one-eighth the total length of the human body [29]–[31]. The body segment length as a fraction of human body height (1H) is shown in Fig. 1(a), where $8 \times 0.13H \approx 1H$ [31]. These ratios are used to provide ranges of eight segments to label SBPs in Stand posture. The human body maintains an approximate Stand posture in activities such as walk, run, skip, etc. However, these activities induce motion in the vertical plane of the human body which is compensated for by selecting a longer range from the eight segments providing accurate labeling and tracking of SBPs.

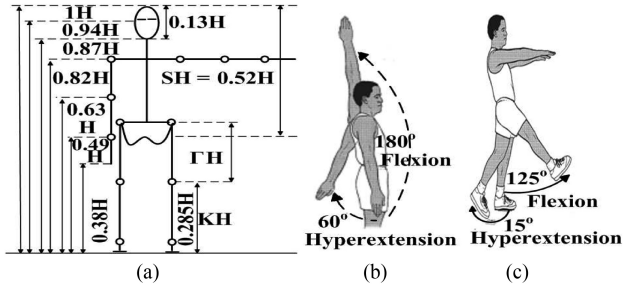


Fig. 1. (a) Body segment lengths as a fraction of the body height (1H). (b) and (c) Arm and leg range of motion based on anthropometric [25], [32], [33] and kinesiology studies [34], respectively.

Thus, the Stand body model is divided into seven segments as shown in Fig. 2(a) (see Section IV-A2).

Anthropometric studies show that in Sit posture the thigh becomes horizontal to the ground and human body height decreases (i.e., head length is not one-eighth the total human body length) [26], [30]. As a result, the Sit posture cannot be divided into eight segments based on empirical anthropometric studies. Note that the body part positioning, (i.e., head, shoulder, arms, knee, and feet above each other, respectively) is somewhat maintained in Sit posture [30]. This problem is resolved by finding the relationship between the segmentation of Sit and Stand postures based on anthropometric studies [26], [30], [31]. According to Fig. 1(a)

$$\Gamma\mathrm{H} = 1H - SH - KH = 1H - 0.52H - 0.285H = 0.195H \quad (1)$$

where ΓH and KH are respectively the thigh length and knee height in the Stand posture. SH is the sitting height (i.e., measured from head to buttocks) in the Sit posture [30].

The number of segments is

$$N_{seg} = \frac{8 \times (1H - \Gamma H)}{H} = \frac{8 \times (1H - 0.195H)}{H} \approx 6. \quad (2)$$

By substituting (1) in (2), for Sit posture N_{seg} should be six, hence, the Sit body model is divided into six horizontal segments as shown in Fig. 2(b). The lie body model is considered as the Stand body model rotated by 90° based on geometry, thus it is divided into seven vertical segments. The Lie body model is further divided into five horizontal segments to account for head leaning [32], [34] in the sagittal plane as shown in Fig. 2(c). These three IBMs can be used to label SBPs in cyclic activities (e.g., walk, side, and skip), and in Stand, Sit, and Lie postures. In all of these activities, anthropometric body proportions and part positioning are somewhat maintained. However, in activities such as bend, wave, punch, and kick, the anthropometry based positioning of body parts/points is not maintained, i.e., the hand goes above/near the head (in wave, punch) or below the knee (in bend), and the feet go above the knee and center of contour (in kick) [25], [32]–[34].

The IBMs are defined based on a range of motion obtained from anthropometric [25], [32], [33] and kinesiology studies [34], human geometry, and vision constraints. They are used to label and track SBPs in activities that do not exactly maintain anthropometry (see Sections IV-A2 and IV-B4 for details). These models cover a diverse range of motions of the shoulder, hand, arm, elbow, knee, and hip mentioned in kinesiology studies and as shown in Fig. 1(b) and (c) [34]. The Wave IBM in Fig. 3(a) covers a range of motion of shoulder, arm and elbow. The kick IBM in Fig. 3(b) covers a range of motion of knee and leg. The Sit body model slightly overlaps with the bend posture. Finally, the Bend IBM in Fig. 3(b) covers a range of motion of trunk.

B. Inverse Pendulum and Contour Moments

Humans are bipeds and locomote over the ground with the majority of the body mass located two third of the body height above the ground. Due to this reason a human body can be represented as an inverted pendulum which is capable of moving in anterior-posterior (forward-back movement) and medial-lateral (side-to-side movement) directions [35]–[37]. In a simple pendulum, it is assumed that motion happens only in 2-D, i.e., the point of mass does not draw an ellipse but an arc. This conjecture allows us to apply a 2-D ellipse fitting on the inverted pendulum human body model as shown in Fig. 4(a).

The global angle θ and angle of the human body ϕ from the vertical are computed, respectively, using ellipse fitting and contour moments. The contour moments of a continuous image $f(x, y)$ are defined as [38]

$$m_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^p y^q f(x, y) dx dy \quad (3)$$

where p and q are, respectively, the x-order and y-order moment of the contour, and x and y are coordinates. The center of the ellipse enclosing the human body is an approximation of the center (x_c, y_c) the human contour mass, that is

$$x_c = \frac{m_{10}}{m_{00}}, \; y_c = \frac{m_{01}}{m_{00}} \quad (4)$$

where m_{10}, m_{01}, and m_{00} are, respectively, the first and zero order spatial moments. The center (x_c, y_c) is used to calculate the central moment

$$\mu_{pq} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - x_c)^p (y - y_c)^q f(x, y) dx dy. \quad (5)$$

The global angle of the human body is the angle of the axis with the least moment of inertia in degree as shown in Fig. 4(a), that is

$$\theta = \frac{1}{2} \tan^{-1} \frac{2\mu_{1,1}}{\mu_{2,0} - \mu_{0,2}} \quad (6)$$

where $\mu_{1,1}$ is the first order central moment, and $\mu_{2,0}$ and $\mu_{0,2}$ are the second order central moments. The angle of the human body from the vertical using contour moments is computed as $\phi = 90 - \theta$.

C. Theoretical Basis of Motion Flow Prediction

The direction of the instantaneous angular velocity (which is measured over an extremely small time interval [34]) is the basis for motion flow prediction. Consider the human arm as a pendulum attached at the shoulder joint producing curvilinear motion (incurring an angular displacement). As the pendulum (arm) swings from its equilibrium position (vertical)

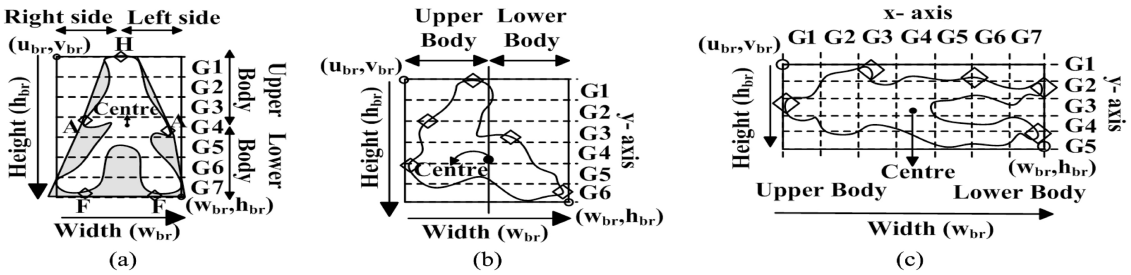


Fig. 2. IBMs for Head (H), Arm (A), and Feet (F) SBP labeling and anthropometry based segmentation [G1–G7] (see Table III) of silhouette contour using bounding rectangle minimum (u_{br}, v_{br}) and maximum points (w_{br}, h_{br}) for (a) Stand (α activities in Table I, convex hull in shaded region), (b) Sit, and (c) Lie.

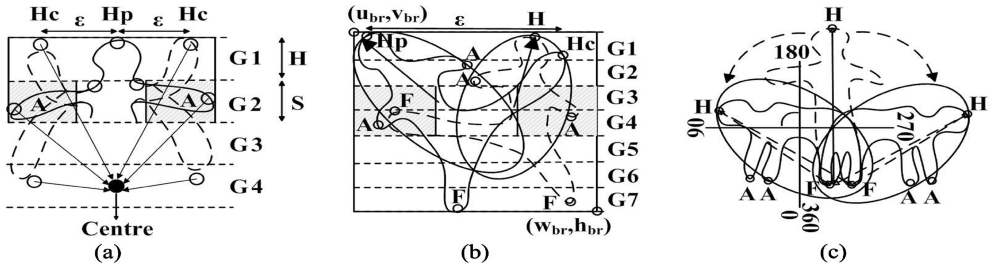


Fig. 3. IBMs based on cues in Section IV-A2 with Smart Search Algorithm (see Section IV-B4) for locating and labeling head (H), arm (A), and feet (F) SBPs in β activities (see Table I). (a) Wave. (b) Kick. (c) Bend.

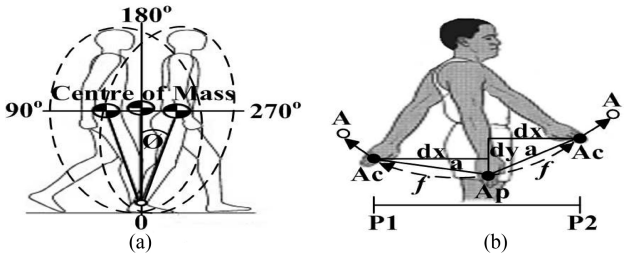


Fig. 4. (a) Inverse pendulum human body model with global angle θ and angle ϕ from the vertical. (b) Motion flow-based arm prediction A using previous arm Ap and current arm Ac during occlusion. (see Section III-C).

to its maximum displacement, the magnitude and direction of angular velocity vector change. Two geometric constraints are proposed for predicting arm location based on pendulum motion. For an extremely small time interval in consecutive time frames:

1) conjecture 1: the direction of the instantaneous angular velocity must be the same until the arm reaches its maximum displacement;
2) conjecture 2: a large instantaneous angular displacement shows that the arm has reached its maximum displacement.

Based on conjecture 1, the point to be predicted should be close to the last arm point and continue in the direction of the previous two arm points, i.e., follow the swing of the arm for cyclic activities as shown in Fig. 4(b). The conjecture 2 identifies the change in the direction of arm swing.

Consider the arm motion as a pendulum swing which draws a small dotted curve f in each frame as shown in Fig. 4(b). Denote (Apx, Apy) and (Acx, Acy), respectively, as the coordinates of labeled arm points in the previous and current frames. For every frame, the linear displacement between the current and previous arm points is

$$dx = Acx - Apx \;\; dy = Acy - Apy. \tag{7}$$

The length L of the entire curve f (i.e., angular displacement) traced by arm movement on the interval [P1-P2] can be approximated as a summation of all the line segments of the entire polygon path. The ath line segment is the hypotenuse of a triangle with base dx and height dy, and has length

$$L_a = \sqrt{(Acx_a - Apx_a)^2 + (Acy_a - Apy_a)^2}. \tag{8}$$

By the mean value theorem, there exists $x_a^* \in [Apx, Acx]$ such that

$$\frac{Acy_a - Apy_a}{Acx_a - Apx_a} = f'(x_a^*) \tag{9}$$

$$Acy_a - Apy_a = f'(x_a^*) \times dx_a. \tag{10}$$

Substituting (10) in (8) gives

$$L_a = \sqrt{1 + [f'(x_a^*)]^2} \times dx_a. \tag{11}$$

Finally, the length of the entire polygon path with k subintervals is

$$\sum_{a=1}^{k} L_a = \sum_{a=1}^{k} \sqrt{1 + [f'(x_a^*)]^2} \times dx_a \tag{12}$$

which has the form of Riemann sum, that is

$$L = \lim_{\Lambda \to 0} \sum_{a=1}^{k} \sqrt{1 + [f'(x_a^*)]^2} \times dx_a = \int_a^k \sqrt{1 + [f'(x)]^2} dx. \tag{13}$$

TABLE I
ACRONYMS FOR ACTIVITIES

Type	Activities (α)
1	Walk
2	Run
3	Skip
4	Side
5	Jump
6	Turn

Type	Activities (β)
7	Jump-in-place-on-two-legs
8	Bend
9	One hand wave
10	Two hand wave
11	Jack
12	Standup
13	Collapse
14	Kick
15	Punch
16	Guard-to-kick
17	Guard-to-punch

Increasing the number of subintervals or line segments of a polygon such that $\Lambda = \max(dx_a) \rightarrow 0$ in (13) proves the approximation that the length of polygon line segments is equal to the length of the curve, i.e., $\sum_{a=1}^{k} L_a \rightarrow L$. This mathematical proof and above-mentioned conjectures lead to the proposed motion flow-based prediction (see Section IV-C2) of arm points as shown in Table IV.

IV. PROPOSED FRAMEWORK

A split approach is developed to simplify the problem and to reduce the search space in order to find the best IBM for labeling the convex points on a silhouette contour as SBPs. This is done using a hierarchical categorization of human posture (Stand, Sit, Lie), movements (Right to left, Left to Right, Stand to Lie, Lie to Stand) and the human body itself (Upper body and lower body, Right side and left side). Stand, Sit, and Lie postures are categorized by considering the human as an inverse pendulum and using contour moments. In Stand, Sit and Lie postures, Upper body and Lower body, and Right side and Left side are respectively distinguished based on the transverse and sagittal planes as shown in Fig. 2 using

$$\begin{aligned} &\text{Stand, Sit}|\delta 1 < y_c \;\&\; \delta 2 > y_c \;\&\; \delta 3 < x_c \;\&\; \delta 4 > x_c \\ &\text{Lie}|\delta 1 < x_c \;\&\; \delta 2 > x_c \;\&\; \delta 3 > C_y \;\&\; \delta 4 < y_c \end{aligned} \tag{14}$$

where body sides $\delta 1$, $\delta 2$, $\delta 3$, and $\delta 4$ are described in Table II.

Initially the Stand to Lie or Lie to Stand movement is ascertained (see Section IV-A1). Fig. 5(a) and (b) is then respectively used to categorize postures in Stand to Lie and Lie to Stand movements according to clockwise and anti-clockwise rotation. Right to Left, Left to Right, and no movement are discerned based on the subjects location in the first frame. In Stand to Lie, for Stand, the movement is further divided into α and β (see Table I). α refers to activities with Right to Left or Left to Right movement, e.g., Walk, Run, Skip, Side, Jump, Turn. β refers to activities in which the subject remains almost at the same place and has Right side or Left side motion, e.g., Jump-in-place-on-two-legs, Bend, One hand wave, Two hand wave, Jack, Standup, Collapse, Kick, Punch, Guard-to-kick, Guard-to-punch. α and β are, respectively, determined using

$$\alpha = \left\{ \gamma 1 | 0.25 \times FR_w > x_c \text{ or } \gamma 2 | x_c > 0.75 \times FR_w \right. \tag{15}$$

TABLE II
ACRONYMS FOR BODY MOVEMENT AND BODY SIDE

Type	Body movement (γ)
1	Right to Left
2	Left to Right
3	Stand to Lie
4	Lie to Stand

Type	Body side (δ)
1	Upper body
2	Lower body
3	Right side
4	Left side

$$\beta = \left\{ 0.25 \times FR_w < x_c < 0.75 \times FR_w. \right. \tag{16}$$

where body movements $\gamma 1$, and $\gamma 2$ are described in Table II. FR_w and FR_h are the frame width and frame height, respectively.

The global angle and the bounding rectangle are respectively used in α and β to select the best IBM for labeling anatomical landmarks. β is further categorized into $\dot{\beta}$ and $\ddot{\beta}$ (see Section IV-A2) to select the appropriate IBM. For any action, the convex points of a human contour are normalized with respect to the bounding rectangle and then filtered. The criteria summarized in Section IV-B from the proposed IBMs are used to label these convex points as SBPs in Stand to Lie, Lie to Stand, α, and β movements. Particle filter (or Motion flow) is used for prediction during occlusion. Finally, the SBPs are connected to generate stick figures for various actions and activities.

A. Silhouette Feature Extraction

1) *Posture Classification:* As in [39] a contour is traced using the freeman chain code [40] on the silhouettes of the Weizmann [41] and multicamera human action video (MuHAVi) datasets [42] (see Section V). A least-squares fitness procedure is used to compute the ellipse global angle θ based on (6) that best approximates the contour.

The maximum flexion and extension range of the trunk in Stand posture, i.e., 140° [33] is used to set the initial global angle θ_{start} parameters such that $255 - 115 = 140°$. This initial global angle is only checked in the first frame of the input video sequence. It is a metric to ascertain the preliminary state of the subject's posture by determining whether the body movement starts from Stand, i.e., Stand to Lie, or from Lie, i.e., Stand to Lie, according to

$$\gamma 3 = \left\{ \text{Stand if } 115 \leq \theta_{start} \leq 255 \right. \tag{17}$$

$$\gamma 4 = \left\{ \text{Lie if } 115 \not\leq \theta_{start} \not\leq 255 \right. \tag{18}$$

where body movements $\gamma 3$ and $\gamma 4$ are described in Table II.

Standard deviation of the global angle has been used to discriminate human shapes, posture-based events and activities [43]. In [1], the difference in angle between the principal and vertical axes is used to detect SBPs but not for posture classification. Biomechanical analysis of human spine show that a complete flexion of the whole trunk occurs due to a rotation of the lumber vertebrae and pelvis, when the difference between the vertical and axis of human body rotation is greater than 50° [33]. A 60° variation in global angle is set to differentiate between Stand and Lie posture for Stand to Lie.

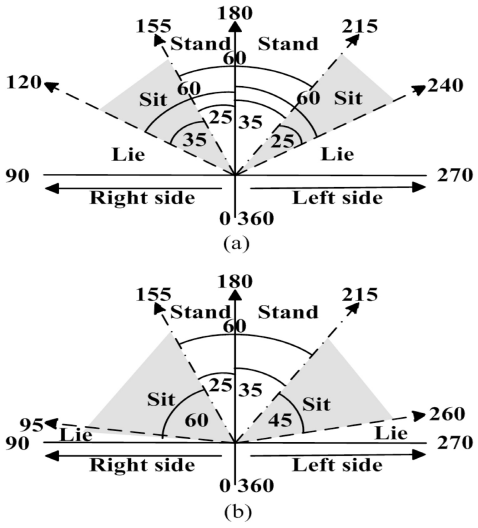


Fig. 5. Stand, Sit, and Lie posture classification using ellipse global angle θ (see Section IV-A1) in movements from (a) Stand to Lie and (b) Lie to Stand.

The reference global angle for Stand is set to 180° in Fig. 5(a). A flexion of more than 60° from the reference in clockwise or anti-clockwise direction is considered as Lie posture, i.e., Lie = $180 \pm 60 = 120°$ or 240°. The human body can flex and extend at a range of $110-140°$ while maintaining a somewhat Stand posture [33]. This yields a variation of 40–70° from the reference global angle with an average of 55°. Thus, the range of angle for Stand posture is set to be $215 - 155 = 60°$, i.e., Stand = $180 + 35 = 215°$ or $180 - 25 = 155°$. The disproportionate division of this range is to cater for the clockwise and anti-clockwise directions leaning ability of the human body while in Stand posture as shown in Fig. 5(a). Sit posture is categorized in the remaining range of angle for clockwise and anti-clockwise directions. It also encompasses intermediate posture such as Bend, manoeuver from Sit to Lie and vice versa.

The range of global angle for Stand in Lie to Stand Fig. 5(b) is kept the same as Stand to Lie, i.e., $215 - 155 = 60°$. However, in trying to Stand from Lie, the body leans forward and the subject remains in intermediate posture (sit) for a longer duration. Thus, a global range of 60° is set for Sit posture in Lie to Stand, i.e., $155 - 95 = 60°$. The Lie posture is categorized in the remaining range of global angle for clockwise and anti-clockwise directions. Fig. 5 illustrates the resulting division of ellipse quadrant used to categorize postures for Stand to Lie and Lie to Stand. A mirror reflection of Fig. 5 is used for the opposite direction of Right side and Left side for Stand to Lie and Lie to Stand. IBM for α activities is selected based on these ranges of global angle.

2) *Posture Segmentation:* The ellipse fitting procedure used in [1] provides approximations, i.e., not body contour points are enclosed by the ellipse as illustrated in Fig. 4(a). The bounding rectangle is used to enclose contour, and obtain its minimum and maximum points, i.e., $P_{min} = (u_{br}, v_{br})$ and $P_{max} = (w_{br}, h_{br})$. u_{br} and v_{br} are respectively the starting x and y coordinates of the bounding rectangle. w_{br} and h_{br} are respectively the width and height of the bounding rectangle.

TABLE III
NORMALIZED SEGMENT VALUES FOR STAND, SIT AND LIE IBM

Model	G1	G2	G3	G4	G5	G6	G7
Stand	0.147	0.295	0.443	0.591	0.738	0.886	1
Sit	0.164	0.328	0.492	0.656	0.742	1	-
Lie	0.194	0.388	0.582	0.776	1	-	-

These points represent the size of the silhouette contour, and are used to divide the body into segments [G1-G7] using anthropometric information [29] (see Section IV-B) defined for IBMs in each of the Stand, Sit and Lie postures as illustrated in Fig. 2. The difference between two segments (which depends on the number of segments N_{seg}) is

$$D_{seg} = (P_{max} - P_{min})/N_{seg} \tag{19}$$

where N_{seg} = 7, 6, 5 and D_{seg} = 30, 21, 22 pixel for horizontal segmentation of Stand, Sit, and Lie, respectively, and N_{seg} = 7 and D_{seg} = 30 pixel for vertical segmentation of Lie. h_{br} and v_{br}, and w_{br} and u_{br} are used in (19) for horizontal and vertical segmentation, respectively. The normalized segments $G[g]$ are determined using

$$G[g+1] = D_{seg} \times (g+1)/(P_{max} - P_{min}), \forall g \in 0 : N_{seg} \tag{20}$$

where $g = 0$ and $g = N_{seg}$ respectively correspond to the minimum and maximum points of the bounding rectangle as shown in Fig. 4(b). Table III shows the normalized segmentation values for Stand, Sit, and Lie posture fixed for all the experiments.

The bounding rectangle along with the angle ϕ from the vertical and global angle θ are used to provide cues to the Smart Search Algorithm (SSA) (see Section IV-B4) for selecting the best IBM for β movements. β is divided into $\dot{\beta}$ and $\ddot{\beta}$, respectively, for $0.7 \times h_{br} > w_{br}$ and $0.7 \times h_{br} < w_{br}$. Thus

$$\beta = \begin{cases} \text{Wave if } \dot{\beta} \text{ and SSA} \\ \text{Kick if } \ddot{\beta} \text{ and } 2 \leq \phi \leq 15 \text{ and SSA} \\ \text{Bend if } \ddot{\beta} \text{ and } 170 > \theta > 190 \\ \qquad \text{and } |H - F| < 1.5 \times D_{seg} \text{ and SSA.} \end{cases} \tag{21}$$

The intermediate postures are selected by wave IBM for labeling, since the subject has yet to attain any defined posture. The Punch action is similar to throwing a ball involving late cocking, acceleration, and follow through. In follow through, the arm moves across the body in a diagonal manner and as a result the angle ϕ of body from the vertical is quite large [33]. Punch action in $\ddot{\beta}$ is labeled using Wave IBM when $\phi > 15$. The range of ϕ in Kick IBM is in between the Stand posture (with tolerance for leaning) and the Punch action. The global angle θ is 170 and 190, respectively, for left and right bend. The bend IBM criteria is formulated based on human vision and kinesiology. The SSA in Section IV-B4 uses (21) in labeling SBPs in Wave, Kick, and Bend IBM.

3) *Convexity Points:* The convex hull method [44] is used to determine SBPs which are located at convex points of a contour, where the line surrounding the silhouette is its convex

hull and the shaded regions are its convexity defects. The convexity defects yield a number of convex points on contour which are marked as head (H), arm (A), feet (F), etc. using the IBM criteria in Section IV-B and as illustrated in Fig. 2(a).

The convex points (x_{cv}, y_{cv}) are normalized with respect to its bounding rectangle to increase the computational speed as follows:

$$x_{nr} = \frac{|x_{cv} - u_{br}|}{w_{br}} \ , \ y_{nr} = \frac{|y_{cv} - v_{br}|}{h_{br}} \tag{22}$$

within [0,1]. The Euclidean distance between convex points is computed as

$$DT_{cv}(i) = \sqrt{(cx_{cv} - px_{cv})^2 + (cy_{cv} - py_{cv})^2} \tag{23}$$

where (cx_{cv}, cy_{cv}) and (px_{cv}, py_{cv}), respectively, denote the current and previous convex points, and i is the number of convex points. Convex points are close to each other in a high resolution video frame but further apart in a low resolution one. This is because in high resolution there are more frequent and sharper edges which will result in more convex points. A threshold Th which is proportional to the frame width FR_w, frame height FR_h and resolution factor Υ are used to remove nearby convex points, where

$$Th = FR_w \times FR_h \times \Upsilon \tag{24}$$

and Υ (determined experimentally) is fixed as follows:

$$\Upsilon = \begin{cases} 0.05 & \text{if } FR_w, FR_h \leq 200 \\ 0.007 & \text{if } FR_w, FR_h \geq 400 \\ 0.01 & \text{if } 200 < FR_w, FR_h < 400. \end{cases} \tag{25}$$

A convex point (x_{cv}, y_{cv}) is selected for labeling by first checking if $CVDT > Th$, where Th is determined by using (24) and (25).

B. SBP Labelling and Tracking

The best IBM is used to label normalized convex points (x_{nr}, y_{nr}) as SBP using Table III as follows. The following SBPs are labeled: head (H), arm/hand (A), knee (K), and feet (F). In the case where multiple criteria are used to label convex points, the abbreviation of a SBP is followed by a numeral, e.g., H1, A1, A2, A3. Convex points (x_{cv}, y_{cv}) upper body, lower body, right side and left side. The ranges for sit and lie have been determined in the MuHAVi dataset since it contains the collapse and Standup activity. Body sides $\delta 1$, $\delta 2$, $\delta 3$, and $\delta 4$ are described in Table II.

1) *Stand:* In Stand posture, Stand to Lie and Lie to Stand, clockwise and anti-clockwise directions, Head and Feet are respectively assigned using

$$\mathrm{H} = \left\{ (x_{nr}, y_{nr}) | y_{nr} < G1 \text{ if } \delta 1 \right. \tag{26}$$

$$\mathrm{F} = \left\{ (x_{nr}, y_{nr}) | y_{nr} > G5 \text{ if } \delta 2. \right. \tag{27}$$

Arm in Stand posture, Stand to Lie, and Lie to Stand for clock and anti-clockwise directions are respectively assigned using

$$\mathrm{A} = \left\{ (x_{nr}, y_{nr}) | G2 < y_{nr} \leq G4 \text{ if } \delta 3/\delta 4 \right. \tag{28}$$

$$\mathrm{A} = \begin{cases} (x_{nr}, y_{nr}) | y_{nr} > G4 & \text{if } \delta 3/\delta 4 \ \& \ \delta 1/\delta 2 \\ (x_{nr}, y_{nr}) | G2 < y_{nr} \leq G4 & \text{if } \delta 3/\delta 4 \ \& \ \delta 2. \end{cases} \tag{29}$$

2) *Sit:* In Sit posture, Stand to Lie and Lie to Stand, clock and anti-clockwise direction, Head and Feet are respectively assigned using

$$\mathrm{H} = \left\{ (x_{nr}, y_{nr}) | y_{nr} < G1 \text{ if } \delta 3/\delta 4 \ \& \ \delta 1 \right. \tag{30}$$

$$\mathrm{F} = \left\{ (x_{nr}, y_{nr}) | y_{nr} > G5 \text{ if } \delta 3/\delta 4 \ \& \ \delta 2. \right. \tag{31}$$

The arm is respectively assigned for Stand to Lie, and Lie to Stand for clockwise and anti-clockwise directions using

$$\mathrm{A} = \left\{ (x_{nr}, y_{nr}) | G1 < y_{nr} \leq G2 \text{ if } \delta 3/\delta 4 \ \& \ \delta 2 \right. \tag{32}$$

$$\mathrm{A} = \left\{ (x_{nr}, y_{nr}) | y_{nr} \geq G5 \text{ if } \delta 3/\delta 4 \ \& \ \delta 2. \right. \tag{33}$$

3) *Lie:* In Lie posture, Stand to Lie and Lie to Stand, clockwise and anti-clockwise directions, Head and Feet are respectively assigned using

$$\mathrm{H} = \begin{cases} (x_{nr}, y_{nr}) | x_{nr} < G1 \text{ if } \delta 1/\delta 3 \ \& \ \delta 4 \\ \ \& \ y_{nr} < G1 \quad \text{if } \delta 1/\delta 3 \ \& \ \delta 4 \\ (x_{nr}, y_{nr}) | x_{nr} < G1 \text{ if } \delta 1/\delta 3 \ \& \ \delta 4 \end{cases} \tag{34}$$

$$\mathrm{F} = \left\{ (x_{nr}, y_{nr}) | x_{nr} > G5 \text{ if } \delta 2. \right. \tag{35}$$

Head is also assigned using

$$\mathrm{H} = \begin{cases} (x_{nr}, y_{nr}) | x_{nr} \geq G2 \ \& \ y_{nr} \geq G4 \text{ if } \delta 1 \\ \text{or } x_{nr} > G2 \ \& \ y_{nr} < G5 \quad \text{if } \delta 1 \\ \text{or } x_{nr} \leq G4 \ \& y_{nr} > G4 \quad \text{if } \delta 2. \end{cases} \tag{36}$$

For Stand to Lie and Lie to Stand, clockwise and anti-clockwise directions, Arm and Head are respectively assigned using

$$\mathrm{A} = \left\{ (x_{nr}, y_{nr}) | G1 < x_{nr} \leq G2 \text{ if } \delta 3/\delta 4 \right. \tag{37}$$

$$\mathrm{H} = \left\{ (x_{nr}, y_{nr}) | x_{nr} < 0.5 \times G1 \text{ if } \delta 1 \ \& \ \delta 3/\delta 4. \right. \tag{38}$$

In Lie to Stand, as the subject is trying to stand, support of arms is used to assist in manoeuvring. (29) for Lie to Stand is utilized for labeling SBPs as the subject is manoeuvring from Sit to Stand. However, during this manoeuvring when $h_{br} > 1.7 \times w_{br}$, (28) is used instead of (29).

4) *Smart Search Algorithm (SSA):* In the β activities, i.e., Wave, Kick and Bend IBMs, SSA is used to label SBPs. Based on (21), SSA is initiated by locating the convex points in the nonanthropometric segment ranges. $\dot{\beta}$ refers to the subject in Stand posture who has yet to attain the posture of models shown in Fig. 3(a)–(c). It is an indication that the subject is likely to perform Wave. In Fig. 3, Hp and Hc are respectively the location of previous (Hpx, Hpy) and current (Hcx, Hcy) head points, and ϵ is the horizontal distance between them. Hx and Hy are respectively the x and y coordinates of head H SBP. SSA divides the wave model into four horizontal

segments, and as the hand goes near or above the head, the following steps are defined for labeling convex points as SBPs in the segment range [G1-G4] as shown in Fig. 3(a).

Step 1: Locate the arm in the segment range $G(1, 2]$ of shoulder S by dividing the bounding rectangle width w_{br} into three equal vertical sections, and reallocate normalized convex points (x_{nr}, y_{nr}) as arm point A if $x_{nr} < w_{br}/3$ or $x_{nr} > 2 \times w_{br}/3$ or $|y_{nr} - Hy| > 0.7 \times D_{seg}$ represented by the shaded region in Fig. 3(a).

Step 2: Verify no arm point was identified using Step 1. Next, every normalized convex point (x_{nr}, y_{nr}) in the head segment range $G[1]$ of Stand to Lie, clockwise and anti-clockwise directions, is reallocated as A if $\epsilon > 0.7 \times D_{seg}$, where $\epsilon = |Hcx - Hpx|$ as shown in Fig. 3(a).

Step 3: Check if no arm point has been labeled using the above two steps. Find two points in the segment range [G1-G4] that are at maximum distance from the center and lie to its right and left, respectively denoted by arrows in Fig. 3(a). These points are then labeled as arm points.

Step 4: If an arm point is labeled using one of the above three criteria then it implies that a wave IBM best represents the activity; hence, the head point is reallocated as follows: $Hx = x_c$, $Hy = y_c - \tau D_{seg}$, where $\tau = 1, 1.7, 2.5$ respectively for resolution factor $\Upsilon = 0.05, 0.007, 0.1$. This is based on the fact that the center of mass moves upward when the human arms are above the head.

In $\ddot{\beta}$ based on (21), for the kick IBM, only Steps 1 and 2 of the SSA are invoked. Steps 1 and 2 are used in the segment range of the arm $G(2, 4]$ and $G[1]$ to reallocate foot point for right and left kick as shown in the shaded region of Fig. 3(b), respectively. In $\ddot{\beta}$ for Bend IBM, the global angle θ is near sit, and the head to feet distance reduces (denoted by dashed arrows) in Fig. 3(c). This model slightly overlaps with the Sit model of Stand to Lie and Lie to Stand, hence, sit criteria stand to lie in Section IV-B2 is used to label SBPs. Depending upon the global angle the proposed framework automatically switches to Lie to Stand using Fig. 5(b).

C. SBP Prediction During Occlusion

1) *Particle Filter-Based Prediction:* A particle filter [5], [45] is able to track and predict SBPs in the presence or absence of occlusion, or missed convex points. Given the current observation of location, i.e, (x_{cv}, y_{cv}), of a SBP at time step $t-1$, the particle filter predicts the location (x'_{cv}, y'_{cv}) of a SBP at time step t. The state vector $X_{t-1} = (x_{cv}, y_{cv}, Vx, Vy)$ is initialized, where (Vx, Vy) are, respectively, the distance between the current and previous SBPs along the x and y directions. A constant-acceleration dynamic model X_t is used to update the state vector, where

$$X_t = M * X_{t-1} \tag{39}$$

$$\mathrm{M} = \begin{bmatrix} 1 & 0 & dt & 0 \\ 0 & 1 & 0 & dt \\ 0 & 0 & dt & 0 \\ 0 & 0 & 0 & dt \end{bmatrix} \tag{40}$$

dt is the time lapse between two frames. For each SBP, particle filter with 100 particles is instantiated for optimum accuracy of prediction with particles ≥ 30 producing good results. During occlusion, the particle filter is initialized with the last known observation to predict the next SBP (x'_{cv}, y'_{cv}). This is achieved by keeping the temporal information of every previous measurement and observation. In the event of occlusion in consecutive frames, the predicted values in the first frame (x'_{cv}, y'_{cv}), $V'x = x'_{cv} - x_{cv}$, and $V'y = y'_{cv} - y_{cv}$ are fed back as observations to initialize particle filter for the subsequent frames.

TABLE IV
PARAMETERS AND THEIR VALUE FOR MOTION FLOW-BASED ARM PREDICTION (α AND β ARE DESCRIBED IN TABLE I)

Act	$\lvert dx\rvert$	Acx	Acy	Ax	Ay
$\alpha 1$	$-, < \zeta$	$\leq Apx$	$\geq Apy$	$Acx \mp dx$	$Acy + dy/0.4\zeta$
$\alpha 1$	$> \zeta$	$-$	$-$	$Acx - 0.4\zeta$	$Acy + dy/0.4\zeta$
$\alpha 2$	$< \zeta$	$\leq Apx$	$\geq Apy$	$Acx \mp dx$	$Acy + dy/0.4\zeta$
$\alpha 2$	$-, \geq \zeta$	$-$	$-$	$Acx \mp 0.8\zeta$	$Acy + dy/0.4\zeta$
$\alpha 3$	$\leq \zeta$	$\leq Apx$	$-$	$Acx \mp dx/0.4\zeta$	Acy
$\alpha 3$	$-$	$-$	$-$	$Hx \pm 1.4\zeta$	$Hy + 4\zeta$
$\alpha 4$	$< \zeta$	$\leq Apx$	$-$	$Acx \mp dx$	Acy
$\alpha 4$	$> \zeta$	$-$	$-$	$Acx \mp dx/\zeta$	Acy
$\alpha 5$	$< \zeta$	$\leq Apx$	$-$	$Acx \mp dx$	Acy
$\alpha 5$	$> \zeta$	$-$	$-$	$Acx \mp dx/\zeta$	Acy
$\beta 7$	$< \zeta$	$-$	$\leq Apy$	Acx	$Acy + dy$
$\beta 7$	$> \zeta$	$-$	$-$	Acx	Acy

2) *Motion Flow-Based Prediction:* Motion flow employs the direction of linear displacement, prior knowledge of the activity, temporal information of an SBP, and geometry of the human body to define criteria for locating, labeling, and tracking SBP, i.e., arm points (Ax, Ay) during occlusion as detailed in Table IV. If the displacement dx between current arm Acx and previous arm Apx point is greater than a threshold $\zeta = D_{seg}/6 = 5$ (where D_{seg}=30, see Section IV-A2), it suggests that the maximum displacement is reached and direction of the arm swing arm has changed. Only dx is used because the horizontal displacement of arm (pendulum) from equilibrium position to maximum displacement is intuitively more than vertical displacement. The direction of the front arm movement is constrained based on the previously labeled front arm points. The criteria in Table IV are used to predict front and back arm points during walk, side, jump-in-place-on-two-legs, jump Left to Right, run Right to Left and skip on the Weizmann dataset.

In Table IV, Hx and Hy, and Ax and Ay, respectively, denote the coordinates of the head and predicted arm points, and Act represents activities (see Table I). The upper polarity is used for Right to Left, and the lower polarity is used for Left to Right. Front arm and Back arm are distinguished, respectively, on Right side and Left side based on (14). For all actions, the arm point is predicted at the center (x_c, y_c) when no conditions are satisfied or when more than three points have been predicted consecutively. In the first row of walk, side, skip, jump-in-place-on-two-legs and run in Table IV, the relational operator and polarity of criteria for current arm (Acx, Acy) and predicted arm (Ax, Ay) are, respectively, reversed for front and back arm prediction in Right to Left

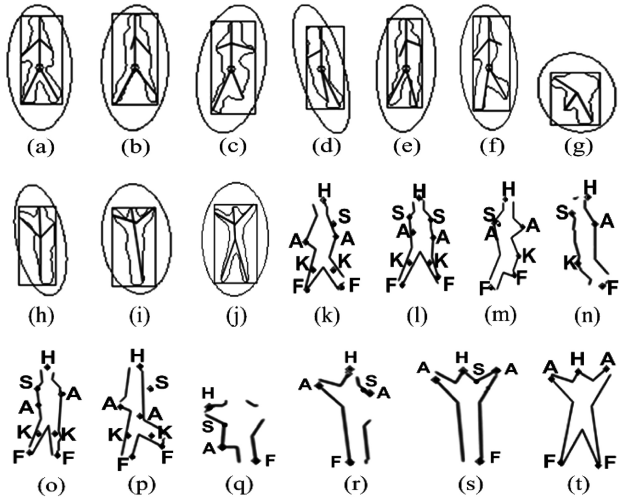


Fig. 6. Weizmann dataset. (a)–(j) Walk, Side, Skip, Jump, Jump-in-place-on-two-legs, Run, Bend, One hand wave, Two hand wave and Jack respectively (contour, bounding rectangle, ellipse, and stick figure). (k)–(t) SBPs labeled as Head (H), Shoulder (S), Arm (A), Knee (K), and Feet (F) in these corresponding actions.

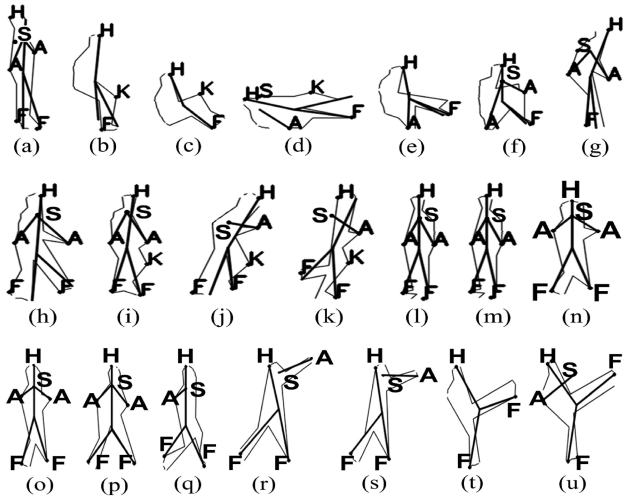


Fig. 7. MuHAVi dataset. SBPs labeled as Head (H), Shoulder (S), Arm (A), Knee (K), and Feet (F) in (a)–(d) Collapse; (d)–(g) Standup; (h) and (i) Walk; (j) and (k) Run; (l) and (m) Turn; (n) and (o) Guard-to-punch; (p) and (q) Guard-to-kick; (r) and (s) Punch; and (t) and (u) Kick.

and Left to Right. The second row of these actions is used to predict back points when they are not predicted by the first row. For walk, dx is not used for front arm point prediction (which is denoted by a dash) but is used to predict back arm point only. For jump, front arm point is predicted at center (x_c, y_c) in occlusion, while the back arm point is predicted using the two rows of jump. However, if $dx > 2\zeta$ pixels then back arm point is predicted at the center.

D. Stick Figure

The proposed framework can be used for the animation of the stick figures of a human body formed by joining the SBPs of every video frame. To form a stick figure, first the maximum distance between shoulder point (Sx, Sy) and head point (Hx, Hy) is computed as

$$Sx = \max(Hx - Sx), \quad Sy = \max(Hy - Sy) \tag{41}$$

for an activity. Noting that a shoulder point is mostly at a constant distance from the head point, (41) is used to find a shoulder point (Sx, Sy) for all activities. According to human anatomy, the head and feet points are connected to the center (x_c, y_c) of the silhouette contour and the arm points are connected to the shoulder point (Sx, Sy).

V. Experimental Results

The Weizmann dataset [41] comprises ninety low-resolution 180×144 video sequences of nine subjects performing ten daily activities as shown in Table I. The MuHAVi dataset [42] comprises nine high resolution 720×576 primitive action classes of two actors with two samples per activity.

A. Qualitative Evaluation

The freeman chain code contours of various subjects enclosed in the bounding rectangle and the rescaled ellipse, with generated stick figures from SBP obtained using the proposed framework on Walk, Side, Skip, Jump, Jump-in-place-on-two-legs, Run, Bend, One hand wave, Two hand wave, and Jack activities are shown in Fig. 6(a)–(j), respectively. Fig. 6(k)–(t) shows the detected SBPs on the corresponding actions. An initial missed or undetected convex point, results in an incomplete stick figure. This is because the proposed framework requires temporal information (at least two convex points) for initialization of prediction using particle filter or motion flow.

The adaptability and generality of the proposed framework is validated by applying it with the same parameter settings on the MuHAVi dataset. Fig. 7(a)–(d) and (e)–(g) respectively show collapse and standup actions with identified SBPs in Stand, Sit, and Lie postures. Fig. 7(h)–(u) illustrate the SBPs identified during Walk, Run, Turn, Guard-to-punch, Guard-to-kick, Punch and Kick, respectively. Figs. 6 and 7 show that the proposed framework successfully labels SBPs and is able to generate stick figures in various actions.

B. Quantitative Evaluation

Most methods in Section II only provide qualitative evaluation. In [1] for computer vision-based human body segmentation and posture estimation (CVHSP), [8] for CBHM, the method in [6] and [9] for star skeletonization, SBPs are detected but the accuracy of their localization with respect to ground truth coordinates of each SBP is not presented. Also, the First Sight [11] detects body parts and not SBPs. Thus, it is not possible to compare the accuracy of SBP localization using the proposed framework with these methods. In Tables V–VIII, the best results are shown in bold.

1) *Accuracy of Localization:* The accuracy of SBP localization is presented in terms of distance in pixels between the manually annotated (i.e., the ground truth) and detected SBPs. Silhouette contours for all activities of the two data sets are skeletonized using the method in [46]. Manual annotation is

TABLE V
AVERAGE ERROR IN PIXELS OF SBPS WITH RESPECT TO GROUND TRUTH

Act	Hx	Hy	FAx	FAy	BAx	BAy	LFx	LFy	RFx	RFy	Avg
	Weizmann Data set with prediction										
$\alpha 1$	2.3	5.5	5.3	7.5	4.8	10.3	4.6	2.4	4.3	2.3	4.93
$\alpha 2$	3.8	5.6	5.3	**3.4**	8.7	8	5	3.7	4	3.4	5.09
$\alpha 3$	4.3	5.4	7	5.9	8.6	6	5	4.1	3.8	2.1	5.22
$\alpha 4$	1.6	5	6.5	6.3	4.5	7.5	3.8	3.1	4	3.5	4.58
$\alpha 5$	3.6	5.1	7.3	11	6.1	7.1	5.3	3.6	5.3	3.6	5.8
$\beta 7$	**1**	4.5	6.5	8.6	3.9	6.5	6.2	2.9	6.2	2.9	4.92
$\beta 8$	7.3	6.5	7.2	9.6	5	6.8	4.2	2.5	4.2	2.5	5.58
$\beta 9$	9.6	5.4	5.2	6	**2.6**	5.2	6	1.7	6	1.7	4.94
$\beta 10$	5.7	**4**	8.5	8.5	8.6	8.7	6	**1.6**	6	**1.6**	5.92
$\beta 11$	5.3	**4**	**3.3**	4.4	2.8	**3.3**	**2.4**	2	**3.2**	2.3	**3.3**
	MuHAVi Data set with prediction										
$\alpha 1$	11	**3.3**	5.7	7.2	8.5	12.3	8	4.6	8.3	4.9	7.38
$\alpha 2$	9.65	3.8	6.4	6.7	9.2	16.3	8.3	5.2	9.7	6	8.12
$\alpha 6$	10.2	3.7	5.7	11.9	**5.3**	14.2	7.7	4.4	8	**4.3**	7.54
$\beta 12$	9	5.2	32	23.5	11.7	13	12	10.4	11.4	7	13.52
$\beta 13$	8.4	5.5	11.6	11.2	7.7	5.6	9.8	8.4	13.1	8.5	8.98
$\beta 14$	10.8	4.9	4.1	5.4	6.5	**5.2**	11.5	9.5	7.2	6.5	7.2
$\beta 15$	8.6	4.9	3.6	6.4	7.5	6.4	4.3	3.3	7.4	4.6	5.7
$\beta 16$	7.3	5.6	**2.9**	4.9	7.9	5.4	3.8	4.3	**6.2**	8	**5.6**
$\beta 17$	**5.5**	5.8	3.3	**3.2**	6.1	10.7	**3.7**	**3.1**	10.3	6.3	5.78

performed on the results of the skeletonized silhouette using mouse cursor to obtain ground truth coordinates of SBPs. Note that the manual annotation of ground truth also involves some guesses of SBPs in cases where these points are not localized by skeletonization or not clearly visible to the human eye.

The location of every SBP obtained using the proposed framework with particle filter is compared with the ground truth in each frame of the video sequence. The overall accuracy of the proposed framework is defined by the average error in detecting each SBP, that is

$$Error(x_{avg}, y_{avg}) = \frac{\sum_{n=1}^{N} |G_n(x, y) - L_n(x, y)|}{N} \quad (42)$$

where $G_n(x, y)$ and $L_n(x, y)$ are respectively the coordinates of each SBP obtained from the ground truth and the proposed framework, and N is the total number of frames.

The average error in x and y coordinates of each SBP, i.e., Head (Hx, Hy), Front arm (FAx, FAy), Back arm (BAx, BAy), Left foot (LFx, LFy), and Right goot (RFx, RFy), in various activities *Act* (see Table I) performed by all subjects of both datasets is shown in Table V. For Jump-in-place-on-two-legs ($\beta 7$), Side ($\alpha 4$) and Walk ($\alpha 1$) of the Weizmann dataset (which have less lateral head movement), the x-coordinate head error is less than other activities whereas the y-coordinate head error is similar in all activities. The front and back arm points are occluded more than any other SBPs, hence they have greater errors. A common average error is obtained for the right and left foot because they are joined in Jump ($\alpha 5$), Jump-in-place-on-two-legs ($\beta 7$), One hand wave ($\beta 9$) and Two hand wave ($\beta 10$). The feet have smaller vertical movement than horizontal movement in consecutive frames in all activities, hence, the average y-coordinate error is less than the x-coordinate for both feet. For the MuHAVi dataset, the y-coordinate head error is less than the x-coordinate average error in all activities. The errors in the front and back arm points are also greater due to occlusion. The highest average error occurs in Collapse and Standup due to severe self occlusion of front and back arms. The right and left feet have similar average errors. The average Avg of five SBP errors per activity is presented in the last column of Table V.

TABLE VI
PARTICLE FILTER AND MOTION FLOW PREDICTION ERROR, RESPECTIVELY, DENOTED BY P AND M

Act	FAx_p	FAy_p	FAx_m	FAy_m	BAx_p	BAy_p	BAx_m	BAy_m
$\alpha 1$	7.7	12.9	**4.2**	**3.3**	9.23	19.4	**3.4**	6.4
$\alpha 2$	7.5	8.1	8.3	**3.3**	9.9	15.4	6.8	8.4
$\alpha 3$	8.5	9.4	4.8	6.3	13	9.2	4.1	**5.7**
$\alpha 4$	5.4	**8**	6.1	5	3.5	11	5	6.6
$\alpha 5$	8.2	14.2	4.1	6.2	6.9	**8.5**	5	6.5
$\beta 7$	**4.4**	12.2	7	6.1	**2.9**	10	4.5	6
Avg	6.9	10.8	**5.8**	**5**	7.1	12.2	**4.8**	**6.6**

Weizmann and MuHAVi datasets have $180 \times 144 = 25920$ pixels and $720 \times 576 = 414720$ pixels per frame, respectively. An overall average error of 5.02 and 7.8 pixels in location of SBPs on all activities for five SBPs, respectively, on two diverse datasets show that the proposed framework with arm prediction using particle filter is accurate and adaptable to data sets of different resolution.

2) *Localization Accuracy of Predicted Arm SBP:* It is vital to verify the accuracy of location of predicted arm SBP versus the ground truth. Table VI shows the error in the location using particle filter and motion flow in occlusion, where the average location error of predicted SBP is

$$ErrorPred(x_{avg}, y_{avg}) = \frac{\sum_{n=1}^{N} |G_n(x, y) - Pred_n(x, y)|}{N} \quad (43)$$

and $Pred_n(x, y)$ are the predicted SBP coordinates.

The particle filter and motion flow are compared for the arm prediction cyclic activities (see Table I), i.e., Walk ($\alpha 1$), Run ($\alpha 2$), Skip ($\alpha 3$), Side ($\alpha 4$), Jump ($\alpha 5$), and Jump-in-place-on-two-legs ($\beta 7$) of both datasets because it is the most occluded SBP. Table VI shows that particle filter and motion flow accurately predict arm point, i.e., close to ground truth location. The y-coordinate error of the front and back arm points using motion flow prediction are consistently smaller than those obtained using particle filter. The x-coordinate error is also smaller in most activities. Hence, motion flow outperforms particle filter which is demonstrated by smaller average Avg errors in all activities in Table VI. However, the lack of necessity for prior information makes particle filter the better choice for prediction. Results on Walk ($\alpha 1$) and Run ($\alpha 2$) activity of both data sets are shown in Table VI.

3) *Accuracy of Detected SBPs Versus Observed:* The accuracy of detection is evaluated in terms of precision (PR), recall (RC) and error (ER), that is

$$PR = \frac{\sum_{1}^{q} CT}{\sum_{1}^{q} DT}, \quad RC = \frac{\sum_{1}^{q} CT}{\sum_{1}^{q} OB} \tag{44}$$

$$ER = \frac{\sum_{1}^{q} DT - \sum_{1}^{q} CT}{\sum_{1}^{q} DT} \tag{45}$$

where DT and CT are respectively the number of detected and correctly detected SBPs. OB is the observed SBPs and q is the number of subjects. The number of detected SBPs includes misclassified SBPs which are manually counted by visual inspection on every frame of video sequence. The number of correctly detected SBPs is obtained by deducting misclassified SBPs from the number of detected SBPs.

The detection accuracy of five SBPs is computed by using the proposed framework first with no prediction and then with particle filter prediction. This demonstrates the impact of prediction on the performance of the framework. In Table VII for SBP detection with no prediction, observed (OB) SBPs are the manually counted visible SBP only with no guess work involved. For SBP detection with prediction in Table VII, observed (OB) SBPs is the manually counted visible SBP with guessed SBPs.

In Table VII, for no prediction, smaller recalls are obtained for Run ($\alpha 2$), Skip ($\alpha 3$), Jump ($\alpha 5$), and Two hand wave ($\beta 10$) that have abrupt human limb movement as compared to Walk ($\alpha 1$), Side ($\alpha 4$), Jump-in-place-on-two-legs ($\beta 7$), Bend ($\beta 8$), and One hand wave ($\beta 9$). The smallest recall and precision respectively occur in Run ($\alpha 2$) and One hand wave ($\beta 9$). The maximum recall and precision respectively occur in Side ($\alpha 4$) and One hand wave ($\beta 9$). The proposed framework with no prediction obtains an overall average $Avg\%$ recall and precision of 95.3% and 96.5%, respectively, for all activities of the Weizmann dataset. On the MuHAVi data set it obtains the smallest recall for Run ($\alpha 2$) but is robust in detecting SBPs in Walk ($\alpha 1$), Standup ($\beta 12$), Punch ($\beta 15$), Guard-to-kick ($\beta 16$), and Guard-to punch ($\beta 17$). In turn ($\alpha 6$), Collapse ($\beta 13$), and Kick ($\beta 14$) it is able to produce SBPs with reasonable accuracy. It has the least precision for complex movement such as Standup ($\beta 12$). It achieves an overall average $Avg\%$ recall and precision of 92.01% and 98.4%, respectively, for all activities of the MuHAVi dataset. The average error for all activities of the Weizmann and MuHAVi datasets computed using (45) are 3.5% and 1.9%, respectively.

In Table VII, for prediction, an overall 2.5% and 2.4% percentage increase in recall and precision, respectively, are obtained in cyclic actions of the Weizmann dataset using particle filter prediction. Specifically, the highest percentage increase of 7.3% in recall is achieved in Run ($\alpha 2$), which has the smallest recall with no prediction. For the MuHAVi dataset, particle filter prediction is only used for Walk ($\alpha 1$) and Run ($\alpha 2$) since they are cyclic actions. A percentage increase of 10.7% in recall is attained in Run ($\alpha 2$). There is a decrease in precision for both Walk ($\alpha 1$) and Run ($\alpha 2$), which suggests an increase in misclassified arm SBPs. However, more importantly particle filter prediction enhances the recall in all cyclic actions of both datasets. The proposed framework with prediction obtains an overall average $Avg\%$ recall and precision of 97.7% and 98.8%, respectively, for all activities of the Weizmann dataset. It achieves an overall average $Avg\%$ recall and precision of 94.2% and 95.7%, respectively, with prediction for all activities of MuHAVi dataset.

TABLE VII
PRECISION AND RECALL OF FIVE SBPS DETECTION OF PROPOSED FRAMEWORK

	Weizmann Data set									
	No prediction			Prediction			No prediction		Prediction	
Act	CT	OB	DT	CT	OB	DT	RC%	PR%	RC%	PR%
$\alpha 1$	2655	2768	2681	3134	3195	3160	95.9	99	**98.1**	99.2
$\alpha 2$	1468	1623	1532	1828	1885	1892	90.4	95.8	**97**	96.6
$\alpha 3$	1566	1664	1585	2108	2170	2127	94.1	98.8	**97.1**	99.1
$\alpha 4$	1726	1786	1726	2183	2220	2183	96.6	**100**	98.3	**100**
$\alpha 5$	1756	1877	1759	2220	2290	2223	93.5	99.8	**97**	99.9
$\beta 7$	2231	2271	2286	2654	2690	2709	98.2	97.6	98.7	98
$\beta 8$	3067	3195	3278	-	-	-	96	93.6	-	-
$\beta 9$	3265	3265	3555	-	-	-	**100**	91.8	-	-
$\beta 10$	2875	3120	3018	-	-	-	92.1	95.3	-	-
$\beta 11$	3157	3370	3201	-	-	-	93.7	98.6	-	-
Avg %	-	-	-	-	-	-	95.3	96.5	**97.7**	**98.8**
	MuHAVi Data set									
$\alpha 1$	1188	1231	1191	1326	1351	1502	96.2	99.8	**98.1**	88
$\alpha 2$	975	1198	985	1080	1198	1160	81.4	99	**90.1**	93.1
$\alpha 6$	868	1046	868	-	-	-	83	**100**	-	-
$\beta 12$	1431	1471	1505	-	-	-	97.4	95	-	-
$\beta 13$	1131	1306	1152	-	-	-	86.6	98.1	-	-
$\beta 14$	828	922	865	-	-	-	89.8	95.7	-	-
$\beta 15$	729	757	739	-	-	-	96.3	98.6	-	-
$\beta 16$	503	512	507	-	-	-	98.2	99.2	-	-
$\beta 17$	529	533	529	-	-	-	**99.2**	**100**	-	-
Avg %	-	-	-	-	-	-	92.01	98.4	**94.2**	95.7

The distance curve method in [1] and [6] is implemented to compare its SBP detection accuracy with the proposed framework. Based on Table VII, the total number of SBPs detected across all activities by the proposed framework is more than the skeletonized and CVHSP or star skeletonization. Hence, it is more consistent in generating stick figures of various activities.

4) *Comparative Evaluation of SBP Detection:* The performance of the proposed framework is compared with state of the art approaches, i.e., FS [11] and CBHM [8], with respect to a similar extent of occlusion and type of activity, respectively. The accuracy of FS to detect five body parts, i.e., head, arms, and feet, is evaluated in terms of the parts observed by the human eye. Five SBPs identified by the proposed framework correspond to the five body parts detected by First Sight. The

TABLE VIII
SBP DETECTION: PROPOSED VERSUS CBHM AND FS

		4 SBPs Accuracy				5 SBPs Error		
Classification		CBHM		Proposed		Proposed		FS
Occlusion	Act	RC%	PR%	RC%	PR%	ER%	Avg%	Avg%
Mild	$\alpha 1$	95.2	100	**97.4**	99.2	0.6		
Mild	$\alpha 2$	76.8	90.8	**97**	**97**	2.59		
Mild	$\alpha 4$	-	-	98.1	**100**	**0**		
Mild	$\alpha 6$	-	-	80.2	**100**	**0**		
Mild	$\beta 7$	-	-	98.3	97.5	2.4		
Mild	$\beta 14$	-	-	87.2	94.5	4.2		
Mild	$\beta 15$	-	-	95.5	98.3	1.35		
Mild	$\beta 16$	-	-	97.8	99	0.79		
Mild	$\beta 17$	-	-	**99.1**	**100**	**0**	**1.33**	15
Severe	$\alpha 5$	88.5	70.4	**97**	**99.8**	**0.17**		
Severe	$\beta 12$	99.7	82.6	95.9	**94.4**	4.91		
Severe	$\beta 13$	83.3	83	**85.7**	**97.6**	1.82		
Severe	$\beta 8$	-	-	97.6	92.2	6.43		
Severe	$\beta 9$	-	-	**100**	89.6	8.15		
Severe	$\beta 10$	-	-	91	94	4.73		
Severe	$\beta 11$	-	-	92.1	98.3	1.37		
Severe	$\alpha 3$	-	-	94.8	97.1	1.19	**3.59**	21

activities used by First Sight differ with respect to no, mild, and severe self occlusion. In the data sets for this paper, Walk ($\alpha 1$), Run ($\alpha 2$), Side ($\alpha 4$), Turn ($\alpha 6$), Jump-in-place-on-two-legs ($\beta 7$), Punch ($\beta 15$), Guard-to-kick ($\beta 16$), and Guard-to-punch ($\beta 17$) have mild self occlusion, whereas Skip ($\alpha 3$), Jump ($\alpha 5$), Bend ($\beta 8$), One hand wave ($\beta 9$), Two hand wave ($\beta 10$), Standup ($\beta 12$), and Collapse ($\beta 13$) have severe self occlusion. Table VIII shows the performances of the proposed framework and FS (as reported in [11]) on activities with mild and severe occlusion on all subjects of the Weizmann and MuHAVi datasets. In Table VIII, results on Walk ($\alpha 1$) and Run ($\alpha 2$) activity of both datasets are presented collectively. The average $Avg\%$ five SBPs error computed using (45) is clearly much less than FS.

Due to unavailability of the data set used by CBHM, Table VIII compares the average precision and recall of the proposed framework in detecting four SBPs (i.e., hands and feets) in similar activities with those of CBHM as reported in [8]. It shows that the proposed framework obtains better recall and precision than CBHM in Run ($\alpha 2$), Jump ($\alpha 5$), and Collapse ($\beta 13$). It also achieves a slightly better recall for Walk ($\alpha 1$). The recall obtained for Standup ($\beta 12$) is close to this approach, thus, overall the proposed framework performs better than CBHM.

C. Computational Complexity

The proposed framework runs in real time due to its computational simplicity. The computational time of the proposed framework implemented in Microsoft Visual Studio 2010 Express Edition environment with OpenCV 2.4.6 on an Intel (R) Core (TM) i7 processor working at 2.93 GHz with 4 GB RAM running Windows 7 operating system is measured using the computer system clock. The proposed framework labels SBPs in 0.031 s per image frame on the Weizmann dataset at 20–30 frames/s. It labels SBPs in 0.071 seconds per image frame on the MuHAVi dataset.

The convex hulll is computed using the Sklansky's algorithm [44] which has a computational complexity of $O(N)$, where N in the number of convex points. The contour moments algorithm is based on the Green theorem [38] which has a computational complexity of $O(L)$, where L is the length of the boundary of the object. The performance of the particle filter enhances with the increase in number of particles. It is formally $O(N \log N)$, however, it can be made $O(N)$ with minor modifications to the sampling procedure. In the proposed framework, the particle filter is initialized with 100 particles with a state vector constituting of four parameters. As a result its computational speed can be considered to be real time. This is similar to [45] where a 6–12 degree of freedom model with 100 particles run in real time.

VI. CONCLUSION

In this paper, an automated video-based human SBP labeling and tracking framework is presented. It employs IBMs based on anthropometry, kinesiology, and human vision inspired criteria to label SBPs. The classification of postures based on global angle is combined with the convexity hull and bounding rectangle to select the best IBM for labeling convex points as SBPs. Particle filter and motion flow are proposed for prediction in occlusion. Stick figures are generated by connecting SBPs. The results demonstrate that the proposed framework robustly locates, labels, and tracks SBPs in several actions on two datasets of low and high resolution. The results also show better it achieves better detection performance than the state of the art approaches. In future, manual counting of misclassified points can be automated and particle filter can be extended to predict SBPs for more actions.

REFERENCES

[1] C. F. Juang, C. M. Chang, J. R. Wu, and D. Lee, "Computer vision-based human body segmentation and posture estimation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans.*, vol. 39, no. 1, pp. 119–133, Jan. 2009.
[2] M. Goffredo, M. Schmid, S. Conforto, M. Carli, A. Neri, and T. D'Alessio, "Markerless human motion analysis in Gauss-Laguerre transform domain: An application to sit-to-stand in young and elderly people," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 207–16, Mar. 2009.
[3] J. Candamo, M. Shreve, D. Goldgof, D. Sapper, and R. Kasturi, "Understanding transit scenes: A survey on human behavior-recognition algorithms," *IEEE Trans. Intell. Transp. Syst.*, vol. 11, no. 1, pp. 206–224, Mar. 2010.
[4] R. Cucchiara, A. Prati, R. Vezzani, and R. Emilia, "A multi-camera vision system for fall detection and alarm generation," *Expert Syst.*, vol. 24, no. 5, pp. 334–345, 2007.
[5] G. Bradski and A. Kaehler., *Learning OpenCV Computer Vision With the OpenCV Library*. Sebastopol, CA, USA: O'Reilly Media, Sep. 2008.
[6] H. Fujiyoshi, A. J. Lipton, and T. Kanade, "Real-time human motion analysis by image skeletonization," *IEICE Trans. Inf. Syst. E Ser. D.*, vol. 87, no. 1, pp. 113–120, 2004.
[7] C. C. Yu, J. N. Hwang, G. F. Ho, and C. H. Hsieh, "Automatic human body tracking and modeling from monocular video sequences," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, vol. 1. Apr. 2007, pp. 917–920.
[8] C. C. Yu, Y. N. Chen, H. Y. Cheng, J. N. Hwang, and K. C. Fan, "Connectivity based human body modeling from monocular camera," *J. Inf. Sci. Eng.*, vol. 26, no. 2, pp. 363–377, Dec. 2010.

[9] W. Lao, J. Han, and P. H. With, "Fast detection and modeling of human-body parts from monocular video," in *Proc. 5th Int. Conf. Articulated Motion Deformable Objects*, 2008, pp. 380–389.
[10] W. Lao, J. Han, and P. H. N. de With, "Flexible human behavior analysis framework for video surveillance applications," *Int. J. Digit. Multimedia. Broadcast.*, vol. 2010, pp. 1–10, Jan. 2010.
[11] M. K. Leung and Y. H. Yang, "First sight: A human body outline labeling system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 4, pp. 359–377, Apr. 1995.
[12] C. Barrón and I. A. Kakadiaris, "On the improvement of anthropometry and pose estimation from a single uncalibrated image," *Mach. Vision Appl.*, vol. 14, no. 4, pp. 229–236, 2003.
[13] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: Real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, Aug. 2000.
[14] C. F. Juang and C. M. Chang, "Human body posture classification by a neural fuzzy network and home care system application," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans.*, vol. 37, no. 6, pp. 984–994, Nov. 2007.
[15] F. Huo, E. Hendriks, P. Paclik, and A. H. J. Oomes, "Markerless human motion capture and pose recognition," in *Proc. Image Anal. Multimedia Interactive Serv.*, May 2009, pp. 13–16.
[16] K. Takahashi and T. Kodama, "Remarks on simple motion capture using heuristic rules and Monte Carlo filter," in *Proc. Int. Conf. Image Graph.*, Sep. 2009, pp. 808–813.
[17] L. Huang, S. Tang, Y. Zhang, S. Lian, and S. Lin, "Robust human body segmentation based on part appearance and spatial constraint," *Neurocomputing*, vol. 118, pp. 191–202, Oct. 2013.
[18] L. Ladicky, P. H. S. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2013, pp. 3578–3585.
[19] M. Dantone, J. Gall, C. Leistner, and L. van Gool, "Human pose estimation from still images using body parts dependent joint regressors," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2013, pp. 3041–3048.
[20] H. Bhaskar, L. Mihaylova, and S. Maskell, "Articulated human body parts detection based on cluster background subtraction and foreground matching," *Neurocomputing*, vol. 100, pp. 58–73, Jan. 2013.
[21] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, *et al.*, "Real-time human pose recognition in parts from single depth images," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, Jun. 2011, pp. 1297–1304.
[22] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, "Multi-view body part recognition with random forests," in *Proc. IEEE Brit. Mach. Vision Conf.*, Sep. 2013, p. 11.
[23] Z. Li and D. Kulic, "Local shape context based real-time endpoint body part detection and identification from depth images," in *Proc. Int. Conf. Comput. Robot Vision*, 2011, pp. 219–226.
[24] A. Gritai, Y. Sheikh, C. Rao, and M. Shah, "Matching trajectories of anatomical landmarks under viewpoint, anthropometric and temporal transforms," *Int. J. Comput. Vision*, vol. 84, no. 3, pp. 325–343, Sep. 2009.
[25] NASA-STD-3000. (1995). Anthropometry and biomechanics [Online]. Available: http://msis.jsc.nasa.gov/sections/section03.htm
[26] R. Easterby, K. Kroemer, and D. B. Chaffin., *Anthropometry and Biomechanics*. New York, NY, USA: Plenum Press, 2010.
[27] C. Barrón and I. A. Kakadiaris, "Estimating anthropometry and pose from a single uncalibrated image," *Comput. Vis. Image Underst.*, vol. 81, no. 3, pp. 269–284, Mar. 2001.
[28] C. Benabdelkader and Y. Yacoob, "Statistical estimation of human anthropometry from a single uncalibrated image," in *Proc. Workshop Biometric Authentication*, 2008, pp. 200–220.
[29] I. F. Leong, J. J. Fang, and M. J. Tsai, "Automatic body feature extraction from a marker-less scanned human body," *Comput. Aided Design*, vol. 39, no. 7, pp. 568–582, 2007.
[30] B. Bogin and M. I. Varela-Silva, "Leg length, body proportion, and health: A review with a note on beauty," *Int. J. Environ. Res. Public Health*, vol. 7, no. 3, pp. 1047–1075, 2010.
[31] D. Winter, *Biomechanics and Motor Control of Human Movement*. New York, NY, USA: Wiley, 2009.
[32] A. E. Chapman, *Biomechanical Analysis of Fundamental Human Movement*. Champaign, IL, USA: Human Kinetics, 2008.
[33] J. Hamill and K. M. Knutzen, *Biomechanical Basis of Human Movement*. Alphen aan den Rijn, The Netherlands: Wolters Kluwer, 2009.
[34] N. Hamilton, W. Weimar, and K. Luttgens, *KINESIOLOGY Scientific Basis of Human Motion*. New York, NY, USA: McGraw-Hill, 2011.
[35] H. Wang and K. Kosuge, "Control of a robot dancer for enhancing haptic human-robot interaction in Waltz," *IEEE Trans. Haptics*, vol. 5, no. 3, pp. 264–273, Jul. 2012.
[36] T. Kwon and J. Hodgins, "Control systems for human running using an inverted pendulum model and a reference motion capture sequence," in *Proc. Eur. Symp. Comput. Animation*, 2010, pp. 129–138.
[37] I. D. Loram, S. M. Kelly, and M. Lakie, "Human balancing of an inverted pendulum: Is sway size controlled by ankle impedance?" *J. Physiol.*, vol. 532, no. 3, pp. 879–891, 2001.
[38] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Fall detection from human shape and motion history using video surveillance," in *Proc. 21st Int. Conf. Adv. Inf. Netw. Appl. Workshops*, vol. 2. 2007, pp. 875–880.
[39] Y. L. Lin and M. J. J. Wang, "Automated body feature extraction from 2D images," *Expert Syst. with Applicat.*, vol. 38, no. 3, pp. 2585–2591, 2011.
[40] H. Freeman, "On the classification of line drawing data," in *Models for the Perception of Speech and Visual Form*. E. W. Wather Dunn, Ed. Cambridge, MA, USA: MIT Press, 1967, pp. 408–412.
[41] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.
[42] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *Proc. IEEE Int. Conf. Adv. Video and Signal Based Surveillance*, Sep. 2010, pp. 48–55.
[43] H. Foroughi, M. Alishah, H. Pourreza, and M. Shahinfar, "Distinguishing fall activities using human shape characteristics," in *Technological Developments in Education and Automation*. M. Iskander, V. Kapila, and M. A. Karim, Eds. Amsteradam, The Netherlands: Springer, 2010, pp. 523–528.
[44] K. Homma and E. Takenaka, "An image processing method for feature extraction of space-occupying lesions," *J. Nucl. Med.*, vol. 26, no. 12, pp. 1472–1477, 1985.
[45] M. Isard and A. Blake, "Condensation: Conditional density propagation for visual tracking," *Int. J. Comput. Vision.*, vol. 29, no. 1, pp. 5–28, 1998.
[46] R. Telea and J. J. V. Wijk, "An augmented fast marching method for computing skeletons and centerlines," in *Proc. Symp. Data Vis.*, 2002, pp. 251–259.

Faisal Azhar (S'14) received the B.Sc. degree in bio-medical engineering and the M.S. degree in electrical engineering, respectively from Sir Syed University, Karachi, Pakistan and the National University of Sciences and Technology, Karachi, and is currently pursing the Ph.D. degree in human motion analysis and tracking for activity recognition with the University of Warwick, Coventry, U.K.

His career projects include super resolution medical image enhancement, iris recognition, LEGO robot control system, and human activity recognition. His current research interests include computer vision and machine learning.

Tardi Tjahjadi (SM'02) received the B.Sc. degree in mechanical engineering from University College London, London, U.K., in 1980 and the M.Sc. degree in management sciences and the Ph.D. degree in total technology from UMIST, Manchester, U.K., in 1981 and 1984, respectively.

He is currently a Reader with the University of Warwick, Coventry, U.K. His current research interests include image processing and computer vision.

Hierarchical relaxed partitioning system for Activity Recognition

Faisal Azhar, *Student Member, IEEE* and Chang-Tsun Li, *Senior Member, IEEE*

***Abstract*—A hierarchical relaxed partitioning system (HRPS) method is proposed for recognizing similar activities which have a feature space with multiple overlaps. Two feature descriptors are built from the human motion analysis of a 2D stick figure to represent cyclic and non-cyclic activities. The HRPS first discerns the pure and impure activities, i.e., with no overlaps and multiple overlaps in the feature space respectively, then tackles the multiple overlaps problem of the impure activities via an innovative majority voting scheme. The results show that the proposed method robustly recognizes various activities of two different resolution, i.e., low and high (with different views), data sets. The advantage of HRPS lies in the real-time speed, ease of implementation and extension, and non-intensive training.**

***Index Terms*—Hierarchical Relaxed Partition, Decision Tree, Model, Activity Recognition**

I. Introduction

Human activity recognition is important due to potential applications in video surveillance, assisted living, animation etc [1] [2]. In general, a standard activity recognition framework consists of the feature extraction, feature selection (dimension reduction) and pattern classification. The feature extraction can be broadly categorized into the holistic (shape or optical flow) [3]–[6], local feature (descriptors of local regions) [7]–[10] and model-based (prior model) or model-free (no prior model) approaches. Techniques such as Principal component analysis (PCA) [11] or Linear Discriminant Analysis (LDA) [12] are commonly used to select the most prominent features. Decision tree (DT) [3] or Support Vector Machines (SVMs) [2] are used for efficient classification.

The current state-of-the-art human activity recognition method varies with respect to application scenario as each method has been designed and verified for data sets containing different challenges such as similar activities, industrial environment, illumination variation, varying clothing, complex backgrounds, multiple actors, person-to-person interaction, human object interaction, multiple views etc. (see [13] for details on datasets). Also, it has been noted in literature [14] that human activity recognition methods have different performances on different data sets. The apparent reason for this lies in the feature extraction approach, i.e., holistic, local feature and model-based/model-free, and the different characteristics of the activities in the data sets [14]. The local features approach that extract the neighbourhood information of the regions or interest points focus more on the local motion than on the figure shape. Hence, it is suitable for activities with more intra-class dissimilarity in the shape of figures. In contrast, the holistic and model-based/model-free approach are focused on figure shape characteristics which makes them suitable for activities with more inter-class similarity in the local motion, i.e., similar activities such as Walk, Run etc.

Recognizing similar activities still remains a challenge (see Section II). The local feature and holistic approaches are computationally expensive and require intensive training while the model-based/model-free approach is efficient but less accurate. Therefore, the robust and efficient implicit body model based approach for significant body point (SBP) detection described in [15] is used for feature extraction. In this context, the work in [16] that extracts the leg frequency and torso inclination is extended to determine two more features, i.e., the leg power and torso power. Also, the SBP detection method is augmented to extract features (similar to [6]) that extract variations in the movement of different body parts at different directions, i.e., up, down, right, and left, during an activity. As in [6] PCA or LDA is not used as we extract less than 15 features. These features are used to create two feature descriptors.

For efficient classification, mostly researchers use off-the-shelve classifier such as SVM and DT but with a trade-off of performance, e.g., SVM struggles due to the lack of generalized information, i.e., each test activity is compared with the training activity of one subject [6]. On the other hand DT imposes hard constraint that lead to separation problems when the number of categories increases or when categories are similar, i.e., a lack of clear separation boundary [17]. To achieve high accuracy while being fast the Relaxed Hierarchy (RH) method in [17] uses relaxed constraint, i.e., postpone decisions on confusing classes, to tackle the increased number of categories but still remains prone to accurately discerning similar categories. The Hierarchical Strategy (HS) method in [18] uses the RH and group together easily confused classes to improve the classification performance. RH and HS has only been applied to the spatial domain. Hierarchical methods [19], [20] are also used at lower levels for feature-wise classification. Note, however, similar to [17] this work focuses on building high-level class hierarchies and look into the problem of class-wise partitioning.

In order to recognize similar human activities efficiently and accurately, we propose a hierarchical relaxed partitioning system (HRPS) (see Section III for details). This is a system that classifies and organizes activities in a hierarchical manner according to their type, i.e., pure activities (easily separable) and impure activities (easily confused). Subsequently, it ap-

Manuscript received April 2015. This work was supported by Warwick PostGraduate Research Scholarship.

The authors are with the Department of Computer Science, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK. E-mail: {faisal.azhar, c-t.li}@warwick.ac.uk, faisal108988@gmail.com.

plies relaxed partitioning to all the easily confused activities by postponing the decisions on them until the last level of the hierarchy, where they are labelled by using a novel majority voting scheme (MVS). As opposed to a conventional multi-class classifier as in [18] that can distinguish between only two similar activities, i.e., two classes overlap simultaneously, the proposed MVS is able to discern between three or more similar activities, i.e., three classes overlap concurrently. Thus, making the HRPS more robust and suitable for identifying activities in real world scenarios.

The proposed method is distinguished from our work in [15], for significant body point labelling and tracking, in the following respects: (a) activity recognition is addressed in this paper, (b) the work in [15] and [16] is augmented to built two feature descriptors, (c) the HRPS with the majority voting scheme is proposed to recognize similar activities.

This paper is organized as follows. Section II reviews related methods. Section III and Section IV present the foundation of HRPS and its application to activity recognition, respectively. Experiments are shown in Section V.

II. Literature review

Several human activity recognition methods, e.g., [3], [7], [8], [14], [21]–[25] verified on the benchmark data sets (see [13] for data sets) struggle in correctly classifying similar activities of the Weizmann data set. The methods [3], [5], [6], [10] that are able to correctly classify similar activities of the Weizmann data set are either computationally expensive or require intensive training or need to learn a large set of features. These methods require tuning of parameters with respect to the data set. Therefore, they require extensive re-training for new activities. Some methods [5], [7], [25] require more number of frames (approximately 100 to 200 frames) for training, thus duplicate or up-sample the training data.

A. Holistic and local feature approaches

In [3], a shape-motion prototype-based method is presented for action recognition. In the training phase, it extracts shape-motion descriptors to learn action prototypes which are represented via a binary hierarchical tree. In the testing phase, the shape-motion descriptor is used to recognize human actions via tree-based prototype matching and look-up table indexing. Both shape and motion cues are required to recognise similar activities accurately. In [5], a learning-based method is proposed which uses time series of optical flow motion features for human action recognition. In the learning stage, the optical flow motion features extracted from the given action sequences are concatenated to construct motion curves. Each human action is represented by a cluster of motion curves which are clustered by using a Gaussian mixture model. In the recognition stage, the cluster of optical flow motion curves of the probe sequence is matched to the learned motion curves using a similarity function. In [6] the optical flow and random sample consensus methods are used to localize the subject. Next, it extracts a feature vector that contain variations in the movement of different body parts at different directions during an activity. Euclidean distance or SVM is used with the feature vector for action recognition. In [10] the locality preserving projection method (that learns a projection onto a low dimensional space while optimally preserving the neighbourhood structure) is supervised to recognize similar activities by not ignoring the local information of the data. These methods are either computationally expensive or require intensive training or tuning of multiple parameter on a data set.

In [7], the kinematic features from the optical flow extracted from videos are converted into kinematic modes using principal component analysis. These kinematic modes are then used in a bag of kinematic mode representation with a nearest neighbour classifier for human action recognition. It has high computational cost, requires intensive training and confuses similar activities. In [8], videos are represented as word $\times$ time tables and the extracted temporal patterns are used with supervised time-sensitive topic models for action recognition. It also confuses similar activities.

B. Model-free and Model-based approaches

A star is a shape that is formed by connecting the centre of mass of a human silhouette contour to the extreme boundary points. The method in [16] creates a one-star by using a local maximum on the distance curve of the human contour to locate the SBPs which are at the extremities. It uses two motion features, i.e., leg frequencies and torso angles, to recognize only the Walk and Run activities. A two star method [26] extends [16] by adding the highest contour point as the second star. It uses a 5D feature descriptor with a hidden Markov model (HMM) to detect the fence climbing activity. The method in [24] extends [26] by using the medial axis [27] to generate the junction points from which variable star models are constructed. It is compared with [16] and [26] on the fence climbing activity, and evaluated on the Weizmann data set. In [28], multiple cues such as the skin colour, principal and minor axes of the human body, the relative distances between convex points, convex point curvature, etc., are used to enhance the method in [16] for the task of posture estimation. It does not provide quantitative results, and uses a non-standard and non-publicly available data set. Thus, it requires extensive further work to validate and apply it to activity recognition. The method in [25] assumes that SBPs are given and uses the chaotic invariant for activity recognition on the Weizmann data set. It uses the trajectories of SBPs to reconstruct a phase space, and applies the properties of this phase space such as the Lyapunov exponent, correlation integral and dimension, to construct a feature vector, for activity recognition. The above-described distance curve based methods are sensitive to the silhouette contour, occlusion, resolution, etc., which affects their accuracy for activity recognition. The method in [24] and [25] confuse similar activities while only two features of the method in [16] are not sufficient for recognizing more than two similar activities.

The method in [29] uses the Poisson equation to obtain the torso, and negative minimum curvature to locate the SBPs. An 8D feature descriptor from the articulated model is used with the HMM to recognize six activities. In [30], the dominant points along the convex hull of a silhouette contour are used

with the body ratio, appearance, etc., to fit a predefined model. It is extended in [31] for activity recognition. These methods are evaluated on non-standard and publically unavailable data sets. The method in [32] uses the convex hull to identify the SBPs. However, it is designed to be used for surveillance purposes. In [15] implicit body models are used with the convex hull of a human contour to label SBPs. It tracks the SBPs by using a variant of the particle filter. This method works in real-time by fitting the knowledge from the implicit body models. It outperforms most of the cutting edge methods that use the distance curve method. Thus, we are motivated to extend and apply it for activity recognition.

III. Foundation of Proposed Method

A DT learns from a data and features the best class separation based on an optimization criteria. Let $p(m|t)$ denote the fraction of samples belonging to a class m at a given node t. Then, for M number of classes, $Entropy(t) = -\sum_{m=0}^{M-1} p(i|t) \log_2 p(m|t)$, can be used as an optimization criteria to determine the best split at each node by measuring the class distribution before and after the split. Techniques such as pruning that optimizes tree depth (leafiness) by merging leaves on the same tree branch can then be used to avoid over-fitting. Random Forest (RDF) is an ensemble learning method that generates many DT classifiers and aggregate their result to avoid over-fitting issue of DT and improve classification performance [33]. Methods like DT and RDF assume that at each node the feature-space can be partitioned into disjoint subspaces, however as mentioned in [17] this does not hold when there are similar classes or when there are large number of classes. In this case finding a feature-space partitioning that reflects the class-set partitioning is difficult as observed in [17]. Therefore, similar to [17], [18] the goal of this work is to establish a class hierarchy and then train a classifier such as simple binary classifier at each node of the class hierarchy to perform efficient and accurate classification. This allows us to define different set of rules for classifying different types of activities. This is important as different feature sets are useful for discerning different types of activities [34].

In this context, a class hierarchy is created and at each node a binary decision rule is learned that ignores easily confused categories. At the bottom node of the hierarchy a MVS is used to perform decisions on easily confused categories. Let us demonstrate the concept of creating a HRPS using a simple example with three overlapping classes that represent similar categories as shown in Fig. 1(a). It can be seen from Fig. 1(a) that it is not possible to clearly distinguish between only two overlapping classes by using the RH method as it assumes that only two classes overlap simultaneously. This is because now the overlap is among three classes concurrently, i,e., the overlap between the two classes A and B also contain some overlap with the third class C. Similar phenomena occurs for B and C, and A and C classes. In addition, a combined overlap occurs, i.e, $A \cap B \cap C \neq \emptyset$. Hence, the RH method is not capable of tackling the multiple overlaps class separation problem.

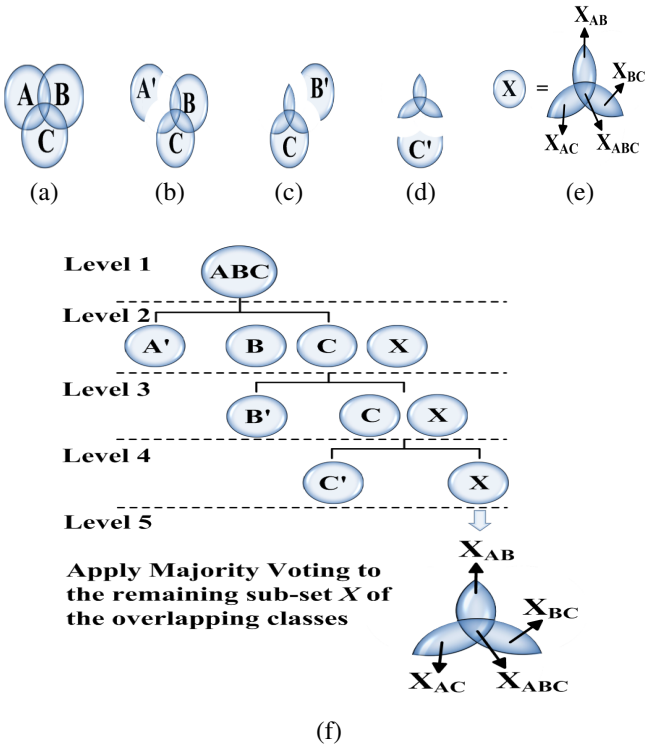


Fig. 1. (a) Example of three classes to illustrate multiple overlaps class separation problem, (b)-(e) Hierarchical relaxed partitioning system: (b), (c) and (d) Partition non-overlapping samples from class A, B and C respectively, (e) Remaining overlapping samples of all the three classes discerned using the majority voting scheme (see Section IV-B for details), and (f) the corresponding class hierarchy structure.

The proposed HRPS method addresses this deficiency in the RH method by splitting the set of classes $K = A' \cup B' \cup C' \cup X$, where $X = \{X_{AB} \cup X_{BC} \cup X_{AC}\}$ and $X_{AB} = A \cap B - A \cap B \cap C$, $X_{BC} = B \cap C - A \cap B \cap C$, $X_{AC} = A \cap C - A \cap B \cap C$ and $X_{ABC} = A \cap B \cap C$. X contains samples from two or more overlapping classes. First, at each level of the hierarchy the clearly separable samples of each class are partitioned into the A' or B' or C' as shown in Fig. 1(b)-(d).

$$A' = A - X_{AB} - X_{AC} - X_{ABC} \quad (1)$$

$$B' = B - X_{AB} - X_{BC} - X_{ABC} \quad (2)$$

$$C' = C - X_{AC} - X_{BC} - X_{ABC}. \quad (3)$$

Next, the overlapping samples of each class as shown in Fig. 1(e) are partitioned into A or B or C via a majority voting scheme (see Section IV-B). The class hierarchy structure for HRPS method is shown in Fig. 1(f). Note that at each level one class is partitioned from the remaining group of easily confused classes [1] [18].

IV. HRPS for Activity Recognition

We present HRPS for the Weizmann data set [35] containing multiple similar activities such as Walk, Run, Side, Skip, etc. that are easily confused by the activity recognition methods in the literature. HRPS for the Multi-camera Human Action Video (MuHAVi) data set [36] containing similar activities

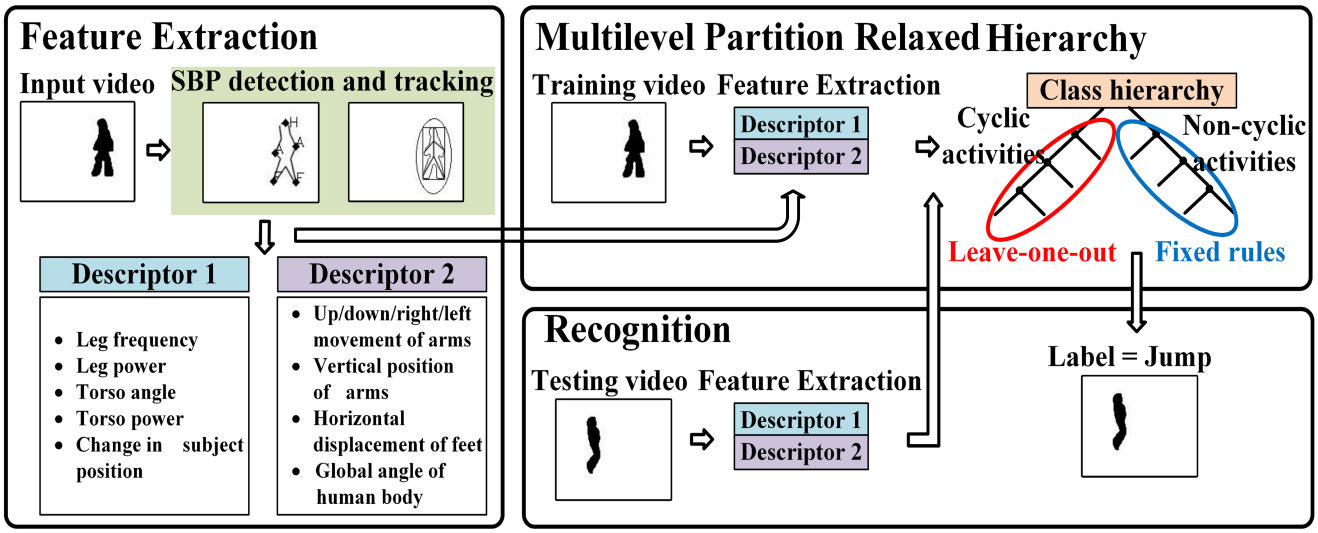
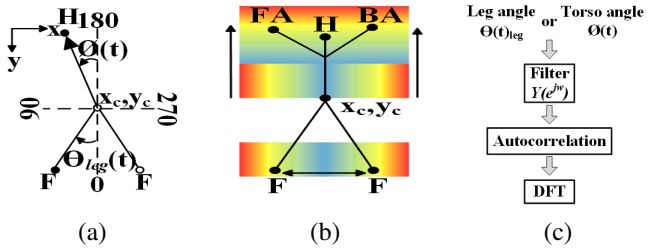


Fig. 2. The main components and work flow of the proposed human activity recognition.



(a) (b) (c)

Fig. 3. Feature extraction. (a) 2D stick figure analysis for cyclic activities, (b) The upper and lower body analysis based on the arm and feet movement, and (c) Process of acquiring D_1 for the cyclic activities. The SBPs labelled as Head (H), Front Arm (FA), Back Arm (BA) and Feet (F).

e.g., walk , run, turn, etc., is also described in order to establish its generality, i.e., adaptability to work on a different data set. The work flow of the proposed activity recognition is shown in Fig. 2.

A. Feature extraction

Distinguishing between the cyclic and non-cyclic activities is vital for activity recognition [37]. Thus, we augment our earlier work in [15] to build two feature descriptors $D_{i,\ i=1,2}$. The 2D stick figure shown in Fig. 3 (a) is used to describe

$$D_1 = [V_1\ \ V_2\ \ V_3\ \ V_4\ \ V_5] \tag{4}$$

for cyclic activities, while the 2D stick figure shown in Fig. 3 (b) is utilized to build

$$D_2 = [V_6\ \ V_7\ \ V_8\ \ V_9\ \ V_{10}\ \ V_{11}\ \ V_{12}\ \ V_{13}] \tag{5}$$

for non-cyclic activities. The $V_{i,\ i=1,2,...12}$ represents the feature elements of the descriptors. In Fig. 3, the SBPs are labelled as the Head (H), Front Arm (FA), Back Arm (BA) and Feet (F). Each SBP abbreviation can be considered as a vector which has a 2D position, e.g, $FA = (x^{FA}, y^{FA})$, $F = (x^F, y^F)$. Here, the superscripts denote the abbreviations of SBP.

The 2D stick figure motion analysis method in [16] uses two motion based features, i.e., the leg power and torso inclination angle, to discern between the Walk and Run activities. This method is suitable for only classifying the cyclic activities with less inter-class similarity, i.e., the activities are not similar to each other. Therefore, we propose two more features, i.e., the torso angle and torso power, to strengthen the method in [16]. Given the global angle from contour moments $V_6 = \theta(t)$ at time t, centre (x_c, y_c), and SBPs from [15], we extend the method in [16] to acquire D_1 which contains four motion based features, i.e., the leg cyclic frequency (V_1) and leg power (V_2), and the torso inclination angle $V_3 = \phi(t) = |90 - (\theta(t)3.14/180)|$ and torso power V_4 for the cyclic activities. The foot point $x^F > x_c$ is used for computing

$$\theta_{leg}(t) = \tan^{-1}(\frac{x^F - x_c}{y^F - y_c}). \tag{6}$$

The computed torso angle $V_3 = \phi(t)$ and leg angle $\theta(t)_{leg}$ are converted into radians. A highpass digital filter $Y(e^{jw})$ is applied to $\theta(t)_{leg}$.

$$Y(e^{jw}) = b(1) - b(2)e^{-jw} \tag{7}$$

Here, $b(1) = 1, b(2) = -0.9$ as in [16]. The filtered leg angles $\theta(t)_{leg}$ are then autocorrelated in order to emphasise the major cyclic components. The discrete Fourier transform (DFT) is applied to the autocorrelated leg angles to quantify the leg frequency V_1 and magnitude expressed as leg power V_2 in decibels [16] as shown in Fig. 3(c). The proposed activity recognition system also applies the high pass digital filter $Y(e^{jw})$ to the torso angle V_3 (in radians) in order to remove the low frequency components in contrast to [16] where this filter is only applied to the leg angle $\theta(t)_{leg}$. Next, the autocorrelation and DFT steps in Fig. 3(c) are performed on the filtered torso angle to compute a new feature, i.e., the torso magnitude expressed as torso power V_4 in decibels. The change in direction of movement or position is incorporated as

$$V_5 = min(x_c^{t+1} - x_c^t) \tag{8}$$

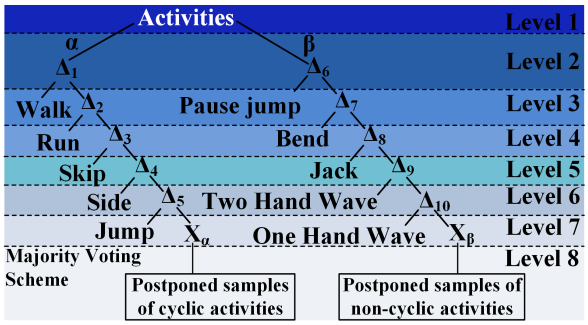


Fig. 4. Hierarchical relaxed partitioning system for the Weizmann data set. $\Delta_{i,\ i=1,2,..10}$ are the decision rules, and X_α and X_β are the unassigned impure cyclic and non-cyclic activities, respectively, with significant multiple overlaps.

TABLE I
ACRONYMS FOR ACTIVITIES.

Type	Activities (α)
1	Walk
2	Run
3	Skip
4	Side
5	Jump
6	Turn

Type	Activities (β)
7	Jump-in-place-on-Two-Legs/Pause Jump
8	Bend
9	One Hand Wave
10	Two Hand Wave
11	Jack
12	Standup
13	Collapse
14	Kick
15	Punch
16	Guard-to-Kick
17	Guard-to-Punch

$\forall\ t \in 1, N-1$, where N is the total number of frames, min gives the minimum value. A positive and negative value of V_5 respectively indicate whether subject moved in the same direction or changed direction (turn around) of movement during an activity.

The feature descriptor D_2 characterises the upper body (torso and arms) and lower body (legs) movements as a proportion of the mean height μ_h at different directions during an activity as shown in Fig. 3 (b) for the non-cyclic activities. The inter-frame displacement (movement) of the front and back arms are described as

$$V_7 = max(|x_{t+1}^{FA} - x_t^{FA}|)/\mu_h, \quad V_8 = max(|y_{t+1}^{FA} - y_t^{FA}|)/\mu_h \tag{9}$$

$$V_9 = max(|x_{t+1}^{BA} - x_t^{BA}|)/\mu_h, \quad V_{10} = max(|y_{t+1}^{BA} - y_t^{BA}|)/\mu_h \tag{10}$$

$\forall\ t \in 1, N-1$,max gives the maximum value. The features V_7, V_8, V_9, and V_{10} do not contain information with respect to the actual positioning of the front and back arm SBPs, i,e., where the arm displacement is being taken place. This information is represented as

$$V_{11} = min(y_t^{FA}), \quad V_{12} = min(y_t^{BA}), \ \forall\ t \in 1, N \tag{11}$$

which uses the vertical position of the front and back arms to represent their maximum height (as the minimum y location of the front and back arms). The variation in the lower body movement due to the leg can be represented by computing the maximum inter-frame horizontal displacement between the two feet as

$$V_{13} = max(|x_{t+1}^{F} - x_t^{F}|)/\mu_h, \ \forall\ t \in 1, N-1. \tag{12}$$

B. Classification: HRPS for the Weizmann data set

The Weizmann data set contain ten activities, i.e., the Walk ($\alpha 1$), Run ($\alpha 2$), Skip ($\alpha 3$), Side ($\alpha 4$), Jump ($\alpha 5$), Jump-in-place-on-two-legs or Pause Jump ($\beta 7$), Bend ($\beta 8$), One Hand Wave ($\beta 9$), Two Hand Wave ($\beta 10$) and Jack ($\beta 11$). In [38], a binary decision tree splits the activities into still and moving categories at the root node in order to obtain better classification. Therefore, an expert knowledge motivated from [38] is added at the root node level 1 to automatically split the above-mentioned ten activities in two groups, i.e., significant translation (α) and no significant translation (β) by using

$$\begin{aligned} \alpha &= 0.25I_w > x_c \text{ or } x_c > 0.75I_w \\ \beta &= 0.25I_w < x_c \text{ or } x_c < 0.75I_w \end{aligned} \tag{13}$$

as shown in level 2 of Fig. 4. I_w and I_h are the frame width and frame height, respectively. Thus, most cyclic activities, i.e., the Walk ($\alpha 1$), Run ($\alpha 2$), Skip ($\alpha 3$), Side ($\alpha 4$) and Jump ($\alpha 5$), which have significant translation of the subject and repetitive nature are grouped together under α. The activities, i.e., the Pause Jump ($\beta 7$), Bend ($\beta 8$), One Hand Wave ($\beta 9$), Two Hand Wave ($\beta 10$) and Jack ($\beta 11$), which have no significant translation of the subject are grouped under β. A HRPS with 8 levels is created with decision rules $\Delta_{i,\ i=1,2,...10}$ as shown in Fig. 4. The decision rules $\Delta_{i,\ i=1,2,...5}$ for cyclic activities are learned by using Algorithm. IV.1 on the training data set that contains the activities performed by eight subjects. The last subject is used as the testing data set in a leave-one-person-out cross validation approach to determine the performance of the HRPS for cyclic activities. The Algorithm. IV.1 postpone decisions on those samples of an activity that are closer to the samples of all the remaining activities by updating the decision rules $\Delta_{i,\ i=1,2,...5}$ by using variable adjustment κ. In [15], SBPs were accurately detected by using implicit body models (IBMs) that are based on the human kinesiology and anthropometric studies, and observed human body characteristics. This inspired us to define decision rules $\Delta_{i,\ i=6,8,...10}$ that are fixed based on the human kinesiology (torso flexion or extension V_6) [39] and anthropometric studies (upper body motion V_7, V_8, V_9, V_{10} and leg motion V_{13}) [40], and individual arm location V_{11} and V_{12}), observed human body characteristics and experimental cues for non-cyclic activities. The Pause Jump ($\beta 7$) is a cyclic activity with no significant translation but has repetitive nature. Thus, it is first separated using V_6 from the non-cyclic activities, i.e., Bend ($\beta 8$), One Hand Wave ($\beta 9$), Two Hand Wave ($\beta 10$), Jack ($\beta 11$), and then dealt with in a similar manner as the remaining cyclic activities for classification. This knowledge will assure an increase in the accuracy and reliability of the activity classification.

$$\Delta_6 = \begin{cases} \beta 7 & \text{if } |90 - V_6| < 9 \\ \Delta 7 & \text{Otherwise.} \end{cases} \tag{14}$$

Algorithm IV.1: PARTITION LEARNING ALGORITHM(D_1)

Input: Training sequences $S_1, ..., S_M$
Corresponding labels $y_1, ..., y_M$
Feature descriptor $D_1 = [V_1\ V_2\ V_3\ V_4\ V_5]$
Output: Decision rules $\Delta_{i,\ i=1,2,...5}$

For each activity, determine the mean μ_j and standard deviation σ_j of feature elements $V_{j,\ j=1,...,5}$ from K training subjects/samples as

$$\mu_j = \sum_{k=1}^{K} V_j^k/K \quad , \ \sigma_j = \sqrt{1/K \sum_{k=1}^{K} (V_j^k - \mu_j)^2}.$$

Learn decision rules as one standard deviation on either side of the mean

$$\Delta_{i,\ i=1,2,...5} = \mu_j - \sigma_j < V_j < \mu_j + \sigma_j.$$

Update decision rules by using a variable adjustment κ to separate clearly separable samples, i.e., pure samples, of an activity from the samples of all the remaining activities

$$\Delta_{i,\ i=1,2,...5} = \mu_j - \sigma_j + \kappa < V_j < \mu_j + \sigma_j + \kappa$$

Accumulate impure samples of an activity that are closer to the samples of all the remaining activities in X_α.

A full flexion of the vertebra in the Bend ($\beta 8$) activity causes a large increase in the torso angle [39]. Based on the experimental observation in Section V-A most training subjects have a torso angle variation greater than 9 degrees, thus,

$$\Delta_7 = \begin{cases} \beta 8 & \text{if } |90 - (V_6 180/3.14)| > 9 \\ \Delta 8 & \text{Otherwise.} \end{cases} \tag{15}$$

The Jack ($\beta 11$) activity which involves a large upper body and lower body movement is determined based on large arm and feet displacement by using

$$\Delta_8 = \begin{cases} \beta 11 & \text{if } V_7 \text{ or } V_8 > 15/\mu_h \text{ and } V_9 \text{ or } V_{10} > 15/\mu_h \\ & \text{and } V_{13} > 20/\mu_h \\ \Delta 9 & \text{Otherwise.} \end{cases} \tag{16}$$

where $\mu_h = 68$ pixels for the Weizmann data set. The human head is one-eighth the human height, i.e., 0.125. Hence, a 15 pixel movement equates to $15/68 = 0.22$ that is almost twice of the height of the human head.

The individual arm motion in the Two Hand Wave ($\beta 10$) and One Hand Wave ($\beta 9$) activities is discerned using the location information. In the Two Hand Wave ($\beta 10$) activity there will be significant movement of both arms while in the One Hand Wave ($\beta 9$) activity there will be significant movement of only one arm. Therefore, the Two Hand Wave ($\beta 10$) and One Hand Wave ($\beta 9$) activities are described below:

$$\Delta_9 = \begin{cases} \beta 10 & \text{if } V_{13} < 20/\mu_h \text{ and } V_8 \geq 5/\mu_h \text{ and} \\ & V_{10} \geq 5/\mu_h \text{and } V_{11} \leq 55 \text{ and } V_{12} < 50 \\ \Delta 10 & \text{Otherwise.} \end{cases} \tag{17}$$

$$\Delta_{10} = \begin{cases} \beta 9 & \text{if } V_{13} < 20/\mu_h \text{ and } V_8 \text{ or } V_{10} \leq 8/\mu_h \\ & \text{and } V_{11} \leq 55 \text{ and } V_{12} > 50 \\ X_\beta & \text{Otherwise.} \end{cases} \tag{18}$$

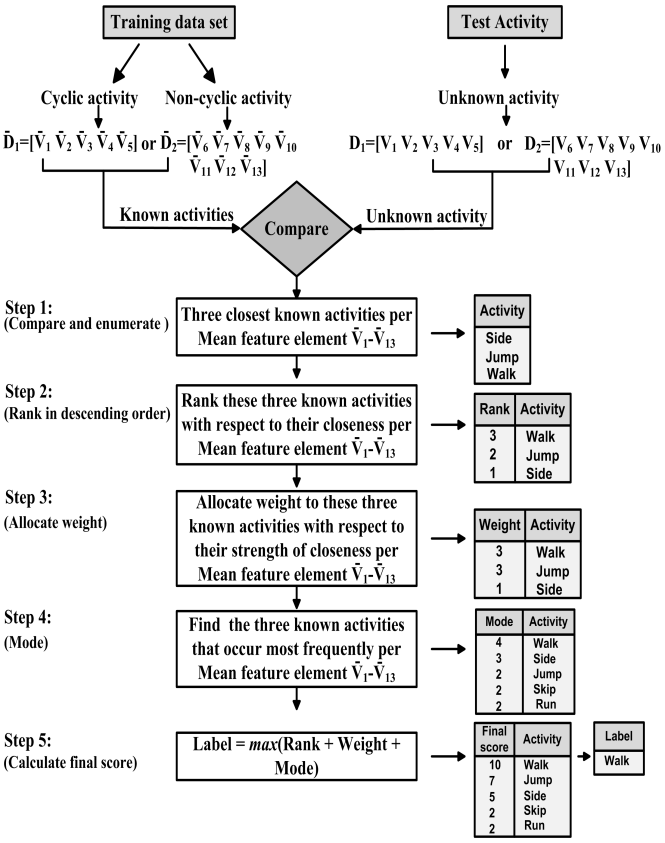


Fig. 5. Proposed majority voting scheme for the unassigned impure activities X_α and X_β using the mean $\bar{D}_{i,\ i=1,2}$.

1) Majority Voting Scheme: The unassigned impure activities X_α and X_β at the second last level of the HRPS (see Fig. 4) are given a label by using a novel majority voting scheme in Fig. 5. This scheme is an integral part of the HRPS and is designed to cater for the increase complexity of multiple overlaps in the feature space of two or more activities. The key idea of this scheme is to accumulate votes based on the rank, assigned weight and frequency (mode) value in order to deduce more accurate decisions at the bottom level of the HRPS.

Given the mean feature descriptors, i.e., $\bar{D}_1 = [\bar{V}_1\ \bar{V}_2\ \bar{V}_3\ \bar{V}_4\ \bar{V}_5]$ and $\bar{D}_2 = [\bar{V}_5\ \bar{V}_6\ \bar{V}_7\ \bar{V}_8\ \bar{V}_9\ \bar{V}_{10}\ \bar{V}_{11}\ \bar{V}_{12}]$, of the known activities of training data set, the goal is to label an unknown impure activity (which contain significant overlaps in the feature space) by extracting the feature descriptors, i.e., $D_1 = [V_1\ V_2\ V_3\ V_4\ V_5]$ and $D_2 = [V_6\ V_7\ V_8\ V_9\ V_{10}\ V_{11}\ V_{12}\ V_{13}]$, in order to calculate the rank, weight and mode as shown in Fig. 5. D_1 and D_2 are used for cyclic and non-cyclic activities, respectively. $V_1 - V_{13}$ represent each feature element of the feature descriptors. The label for the unknown impure activity is determined as follows.

- **Step 1:** Compare each feature element of the feature descriptor, i.e., D_1 or D_2, of one unknown impure activity with the respective mean feature elements of the feature descriptor, i.e., $\bar{D}_1$ or $\bar{D}_2$, for each of the ten known activities in order to enumerate three closest known

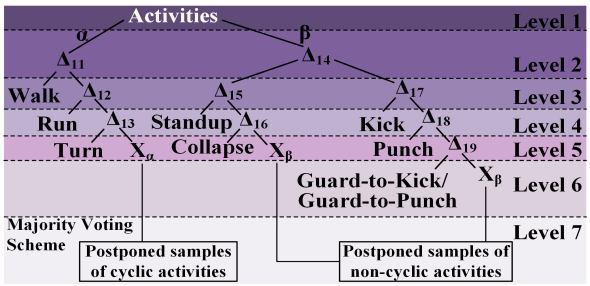


Fig. 6. Hierarchical relaxed partitioning system for the MuHAVi data set. $\Delta_{i,\ i=11,12,..19}$ are the decision rules, and X_α and X_β are the unassigned impure cyclic and non-cyclic activities, respectively, with significant multiple overlaps.

activities per mean feature element.

- **Step 2:** Assign a score (rank) $\nu = 3, 2, 1$ to the three activities enumerated in Step 1 based on their closeness to each of the mean feature elements of $\bar{D}_1$ or $\bar{D}_2$. Next, arrange them in the descending order of their ranks.
- **Step 3:** Allocate a weight $\omega = 3, 2, 1$ to the three ranked activities in Step 2 based on their strength of closeness to the mean feature elements of $\bar{D}_1$ or $\bar{D}_2$.
- **Step 4:** Find the three known activities that occur most frequently (i.e., mode ϖ) per mean feature element of $\bar{D}_1$ or $\bar{D}_2$.
- **Step 5:** Calculate the final score to find the label of the unknown activity. The known activity of the training data set whose rank, weight, and mode yield the maximum score with respect to the unknown activity is assigned as the label for the unknown activity, i.e., Label=$\max(\varpi + \nu + \omega)$.

C. Classification: HRPS for the MuHAVi data set

The robustness of the proposed HRPS method is further validated by applying it with the same feature descriptors $D_{i,\ i=1,2}$ on the MuHAVi dataset [36]. The MuHAVi data set contain eight activities, i.e., the Walk ($\alpha 1$), Run ($\alpha 2$), Turn ($\alpha 6$), Standup ($\beta 12$), Collapse ($\beta 13$), Kick ($\beta 14$), Punch ($\beta 15$) and Guard-to-kick or Guard-to-punch ($\beta 16/\beta 17$). As in Section IV-B the root node is split into α and β activities by using (13). A HRPS with 7 levels is created with decision rules $\Delta_{i,\ i=11,...,19}$ as shown in Fig. 6. Algorithm. IV.1 is used on the 7 training samples of the MuHAVi data set to learn the decision rules $\Delta_{i,\ i=11,12,13}$ for the Walk ($\alpha 1$), Run ($\alpha 2$) and Turn ($\alpha 6$) cyclic activities respectively. The last sample is used as the testing data in a leave-one-out procedure to determine the performance of the HRPS.

Similar to Section IV-B we define decision rules $\Delta_{i,\ i=14,...,19}$ that are fixed based on the human kinesiology [39], anthropometry [40] and body characteristics for non-cyclic activities. Let the reference global angle $V_6 = \theta(t)$ in Stand posture be 90^o. Then, based on biomechanical analysis [41] of human spine the maximum flexion of torso is 60^o, i.e., $(90 - 60 = 30$ or $90 + 60 = 150)$, which causes a significant change in posture. Thus,

$$\Delta_{14} = \begin{cases} \Delta 15 & \text{if } 30 \geq V_6 \geq 150 \\ \Delta 17 & \text{Otherwise} \end{cases} \tag{19}$$

is used to determine whether a transition occurred $\forall\ t \in 1, N$ frames of the activity video. The transition Δ_{15} includes Standup ($\beta 12$) and Collapse ($\beta 13$) activities which contain significant change in posture while the non-transition Δ_{16} contain Kick ($\beta 14$), Punch ($\beta 15$) and Guard-to-kick or Guard-to-punch ($\beta 16/\beta 17$) which do not have significant change in posture. The decision rules for the Standup ($\beta 12$) and Collapse ($\beta 13$), i.e., Δ_{15} and Δ_{16}, respectively are defined as

$$\Delta_{15} = \begin{cases} \beta 12 & \text{if } 30 \geq V_6 \geq 150, \text{ at } t = 1 \\ & \text{and } 65 \leq V_6 \leq 125, \forall\ t \in 2, N \\ \Delta 16 & \text{Otherwise} \end{cases} \tag{20}$$

$$\Delta_{16} = \begin{cases} \beta 13 & \text{if } 65 \leq V_6 \leq 125, \text{ at } t = 1 \\ & \text{and } 30 \geq V_6 \geq 150, \forall\ t \in 2, N \\ X_\beta & \text{Otherwise} \end{cases} \tag{21}$$

The range $125 - 65 = 60^o$ [41] is selected as it corresponds to the flexion and extension range of human body while maintaining a somewhat Stand posture. We are motivated from [15] to borrow the definition of the Kick and Punch IBM as decision rules for the Kick ($\beta 14$) and Punch ($\beta 15$) activities. Hence,

$$\Delta_{17} = \begin{cases} \beta 14 & \text{if } 2 \leq 90 - V_6 \leq 15 \\ \Delta 18 & \text{Otherwise.} \end{cases} \tag{22}$$

$$\Delta_{18} = \begin{cases} \beta 15 & \text{if } 90 - V_6 > 15 \\ \Delta 19 & \text{Otherwise.} \end{cases} \tag{23}$$

Note that in Punch ($\beta 15$), the arm moves across the body in a diagonal manner and as a result the angle of body from the vertical is quite large. The Guard-to-punch and Guard-to-kick are considered as one class because both primarily have a guard activity with minimal movement of the arms and legs. In Guard-to-kick or Guard-to-punch ($\beta 16/\beta 17$), the human remains in Stand posture with least angle of body from the vertical. Hence,

$$\Delta_{19} = \begin{cases} \beta 16/\beta 17 & \text{if } 90 - V_6 < 2 \\ X_\beta & \text{Otherwise.} \end{cases} \tag{24}$$

The unassigned impure activities X_α and X_β are given a label by using the MVS (see Section IV-B1).

V. Experimental results

The Weizmann dataset [35] comprises ninety low-resolution 180×144 video sequences of nine subjects performing ten daily activities. The MuHAVi dataset [36] comprises eight high resolution 720×576 primitive activity classes of two actors with two samples with two different views (camera 3 and camera 4), i.e., total eight samples, per activity. We use a standard leave-one-out cross validation method. The activities and their acronyms are shown in Table I.

A. Feature descriptors evaluation

The 3D scatter plots of the selected features are shown in Fig. 7 and Fig. 8 to visualize the distribution of the activities of the input data set. It can be seen from Fig. 7 (a) that the Walk activity has the least leg frequency (most blue circles between 2-3 Hz) and the Run activity has the maximum leg frequency

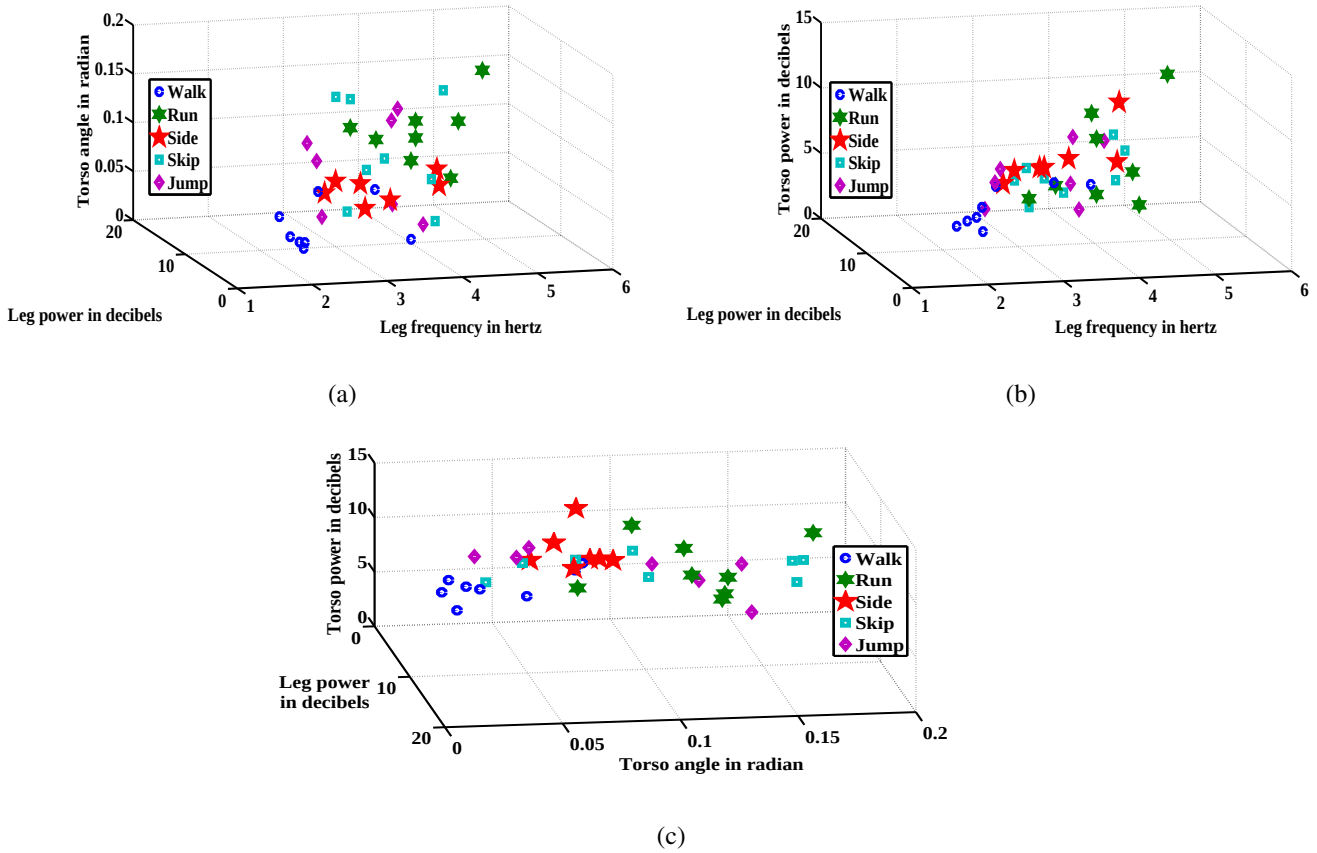


Fig. 7. 3D scatter plots of the selected features that show the distribution of the cyclic activities for the input Weizmann data set.

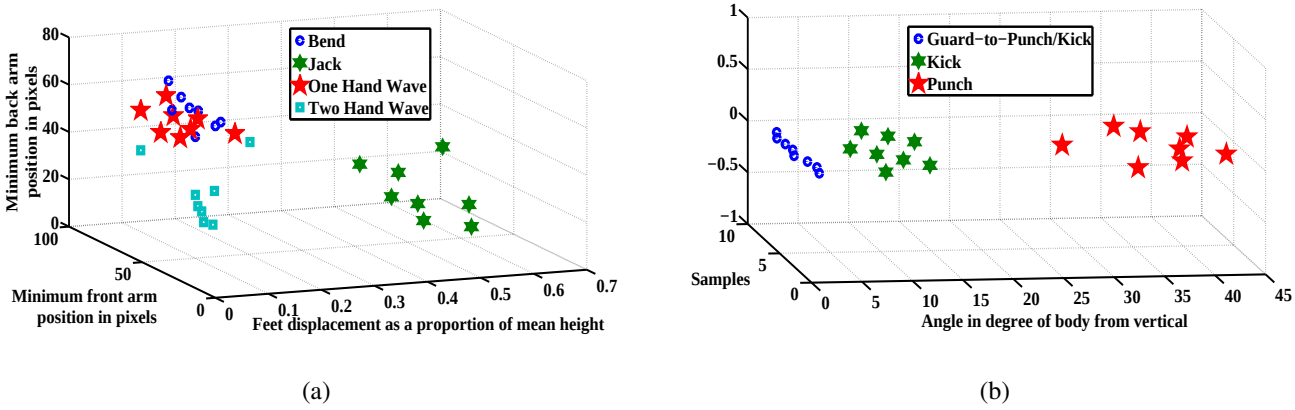


Fig. 8. 3D scatter plots of the selected features that show the distribution of the activities for the input Weizmann and MuHAVi data sets.

(green pentagons lie between 4-6 Hz onwards). Similarly, it can be seen in Fig. 7 (b) that the torso power of the Walk activity is much less than the remaining cyclic activities. In Fig. 7 (c) it can be seen that the torso angle of most of the Run (green pentagons), Jump (purple diamonds) and Skip (light blue square) activities is more than the Walk (blue circles) and Side (red stars) activity. It can be observed from Fig. 7 (c) that the Walk activity has the least torso angle (blue circles between 0-0.05 radian) while the torso angle for the Side (red stars) activity is concentrated between 0.05-0.1 radian.

The Fig. 8 (a) shows the 3D scatter plots of the selected features for the Bend, Jack, One Hand Wave and Two Hand Wave activities of the Weizmann data set. It can be seen that the Jack activity has the maximum displacement of the feet as a proportion of the mean height of subject. Also, it can be seen that in the Two Hand Wave (light blue square) activity both front and back arm have minimum position in pixels, and is well separate from the One Hand Wave (red star) activity. The Fig. 8 (b) shows the 3D scatter plots of a selected feature for the Guard-to-Punch or Guard-to-Kick,

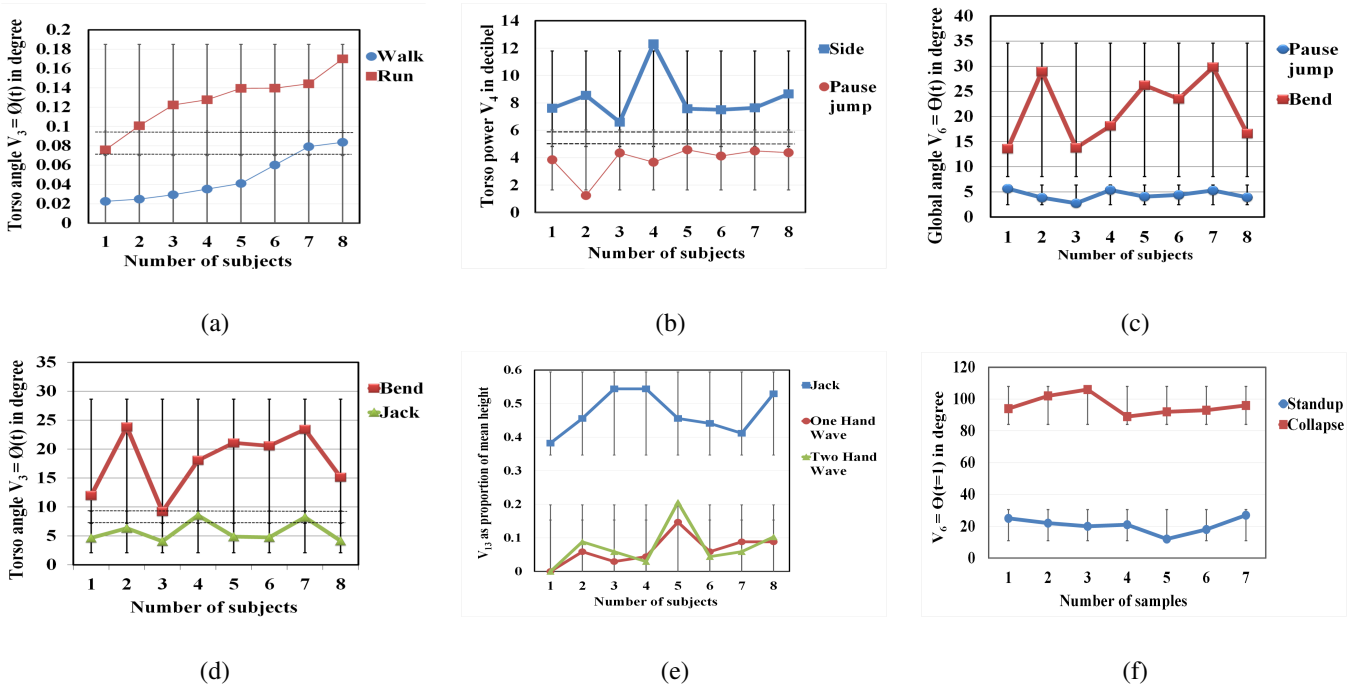


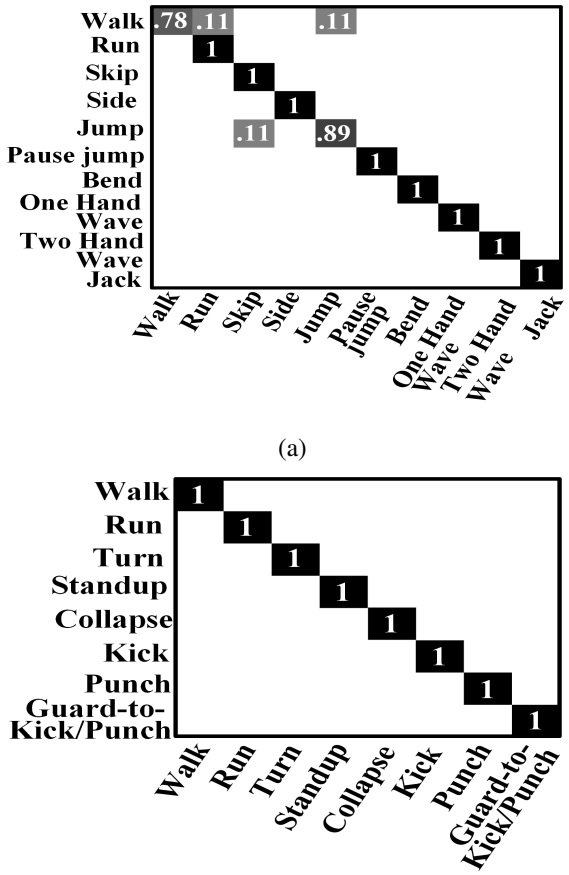
Fig. 9. Significance of the extracted features for discerning activities. Error bars show 95% confidence intervals on selected features with two standard deviation as an error metric. (a)-(e) Weizmann data set and (f) MuHAVi data set.

Kick and Punch activities of the MuHAVi data set. It can be seen that the Guard-to-Punch or Guard-to-Kick has the least variation in the angle of body from the vertical and the Punch has the maximum angle of body from the vertical. The angle of body from the vertical for the Kick activity lies in between the Guard-to-Punch or Guard-to-Kick and Punch activity.

In Fig. 9, we illustrate the ability of some of the features from $D_{i,\ i=1,2}$ to discern various human activities of the Weizmann and MuHAVi data sets. The error bars show 95% confidence intervals on selected features with two standard deviation as an error metric. Although the leg frequency, i.e., V_1, of the Walk ($\alpha 1$) and Run ($\alpha 2$) activity is dissimilar based on speed of the leg movement but anomalies like some subjects walking faster causes misclassification. However, it can be seen from Fig. 9 (a) that the torso angle $V_3 = \phi(t)$ provides a good separation to discern the Walk ($\alpha 1$) and Run ($\alpha 2$) activities. Similarly, the newly introduced torso power feature V_4 provides a reasonable distinction between the Side ($\alpha 4$) and Pause Jump ($\beta 7$) activities as shown in Fig. 9 (b). In Fig. 9 (c), the global angle $V_6 = \theta(t)$ provides clear separation between the Pause Jump ($\beta 7$) and Bend ($\beta 8$) activity while in Fig. 9 (d) the torso angle $V_3 = \phi(t)$ provides sufficient discerning ability between the Bend ($\beta 8$) and Jack ($\beta 11$) activity. It can be observed from Fig. 9 (e) that the distance between the legs, i.e., V_{13}, gives a very good separation among the Jack ($\beta 11$), One Hand Wave ($\beta 9$) and Two Hand Wave ($\beta 10$) activities. Finally, in Fig. 9 (f) the global angle $V_6 = \theta(t = 1)$ easily discern the Standup ($\beta 12$) and Collapse ($\beta 12 = 3$) activities. Thus, the $D_{i,\ i=1,2}$ acquires meaningful information. However, there is slight overlap in the confidence intervals of some of the features, e.g., Fig. 9 (a), (b) and (d). This illustrate the importance of using HRPS to postpone decisions on such samples that lie closer to the samples of another activity. Also, for these samples the MVS is better suited because it takes into account multiple criteria based on the average values of all the feature elements obtained from the training data set to assign a label to an unknown activity. As stated in [6] the average features provide more generalized information about the movement pattern of body during an activity.

B. Classification evaluation

The confusion tables for the HRPS method on the Weizmann and MuHAVi data set are shown in Fig. 10 (a) and (b) respectively. We obtained a mean classification accuracy of 96.7% for ten activities of the Weizmann data set (see Table II and details below for significance in comparison to other methods). This shows that our method robustly recognises activities that have significant multiple overlaps in the feature space. In particular, our method recognises four activities, i.e., Run ($\alpha 2$), Side ($\alpha 4$), Jump ($\alpha 5$) and Pause Jump ($\beta 13$), out of the six cyclic activities with a mean classification accuracy of 100%. Thus, our method robustly discerns similar cyclic activities. It obtains a mean classification accuracy of 94.5% for all the six cyclic activities, i.e, Walk ($\alpha 1$), Run ($\alpha 2$), Side ($\alpha 4$), Jump ($\alpha 5$), Skip ($\alpha 3$) and Pause Jump ($\beta 13$). The decomposition of the Walk ($\alpha 1$) into the Run ($\alpha 2$) and Jump ($\alpha 5$) activities is reasonable due to similar motion. Also, the Skip ($\alpha 3$) and Jump ($\alpha 5$) activities are similar in the way the subject bounces across the video. The non-cyclic activities, i.e., Bend ($\beta 14$), Jack ($\beta 11$), Two Hand Wave ($\beta 10$) and One Hand Wave ($\beta 15$) are robustly classified with a mean classification accuracy of 100%. This proves that the decision



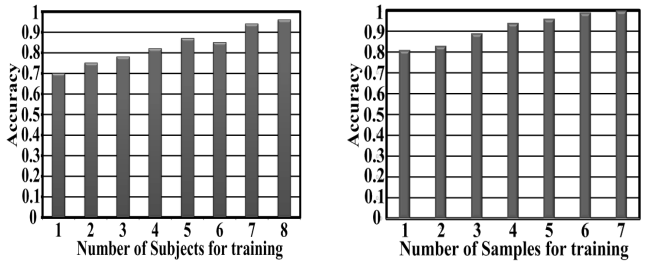
(a)

(b)

Fig. 10. Confusion table (see Table I for α and β). (a) Weizmann data set and (b) MuHAVi data set.

rules based on human kinesiology and body characteristics work well. We obtained a mean classification accuracy of 100% for eight activities of the MuHAVi data set as shown in Fig. 10 (b). The results demonstrate that the proposed HRPS method can robustly distinguish various activities in two different (low and high) resolution data sets. It also show that our method perform well under different views, i.e., camera 3 and camera 4, for the MuHAVi data set. A high accuracy on the Standup ($\beta 12$), Collapse ($\beta 13$), Kick ($\beta 14$), Punch ($\beta 15$) and Guard-to-kick or Guard-to-punch ($\beta 16/\beta 17$) activities demonstrate the importance of decision rules based on human kinesiology and body characteristics.

Fig. 11 (a) shows classification performance with respect to training subjects of the Weizmann data set. It can be seen that the classification accuracy of the proposed method is about 70% with only one training subject. However, as the number of training subjects increase the classification accuracy also improves. The classification accuracy becomes slightly stable when the number of training subjects is four, five and six. The best performance is achieved with eight training subjects. The classification performance with respect to training samples of the MuHAVi data set is shown in Fig. 11 (b). It can be seen that the classification performance increases steadily till it reaches 100% with seven samples used for training.



(a) (b)

Fig. 11. Classification performance. (a) Weizmann data set and (b) MuHAVi data set.

TABLE II
COMPARISON ON THE WEIZMANN DATA SET.

Method	Accuracy%	Real-time	Intensive training	Year
Michalis, et al. [5]	100	**No**	**Yes**	2014
Marlon, et al. [23]	96.7	Yes	No	2014
Mahbub, et al. [6]	100	**No**	No	2014
Ma, et al. [10]	100	**No**	**Yes**	2013
Romain, et al. [8]	82.79	**No**	**Yes**	2013
Zhuolin, et al. [3]	100	Yes	**Yes**	2012
Saad, et al. [7]	95.75	**No**	**Yes**	2010
Elden, et al. [24]	93.6	Yes	No	2009
Saad, et al. [25]	92.6	-	No	2007
Our method	**96.7**	**Yes**	**No**	2014

Table II compares the HRPS with relevant state-of-the-art methods (see Section II) for activity recognition on the Weizmann data set. It shows that the our method outperforms the methods in [7], [8], [24], [25] in terms of accuracy. Saad et al. [25] only deals with nine activities. The method in [5], [7], [8], [6] and [10] are not real-time since they require intensive training for learning vocabulary. Zhuolin, et al. [3] required both shape and motion features to achieve 100% accuracy. On a similar basis, i.e., using motion features, they obtain 88.89% accuracy while our method obtains 96.7%. Their method is reported to be fast but requires intensive training and uses optical flow which is usually computationally expensive. Hence, these methods are not suitable for real-world applications. In contrast, our method operates in real-time, avoid intensive training, and it is simple to implement and extend for new activity categories (i.e., for each new category new features can be added to the HRPS). This makes it more suitable for real world applications. The model-free method in [16] recognizes only two activities, i.e., the Walk and Run with 97% accuracy. On similar activities, i.e., Walk ($\alpha 1$), Run ($\alpha 2$), and Jump ($\alpha 5$), the method in [29] has mean classification accuracy of 82.4% while we obtain 92.7% mean classification accuracy. The method in [43] although real-time and non-intensive but achieves only 90.32% on the Weizmann data set.

In Table III, our HRPS method is compared with recent methods on the MuHAVi data set. Our method achieved better recognition rate than most of the methods and works in real-time with no intensive training. On both data sets our method is comparable to the method in [23]. On Intel (R) Core (TM)

TABLE III
COMPARISON ON THE MUHAVI DATA SET.

Method	Accuracy%	Real-time	Intensive training	Year
Alexandros, et al. [42]	100	Yes	No	2014
Marlon, et al. [23]	100	Yes	No	2014
Alexandros, et al. [43]	97.1	Yes	No	2013
Abdalrahman, et al. [44]	98.5	**No**	No	2011
Sanchit, et al. [36]	97.8	Yes	No	2010
Martinez, et al. [45]	98.4	**No**	**Yes**	2009
Our method	**100**	**Yes**	**No**	2014

i7 2.93 GHz with 4 GB RAM and Windows 7, the feature extraction in OpenCV 2.4.6 takes 0.031 and 0.071 seconds per image frame on the Weizmann and MuHAVi data sets respectively. The classification in MatLab takes 0.183 seconds for all activities. Marlon, et al. [23] method takes 4.85 and 2859.29 seconds for feature extraction on the Weizmann and MuHAVi data sets respectively. This demonstrates that the HRPS method works in real-time.

VI. SUMMARY

We proposed a hierarchical relaxed partitioning system to efficiently and robustly recognize activities. Our method first discerns the pure activities from the impure activities, and then tackles the multiple overlaps problem of the impure activities via an innovative majority voting scheme. The results proved that our method not only accurately discerns similar activities, but also obtains real-time recognition on two (low and high) resolution data sets, i.e., Weizmann and MuHAVi respectively. It also performs well under two different views of the MuHAVi data set. These attributes make our method more suitable for real-world applications in comparison to the state-of-the-art methods.

REFERENCES

[1] P. Ribeiro and J. Santos-victor, "Human activity recognition from video: modeling, feature selection and classification architecture," in *Int. Workshop on Human Activity Recognit. and Modeling*, 2005.

[2] H. Qian, Y. Mao, W. Xiang, and Z. Wang", "Recognition of human activities using {SVM} multi-class classifier," *Pattern Recognition Letters*, vol. 31, no. 2, pp. 100–111, 2010.

[3] Z. Jiang, Z. Lin, and L. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 533–547, Mar 2012.

[4] M. Lucena, N. P. de la Blanca, and J. Fuertes, "Human action recognition based on aggregated local motion estimates," *Mach. Vis. Appl.*, vol. 23, no. 1, pp. 135–150, 2012.

[5] M. Vrigkas, V. Karavasilis, and C. Nikou, "Matching mixtures of trajectories for human action recognition," *Comput. Vision and Image Understanding*, vol. 19, pp. 27–40, Jan 2014.

[6] U. Mahbub, H. Imtiaz, and M. Ahad, "Action recognition based on statistical analysis from clustered flow vectors," *Signal, Image and Video Processing*, vol. 8, no. 2, pp. 243–253, 2014.

[7] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, Feb 2010.

[8] R. Tavenard, R. Emonet, and J. Odobez, "Time-sensitive topic models for action recognition in videos," in *Proc. of IEEE Int. Conf. on Image Processing*, 2013.

[9] I. Jargalsaikhan, S. Little, C. Direkoglu, and N. O'Connor, "Action recognition based on sparse motion trajectories," in *Proc. of IEEE Int. Conf. on Image Processing*, 2013.

[10] A. Ma, P. Yuen, W. Zou, and J.-H. Lai, "Supervised spatio-temporal neighborhood topology learning for action recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 8, pp. 1447–1460, Aug 2013.

[11] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?" in *Proc. of IEEE Int. Comput. Vis. and Pattern Recognit.*, June 2008, pp. 1–8.

[12] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *Proc. of IEEE Int. Comput. Vis. and Pattern Recognit.*, June 2008, pp. 1–8.

[13] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comput. Vis. Image Underst.*, vol. 117, no. 6, pp. 633–659, Jun 2013.

[14] X. Sun, M. Y. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in *Proc. of IEEE Int. Comput. Vis. and Pattern Recognit.*, 2009.

[15] F. Azhar and T. Tjahjadi, "Significant body point labelling and tracking," *IEEE Trans. on Cybern.*, vol. 44, no. 9, pp. 1673–1685, Sep 2014.

[16] H. Fujiyoshi, A. J. Lipton, and T. Kanade, "Real-time human motion analysis by image skeletonization," *IEICE Trans. Inf. Syst. E SERIES D.*, vol. 87, no. 1, pp. 113–120, 2004.

[17] M. Marszaek and C. Schmid, "Constructing category hierarchies for visual recognition," in *Proc. of European Conf. on Comput. Vision*, 2008, pp. 479–491.

[18] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *Proc. of IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2008.

[19] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognit.*, 2006, pp. 2161–2168.

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognit.*, 2007.

[21] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3{D}-gradients," in *Proc. of British Machine Vision Conf.*, 2008, pp. 99.1–99.10.

[22] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognit.*, June 2008, pp. 1–8.

[23] M. Alcantara, T. Moreira, and H. Pedrini, "Real-time action recognition based on cumulative motion shapes," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 2014, pp. 2917–2921.

[24] E. Yu and J. Aggarwal, "Human action recognition with extremities as semantic posture representation," *Proc. of IEEE Conf. on Comput. Vision and Pattern Recognit. Workshops*, pp. 1–8, 2009.

[25] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *Proc. of IEEE 11th Int. Conf. on Comput. Vision*, Oct 2007, pp. 1–8.

[26] E. Yu and J. Aggarwal, "Detection of fence climbing from monocular video," in *Proc. of 18th Int. Conf. on Pattern Recognit.*, vol. 1, 2006, pp. 375–378.

[27] R. Telea and J. J. V. Wijk, "An augmented fast marching method for computing skeletons and centerlines," in *Proc. of Symp. on Data Visualisation*, 2002, pp. 251–259.

[28] C. F. Juang, C. M. Chang, J. R. Wu, and D. Lee, "Computer vision-based human body segmentation and posture estimation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans.*, vol. 39, no. 1, pp. 119–133, 2009.

[29] C. C. Yu, Y. N. Chen, H. Y. Cheng, J. N. Hwang, and K. C. Fan, "Connectivity based human body modeling from monocular camera," *J. Inf. Sci. Eng.*, vol. 26, pp. 363–377, 2010.

[30] W. Lao, J. Han, and P. H. With, "Fast detection and modeling of human-body parts from monocular video," in *Proc. of 5th Int. Conf. Articulated Motion and Deformable Objects*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 380–389.

[31] W. Lao, J. Han, and P. H. N. de With, "Flexible human behavior analysis framework for video surveillance applications," *Int. J. Digit. Multimedia. Broadcast.*, pp. 1–10, 2010.

[32] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, 2000.

[33] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[34] M. Bregonzio, S. Gong, and T. Xiang, "Action recognition with cascaded feature selection and classification," in *Proc. of Int. Conf. on Imaging for Crime Detection and Prevention*, 2009.

[35] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.
[36] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *Proc. of IEEE Int. Conf. Adv. Video and Signal Based Surveillance*, Sep 2010, pp. 48–55.
[37] J. Bent, "Data-driven batch scheduling," Ph.D. dissertation, University of Wisconsin, Madison, May 2005.
[38] B. Arbab-Zavar, I. Bouchrika, J. N. Carter, and M. S. Nixon, "On supervised human activity analysis for structured environments." in *Proc. of 6th Int. Symposium on Visual Computing*, ser. Lecture Notes in Computer Science, vol. 6455. Springer, 2010, pp. 625–634.
[39] N. Hamilton, W. Weimar, and K. Luttgens, *KINESIOLOGY Scientific Basis of Human Motion*. McGraw-Hill, 2011.
[40] R. Easterby, K. Kroemer, and D. B. Chaffin., *Anthropometry and Biomechanics*. New York: Plenum Press, 2010.
[41] J. Hamill and K. M. Knutzen., *Biomechanical basis of Human Movement*. Lippincott Williams and Wilkins, Wolters Kluwer, 2009.
[42] "A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views," *Int. Scholarly Research Notices*, pp. 1–36, July.
[43] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognit. Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.
[44] A. Eweiwi, S. Cheema, C. Thurau, and C. Bauckhage, "Temporal key poses for human action recognition," in *Proc. of IEEE Int. Conf. on Computer Vision Workshops*, Nov 2011, pp. 1310–1317.
[45] F. Martinez-Contreras, C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, and S. Velastin, "Recognizing human actions using silhouette-based hmm," in *Proc. of IEEE Int. Conf. on Adv. Video and Signal Based Surveillance*, Sept 2009, pp. 43–48.

References

[1] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Underst.*, vol. 104, no. 2, pp. 90–126, Nov 2006.

[2] L. Sigal, A. O. Balan, and M. J. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *Int. J. Comput. Vision*, vol. 87, no. 1-2, pp. 4–27, Mar 2010.

[3] R. Poppe, "Vision-based human motion analysis: An overview," *Comput. Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, 2007.

[4] M. Goffredo, M. Schmid, S. Conforto, M. Carli, A. Neri, and T. D'Alessio, "Markerless human motion analysis in Gauss-Laguerre transform domain: an application to sit-to-stand in young and elderly people." *IEEE Trans. Information Technol. in Biomedicine*, vol. 13, no. 2, pp. 207–16, 2009.

[5] NASA-STD-3000, "Anthropometry and biomechanics," http://msis.jsc.nasa.gov/sections/section03.htm, 1995.

[6] R. Easterby, K. Kroemer, and D. B. Chaffin., *Anthropometry and Biomechanics.* New York: Plenum Press, 2010.

[7] C. Barrón and I. A. Kakadiaris, "Estimating anthropometry and pose from a single uncalibrated image," *Comput. Vis. Image Underst.*, vol. 81, no. 3, pp. 269–284, Mar 2001.

[8] C. Benabdelkader and Y. Yacoob, "Statistical estimation of human anthropometry from a single uncalibrated image," in *Workshop on Biometric Authentication*, 2008, pp. 200–220.

[9] G. Bradski and A. Kaehler., *Learning OpenCV Computer Vision with the OpenCV Library.* O'Reilly Media, Sebastopol, Sep 2008.

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

[11] A. Kläser, "Learning human actions in video," Ph.D. dissertation, Université de Grenoble, Jul 2010.

[12] M. Lucena, N. P. de la Blanca, and J. Fuertes, "Human action recognition based on aggregated local motion estimates," *Mach. Vis. Appl.*, vol. 23, no. 1, pp. 135–150, 2012.

[13] Z. Jiang, Z. Lin, and L. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 3, pp. 533–547, Mar 2012.

[14] M. Vrigkas, V. Karavasilis, and C. Nikou, "Matching mixtures of trajectories for human action recognition," *Comput. Vision and Image Understanding*, vol. 19, pp. 27–40, Jan 2014.

[15] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, Feb 2010.

[16] R. Tavenard, R. Emonet, and J. Odobez, "Time-sensitive topic models for action recognition in videos," in *Proc. of IEEE Int. Conf. on Image Processing*, 2013.

[17] I. Jargalsaikhan, S. Little, C. Direkoglu, and N. O'Connor, "Action recognition based on sparse motion trajectories," in *Proc. of IEEE Int. Conf. on Image Processing*, 2013.

[18] M. Marszaek and C. Schmid, "Constructing category hierarchies for visual recognition," in *Proc. of European Conf. on Comput. Vision*, 2008, pp. 479–491.

[19] I. Haritaoglu, D. Harwood, and L. S. Davis, "W4: real-time surveillance of people and their activities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 809–830, 2000.

[20] C. F. Juang, C. M. Chang, J. R. Wu, and D. Lee, "Computer vision-based human body segmentation and posture estimation," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans.*, vol. 39, no. 1, pp. 119–133, 2009.

[21] H. Fujiyoshi, A. J. Lipton, and T. Kanade, "Real-time human motion analysis by image skeletonization," *IEICE Trans. Inf. Syst. E SERIES D.*, vol. 87, no. 1, pp. 113–120, 2004.

[22] E. Yu and J. Aggarwal, "Human action recognition with extremities as semantic posture representation," *Proc. of IEEE Conf. on Comput. Vision and Pattern Recognit. Workshops*, pp. 1–8, 2009.

[23] S. Ali, A. Basharat, and M. Shah, "Chaotic invariants for human action recognition," in *Proc. of IEEE 11th Int. Conf. on Comput. Vision*, Oct 2007, pp. 1–8.

[24] J. J. Wang and S. Singh, "Video analysis of human dynamics - a survey," *Real-Time Imaging*, vol. 9, no. 5, pp. 321–346, 2003.

[25] F. Cardinaux, D. Bhowmik, C. Abhayaratne, and M. S. Hawley, "Video based technology for ambient assisted living: A review of the literature," *J. Ambient Intell. Smart Environ.*, vol. 3, no. 3, pp. 253–269, Aug 2011.

[26] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behaviour analysis for ambient-assisted living," *Expert Syst. Appl.*, vol. 39, no. 12, pp. 10 873–10 888, Sep 2012.

[27] J. Candamo, M. Shreve, D. Goldgof, D. Sapper, and R. Kasturi, "Understanding transit scenes: A survey on human behaviour-recognition algorithms," *IEEE Trans. Intell. Transportation Sys.*, vol. 11, no. 1, pp. 206–224, 2010.

[28] A. Hampapur, "Smart video surveillance for proactive security [in the spotlight]," *Signal Processing Magazine, IEEE*, vol. 25, no. 4, pp. 136–134, July 2008.

[29] S. Fleck and W. Strasser, "Smart camera based monitoring system and its application to assisted living," *Proc. of the IEEE*, vol. 96, no. 10, pp. 1698–1714, Oct 2008.

[30] A. Gritai and M. Shah, "Tracking of human body joints using anthropometry," in *Proc. of IEEE Int. Conf. on Multimedia and Expo*, July 2006, pp. 1037–1040.

[31] R. Cucchiara, A. Prati, R. Vezzani, and R. Emilia, "A multi-camera vision system for fall detection and alarm generation," *Expert Syst.*, vol. 24, no. 5, pp. 334–345, 2007.

[32] C. C. Yu, J. N. Hwang, G. F. Ho, and C. H. Hsieh, "Automatic human body tracking and modelling from monocular video sequences," in *Proc. of IEEE Int. Conf. Acoust., Speech and Signal Process.*, vol. 1, Apr 2007, pp. I–917–920.

[33] C. C. Yu, Y. N. Chen, H. Y. Cheng, J. N. Hwang, and K. C. Fan, "Connectivity based human body modelling from monocular camera," *J. Inf. Sci. Eng.*, vol. 26, pp. 363–377, 2010.

[34] E. Yu and J. Aggarwal, "Detection of fence climbing from monocular video," in *Proc. of 18th Int. Conf. on Pattern Recognit.*, vol. 1, 2006, pp. 375–378.

[35] K. Takahashi and T. Kodama, "Remarks on simple motion capture using heuristic rules and monte carlo filter," in *Proc. of Int. Conf. Image and Graphics.*, Sep 2009, pp. 808–813.

[36] C. Wu and H. Aghajan, "Real-time human pose estimation: A case study in algorithm design for smart camera networks," *Proceedings of the IEEE*, vol. 96, no. 10, pp. 1715–1732, Oct 2008.

[37] M. K. Leung and Y. H. Yang, "First sight: A human body outline labelling system," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 4, pp. 359–377, 1995.

[38] W. Lao, J. Han, and P. H. With, "Fast detection and modeling of human-body parts from monocular video," in *Proc. of 5th Int. Conf. Articulated Motion and Deformable Objects.* Berlin, Heidelberg: Springer-Verlag, 2008, pp. 380–389.

[39] F. Huo, E. Hendriks, P. Paclik, and A. H. J. Oomes, "Markerless human motion capture and pose recognition," in *Proc. Image Anal. Multimedia Interactive Serv.*, May 2009, pp. 13–16.

[40] N. Thome, D. Merad, and S. Miguet, "Human body part labelling and tracking using graph matching theory," in *Proc. of IEEE Int. Conf. on Video and Signal Based Surveillance*, Nov 2006, pp. 38–38.

[41] R. D. Green and L. Guan, "Quantifying and recognizing human movement patterns from monocular video images - part I: A new framework for modeling human motion," *IEEE Trans. on Circuits and Syst. for Video Technol.*, vol. 14, pp. 179–190, 2003.

[42] J. M. del Rinco, D. Makris, C. O. Urunuela, and J. C. Nebel, "Tracking human position and lower body parts using kalman and particle filters constrained by human biomechanics," *IEEE Trans. on Syst., Man, and Cybernetics, Part B: Cybernetics*, vol. 41, no. 1, pp. 26–37, Feb 2011.

[43] M. W. Lee, I. Cohen, and S. K. Jung, "Particle filter with analytical inference for human body tracking," in *Proc. of Motion and Video Computing*, Dec 2002, pp. 159–165.

[44] H. Sidenbladh, M. J. Black, and D. J. Fleet, "Stochastic tracking of 3D human figures using 2D image motion," in *Proc. of the 6th European Conf. on Computer Vision-Part II*, 2000, pp. 702–718.

[45] J. Humpherys, P. Redd, and J. M. West, "A fresh look at the kalman filter." *SIAM Review*, vol. 54, pp. 801–823, 2012.

[46] D. B. Rubin, "The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The SIR algorithm," *Journal of the American Statistical Association*, vol. 82, no. 398, pp. 543–546, 1987.

[47] J. Maccormick and M. Isard, "Partitioned sampling, articulated objects, and interface-quality hand tracking," in *Proc. of the 6th European Conf. on Comput. Vision-Part II*, 2000.

[48] J. Sullivan, A. Blake, M. Isard, and J. MacCormick, "Bayesian object localisation in images," *Int. J. of Comput. Vision*, vol. 44, 2001.

[49] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *Int. J. on Computer Vision*, vol. 61, no. 2, pp. 185–205, Feb 2005.

[50] J. Deutscher, A. Blake, and I. Reid, "Articulated body motion capture by annealed particle filtering," in *Proc. of IEEE Conf. on Comput. Vision and Pattern Recognit.*, vol. 2, Aug 2002, pp. 126–133.

[51] A. D. Bagdanov, A. Del Bimbo, F. Dini, and W. Nunziati, "Improving the robustness of particle filter-based visual trackers using online parameter adaptation," in *Proc. of IEEE Conf. on Adv. Video and Signal Based Surveillance*, Sep 2007, pp. 218–223.

[52] K. Nummiaro, E. Koller-Meier, and L. J. V. Gool, "Object tracking with an adaptive color-based particle filter," in *Proc. of the 24th DAGM Symposium on Pattern Recognit.* London, UK, UK: Springer-Verlag, 2002, pp. 353–360.

[53] E. Maggio and A. Cavallaro, "Hybrid particle filter and mean shift tracker with adaptive transition model," in *Proc. of Int. Conf. Acoustics, Speech, and Signal Processing*, 2005, pp. 221–224.

[54] C. Shan, T. Tan, and Y. Wei, "Real-time hand tracking using a mean shift embedded particle filter," *Pattern Recogn.*, vol. 40, no. 7, pp. 1958–1970, Jul 2007.

[55] Z. Wang, X. Yang, Y. Xu, and S. Yu, "Camshift guided particle filter for visual tracking," in *IEEE Workshop on Signal Processing Syst.*, Oct 2007, pp. 301 –306.

[56] M. Morshidi and T. Tjahjadi, "Gravity optimised particle filter for hand tracking," *Pattern Recogn.*, vol. 47, no. 1, pp. 194–207, Jan 2014.

[57] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar 2001.

[58] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, 2007.

[59] M. Grundmann, F. Meier, and I. A. Essa, "3D shape context and distance transform for action recognition." in *Proc. of IEEE Int. Conf. on Pattern Recongit.*, 2008, pp. 1–4.

[60] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, "Motion context: A new representation for human action recognition," ser. Lecture Notes in Computer Science, 2008, vol. 5305, pp. 817–829.

[61] I. Laptev, "On space-time interest points," *Int. J. Comput. Vision*, vol. 64, pp. 107–123, Sep 2005.

[62] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, Sep 2009, p. 127.

[63] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004.

[64] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in *Proc. of the 15th Int. Conf. on Multimedia*, 2007, pp. 357–360.

[65] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Conference on Computer Vision & Pattern Recognition*, Jun 2008.

[66] F. Hu, L. Luo, F. Zhang, and J. Liu, "Action recognition using hybrid spatio-temporal bag-of-features," in *5th Int. Con. on Computer Sciences and Convergence Information Technology*, Nov 2010, pp. 812–815.

[67] J. Liu, Y. Yang, I. Saleemi, and M. Shah, "Learning semantic features for action recognition via diffusion maps." *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 361–377, 2012.

[68] A. Doulamis, N. Doulamis, L. v. Gool, and M. Nixon, "Guest editorial: Event-based video analysis/retrieval," *Multimedia Tools and Applications*, vol. 69, pp. 247–251, Mar 2014.

[69] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[70] M. Iglesias-Ham, E. Garca-Reyes, W. Kropatsch, and N. Artner, "Convex deficiencies for human action recognition," *Journal of Intelligent & Robotic Systems*, vol. 64, pp. 353–364, 2011.

[71] W. Lao, J. Han, and P. H. N. de With, "Flexible human behaviour analysis framework for video surveillance applications," *Int. J. Digit. Multimedia. Broadcast.*, pp. 1–10, 2010.

[72] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comput. Vis. Image Underst.*, vol. 117, no. 6, pp. 633–659, Jun 2013.

[73] S. Singh, S. A. Velastin, and H. Ragheb, "Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods," in *Proc. of IEEE Int. Conf. Adv. Video and Signal Based Surveillance*, Sep 2010, pp. 48–55.

[74] C. Barrón and I. A. Kakadiaris, "On the improvement of anthropometry and pose estimation from a single uncalibrated image," *Mach. Vision and Appl.*, vol. 14, no. 4, pp. 229–236, 2003.

[75] C. F. Juang and C. M. Chang, "Human body posture classification by a neural fuzzy network and home care system application," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans.*, vol. 37, no. 6, pp. 984–994, 2007.

[76] L. Huang, S. Tang, Y. Zhang, S. Lian, and S. Lin, "Robust human body segmentation based on part appearance and spatial constraint," *Neurocomputing*, vol. 118, pp. 191–202, 2013.

[77] L. Ladicky, P. H. S. Torr, and A. Zisserman, "Human pose estimation using a joint pixel-wise and part-wise formulation," in *Proc. of IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2013.

[78] M. Dantone, J. Gall, C. Leistner, and L. van Gool, "Human pose estimation from still images using body parts dependent joint regressors," in *Proc. of IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2013.

[79] H. Bhaskar, L. Mihaylova, and S. Maskell, "Articulated human body parts detection based on cluster background subtraction and foreground matching," *Neurocomputing*, vol. 100, pp. 58–73, Jan 2013.

[80] 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," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognit.*, Jun 2011, pp. 1297–1304.

[81] V. Kazemi, M. Burenius, H. Azizpour, and J. Sullivan, "Multi-view body part recognition with random forests," in *Proc. of IEEE British Mach. Vision Conf.*, Sep 2013, p. 11.

[82] Z. Li and D. Kulic, "Local shape context based real-time endpoint body part detection and identification from depth images," *Proc. of Int. Conf. on Comput. and Robot Vision*, pp. 219–226, 2011.

[83] V. M. Zatsiorsky., *Biomechanical basis of Human Movement.* Champaign, lL: Human Kinetics, 2002.

[84] B. Farnell, "Moving bodies, acting selves," *Annual Review of Anthropology*, vol. 28, pp. 341–373, 1999.

[85] I. F. Leong, J. J. Fang, and M. J. Tsai, "Automatic body feature extraction from a marker-less scanned human body," *Comput. Aided Design*, vol. 39, no. 7, pp. 568–582, 2007, human Modelling and Applications.

[86] A. Gritai, Y. Sheikh, C. Rao, and M. Shah, "Matching trajectories of anatomical landmarks under viewpoint, anthropometric and temporal transforms," *Int. J. Comput. Vision*, vol. 84, no. 3, pp. 325–343, Sep 2009.

[87] R. Easterby, K. Kroemer, and D. B. Chaffin., *Anthropometry and biomechanics: theory and application.* New York: Plenum Press, 1982.

[88] B. Bogin and M. I. Varela-Silva, "Leg length, body proportion, and health: A review with a note on beauty," *Int. J. of Environ. Res. and Public Health*, vol. 7, no. 3, pp. 1047–1075, 2010.

[89] D. Winter, *Biomechanics and Motor Control of Human Movement.* John Wiley & Sons, 2009.

[90] N. Hamilton, W. Weimar, and K. Luttgens, *KINESIOLOGY Scientific Basis of Human Motion.* McGraw-Hill, 2011.

[91] A. E. Chapman., *Biomechanical Analysis of Fundamental Human Movement.* Champaign, lll. Human Kinetics, 2008.

[92] J. Hamill and K. M. Knutzen., *Biomechanical basis of Human Movement*. Lippincott Williams and Wilkins, Wolters Kluwer, 2009.

[93] A. Kuo, "The six determinants of gait and the inverted pendulum analogy: A dynamic walking perspective," *Human Movement Science*, vol. 26, no. 4, pp. 617–656, Aug 2007.

[94] D. A. Winter, "Human balance and posture control during standing and walking," *Gait and Posture*, vol. 3, no. 4, pp. 193–214, Dec 1995.

[95] C. Maurer and R. Peterka, "A new interpretation of spontaneous sway measures based on a simple model of human postural control." *J Neurophysiol*, vol. 93, no. 1, pp. 189–200, 2005.

[96] H. Wang and K. Kosuge, "Control of a robot dancer for enhancing haptic human-robot interaction in waltz," *IEEE Trans. on Haptics*, vol. 5, pp. 264–273, 2012.

[97] T. Kwon and J. Hodgins, "Control systems for human running using an inverted pendulum model and a reference motion capture sequence," in *Eurographics Symposium on Comput. Animation*. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2010, pp. 129–138.

[98] I. D. Loram, S. M. Kelly, and M. Lakie, "Human balancing of an inverted pendulum: is sway size controlled by ankle impedance?" *J. Physiol*, vol. 532, no. Pt 3, pp. 879–891, 2001.

[99] H. Herr and M. Popovic, "Angular momentum in human walking," *The J. of Experimental Biology*, pp. 467–481, 2008.

[100] C. Rougier, J. Meunier, A. St-Arnaud, and J. Rousseau, "Fall detection from human shape and motion history using video surveillance," in *Proc. of 21st Int. Conf. Adv. Information Networking and Appl. Workshops*, vol. 2, 2007, pp. 875–880.

[101] M. Nixon and A. S. Aguado, *Feature Extraction & Image Processing for Computer Vision, Third Edition*, 3rd ed. Academic Press, 2012.

[102] Y. L. Lin and M. J. J. Wang, "Automated body feature extraction from 2D images," *Expert Syst. with Applications*, vol. 38, no. 3, pp. 2585–2591, 2011.

[103] H. Freeman, "On the classification of line drawing data," in *Models for the Perception of Speech and Visual Form*, E. W. Wather Dunn, Ed. MIT Press, Cambridge, MA, 1967, pp. 408–412.

[104] H. Foroughi, M. Alishah, H. Pourreza, and M. Shahinfar, "Distinguishing fall activities using human shape characteristics," in *Technol.l Develop. in Educ. and Automation*, M. Iskander, V. Kapila, and M. A. Karim, Eds. Springer Netherlands, 2010, pp. 523–528.

[105] K. Homma and E. Takenaka, "An image processing method for feature extraction of space-occupying lesions," *J. Nucl. Med.*, vol. 26, no. 12, pp. 1472–1477, 1985.

[106] R. Telea and J. J. V. Wijk, "An augmented fast marching method for computing skeletons and centerlines," in *Proc. of Symp. on Data Visualisation*, 2002, pp. 251–259.

[107] D. Valdes-Amaro and A. Bhalerao, "To boldly split: Partitioning space filling curves by markov chain monte carlo simulation." in *Proc. of Mexican Conference on artificial intelligence*, ser. Lecture Notes in Computer Science, vol. 5317. Springer, 2008, pp. 543–553.

[108] J. Sherrah, B. Ristic, and N. J. Redding, "Evaluation of a particle filter to track people for visual surveillance," in *Digit. Image Comput.: Techniques and Applications.*, Dec 2009, pp. 96–102.

[109] J. Czyz, B. Ristic, and B. Macq, "A particle filter for joint detection and tracking of color objects," *Image Vision Comput.*, vol. 25, no. 8, pp. 1271–1281, Aug 2007.

[110] N. Whitekey and A. Lee., "Twisted particle filters," *The Annals of Statistics.*, vol. 42, no. 1, pp. 115–141, 2014.

[111] A. S. Montemayor, J. J. Pantrigo, and J. Hernández, "A memory-based particle filter for visual tracking through occlusions," in *Proc. of the 3rd Int. Work-Conference on The Interplay Between Natural and Artificial Computation: Part II: Bioinspired Applications in Artificial and Natural Computation.* Berlin, Heidelberg: Springer-Verlag, 2009, pp. 274–283.

[112] J. Scharcanski, A. B. de Oliveira, P. G. Cavalcanti, and Y. Yari, "A particle-filtering approach for vehicular tracking adaptive to occlusions," *IEEE Trans. Veh. Technol.*, vol. 60, no. 2, pp. 381–389, 2011.

[113] P. Ribeiro and J. Santos-victor, "Human activity recognition from video: modelling, feature selection and classification architecture," in *Int. Workshop on Human Activity Recognit. and Modeling*, 2005.

[114] H. Qian, Y. Mao, W. Xiang, and Z. Wang", "Recognition of human activities using SVM multi-class classifier," *Pattern Recognition Letters*, vol. 31, no. 2, pp. 100–111, 2010.

[115] U. Mahbub, H. Imtiaz, and M. Ahad, "Action recognition based on statistical analysis from clustered flow vectors," *Signal, Image and Video Processing*, vol. 8, no. 2, pp. 243–253, 2014.

[116] A. Ma, P. Yuen, W. Zou, and J.-H. Lai, "Supervised spatio-temporal neighbourhood topology learning for action recognition," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 23, no. 8, pp. 1447–1460, Aug 2013.

[117] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?" in *Proc. of IEEE Int. Comput. Vis. and Pattern Recognit.*, June 2008, pp. 1–8.

[118] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *Proc. of IEEE Int. Comput. Vis. and Pattern Recognit.*, June 2008, pp. 1–8.

[119] X. Sun, M. Y. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in *Proc. of IEEE Int. Comput. Vis. and Pattern Recognit.*, 2009.

[120] F. Azhar and T. Tjahjadi, "Significant body point labelling and tracking," *IEEE Trans. on Cybern.*, vol. 44, no. 9, pp. 1673–1685, Sep 2014.

[121] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *Proc. of IEEE Conf. on Comput. Vision and Pattern Recognit.*, 2008.

[122] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognit.*, 2006, pp. 2161–2168.

[123] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognit.*, 2007.

[124] A. Klaeser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. of British Machine Vision Conf.*, 2008, pp. 99.1–99.10.

[125] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *Proc. of IEEE Int. Conf. on Computer Vision and Pattern Recognit.*, June 2008, pp. 1–8.

[126] M. Alcantara, T. Moreira, and H. Pedrini, "Real-time action recognition based on cumulative motion shapes," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, May 2014, pp. 2917–2921.

[127] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[128] M. Bregonzio, S. Gong, and T. Xiang, "Action recognition with cascaded feature selection and classification," in *Proc. of Int. Conf. on Imaging for Crime Detection and Prevention*, 2009.

[129] J. Bent, "Data-driven batch scheduling," Ph.D. dissertation, University of Wisconsin, Madison, May 2005.

[130] B. Arbab-Zavar, I. Bouchrika, J. N. Carter, and M. S. Nixon, "On supervised human activity analysis for structured environments." in *Proc. of 6th Int. Symposium on Visual Computing*, ser. Lecture Notes in Computer Science, vol. 6455. Springer, 2010, pp. 625–634.

[131] A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognit. Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[132] "A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views," *Int. Scholarly Research Notices*, pp. 1–36, July.

[133] A. Eweiwi, S. Cheema, C. Thurau, and C. Bauckhage, "Temporal key poses for human action recognition," in *Proc. of IEEE Int. Conf. on Computer Vision Workshops*, Nov 2011, pp. 1310–1317.

[134] F. Martinez-Contreras, C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, and S. Velastin, "Recognizing human actions using silhouette-based hmm," in *Proc. of IEEE Int. Conf. on Adv. Video and Signal Based Surveillance*, Sept 2009, pp. 43–48.