CHAPTER 2

LITERATURE REVIEW

The analysis of human movements found in videos has been researched by computer vision community for quite some time. However, the focus of this research work is to employ data fusion classification framework to human gait analysis using computer vision techniques. This chapter reviewed data fusion frameworks and techniques and, state-of-the-art research work on human gait analysis, an overview of the benchmark datasets and followed by a discussion on the relationship of data fusion strategies to human movement analysis. Finally, the chapter concludes with remarks that provide motivation for the research work presented in the following chapters.

2.1. CLASSIFICATION OF DATA FUSION AND TECHNIQUES

The existing literature report various multimedia analysis tasks accomplished through data fusion frameworks, according to the requirements of the applications. Different types of data fusion classifications are categorized as:

- JDL data fusion Classification where data fusion levels are defined by the JDL as level 0, level 1, level 2, level 3, level 4 [12].
- Dasarathy’s Classification systems based on the input/output data types and their nature, as proposed by Dasarathy [13].
- Classification based on abstraction levels of data such as fusion-feature, decision and hybrid [7].
2.1.1. JDL Data Fusion Classification

This classification is the most popular conceptual model in the data fusion community. It was originally proposed by JDL and the American Department of Defense (DoD). These organizations classified the data fusion process into five processing levels, an associated database, and an information bus that connects the five components in Fig. 2.1. The five levels could be grouped into two groups, low-level fusion and high-level fusion, which comprise the following components:

![Fig. 2.1 JDL data fusion framework](image)

The five levels of data processing are defined as follows.

- **Level 0—Source preprocessing:**
  
  Source preprocessing is the lowest level of the data fusion process, and it includes fusion at the signal and pixel levels. In the case of text sources, this level also includes the information extraction process. This level reduces the amount of data and maintains useful information for the high-level processes;

- **Level 1—Object refinement:**

  Object refinement employs the processed data from the previous level. Common procedures of this level include spatio-temporal alignment, association, correlation, clustering or grouping techniques, state estimation, and the removal of false positives, identity fusion, and the combining of features that were extracted from images. The output results of this stage are the object discrimination (classification...
and identification) and object tracking (state of the object and orientation). This stage transforms the input information into consistent data structures;

- **Level 2—Situation assessment:**
  This level focuses on a higher level of inference than level 1. Situation assessment aims to identify the likely situations given the observed events and obtained data. It establishes relationships between the objects. Relations (i.e., proximity, communication) are valued to determine the significance of the entities or objects in a specific environment. The aim of this level includes performing high-level inferences and identifying significant activities and events (patterns in general). The output is a set of high-level inferences;

- **Level 3—Impact assessment:**
  This level evaluates the impact of the detected activities in level 2 to obtain a proper perspective. The current situation is evaluated, and a future projection is performed to identify possible risks, vulnerabilities, and operational opportunities. This level includes (1) an evaluation of the risk or threat and (2) a prediction of the logical outcome;

- **Level 4—Process refinement:**
  This level improves the process from level 0 to level 3 and provides resource and sensor management. The aim is to achieve efficient resource management while accounting for task priorities, scheduling, and the control of available resources.

### 2.1.2. Dasarathy’s Classification

Another well-known data fusion classification model was provided by Dasarathy and is composed of the following five categories; based on the input/output data types and their nature (Fig. 2.2):

- **Data in-feature out (DAI-FEO):**
  At this level, the data fusion process employs raw data from the sources to extract features or characteristics that describe an entity in the environment;

- **Feature in-feature out (FEI-FEO):**
  At this level, both the input and output of the data fusion process are features. Thus, the data fusion process addresses a set of features with to improve, refine or obtain new
features. This process is also known as feature fusion, symbolic fusion, information fusion or intermediate-level fusion;

- **Feature in-decision out (FEI-DEO):**
  This level obtains a set of features as input and provides a set of decisions as output. Most of the classification systems that perform a decision based on a sensor’s inputs fall into this category of classification;

![Diagram](image)

Fig. 2.2 Dasarathy’s Input-Output model

- **Decision In-Decision Out (DEI-DEO):**
  This type of classification is also known as decision fusion. It fuses input decisions to obtain better or new decisions. The main contribution of Dasarathy’s classification is the specification of the abstraction level either as an input or output, providing a framework to classify different methods or techniques.

2.1.3. **Classification Based on the Abstraction Levels**
Luo, *et al.*, [14] provide the following data fusion classification based on the abstraction levels.
• *Early fusion:*
  In early fusion approach, the features extracted from input data are combined, to produce new raw data that are expected to be more informative than the inputs initially and represented using compression techniques and later used as input to perform further analysis. This fusion is advantageous in the utilizing of correlation between features. However, it is hard to represent the time synchronization. In addition, to represent features of different modalities in the same format is difficult. Also, increase in the number of modalities it becomes complex to handle.

![Diagram](image)

**Fig. 2.3** Demonstration of classification based on abstraction levels

• *Late fusion:*
  Late fusion, involves learning of combination rules across multiple decision streams (ranked lists or classifier responses) using a certain amount of data with the ground truth as a validation set. The objective is to obtain limited relevant features. This level of fusion allows features from different representations to be combined in the same format of representation. Late fusion allows more flexibility and scalability. On the other hand, one major disadvantage is, late fusion requires a lot of training data sets due to scalability otherwise known as ‘curse of dimensionality.’

• *Hybrid fusion:*
  A hybrid fusion is used to utilize the advantages of both early and late fusion strategies. Therefore, many researchers have opted to use the hybrid fusion strategy to solve various kinds of multimedia analysis problems. One prominent disadvantage this level of data fusion relies much on the data that is preprocessed and can yield ‘mutual disambiguation’.

The following Table 2.1 gives a comparison of the early fusion, late fusion and hybrid fusion[8]:

### Data Fusion Techniques

Data fusion techniques are classified into three types:

1. Data Association (Correlation),
2. State estimation (Prediction) and
3. Decision Fusion

The goal of correlation is to establish a set of measurements that are generated by an object in observation over time. Hall and Llinas [8] provided a definition of data association as”

The process of assign and compute the weights that relates the observations or tracks (A track can be defined as an ordered set of points that follow a path and are generated by the same target.) from one set to the observation of tracks of another set.” The Correlation or data association is applicable to all levels of fusion and usually depends on the type of granularity and objective to be met at each level. There are several techniques used to solve correlation problems. Nearest Neighbor and K-means, the probabilistic data association (PDA) algorithm, Joint probabilistic data association (JPDA) algorithms are some of the techniques used to solve Correlation.

State estimation methods aim to determine the state of the object under observation; typically the position is determined by the measurements determined by correlation. The

---

<table>
<thead>
<tr>
<th>Level of Fusion</th>
<th>Early Fusion</th>
<th>Late Fusion</th>
<th>Hybrid Fusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Level</td>
<td>Feature Level</td>
<td>Feature Level</td>
<td>Feature and Decision Level</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Input Type</th>
<th>Raw data</th>
<th>Closely related modes</th>
<th>Loosely related modes</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Noise/ Failures</th>
<th>Highly susceptible to noise or failures</th>
<th>Less sensitive to noise or failures</th>
<th>Highly resistant to noise or failures</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Usage</th>
<th>Not widely used</th>
<th>Used for particular modes</th>
<th>Most widely used type of fusion</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Drawbacks</th>
<th>No pre-processing may result in concatenation of data rather than fusion, redundant data</th>
<th>Require a lot of training data sets and due to ‘curse of dimensionality.’</th>
<th>Relies much on the data that is pre-processed, can yield ‘mutual disambiguation.’</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Application Examples</th>
<th>Combining two video streams, text and images</th>
<th>speech recognition from voice and lips</th>
<th>Tracking and recognition of human face and gait.</th>
</tr>
</thead>
</table>

---

Table 2.1  Fusion of multimedia data
common methods include maximum likelihood and maximum posterior, the Kalman filter, distributed particle filter and Covariance consistency. Data association (Correlation) is often performed before the state estimation (Prediction) of the detected targets. Moreover, it is a key step because the estimation or classification will behave incorrectly if the data association phase does not work coherently. In general, state estimation techniques fall under the Level 1 of the JDL data fusion framework.

Decision Fusion techniques aim to derive high-level meaning from the knowledge of the situations which is provided by different sources by way of reasoning. The Level 4 of JDL data fusion framework deals with the decision fusion techniques. Some of the techniques include Bayesian methods, Dempster-Shafer Inference, Abductive reasoning and semantic methods [10].

2.2. OVERVIEW OF HUMAN MOTION ANALYSIS

This section focuses on vision based human motion analysis. The work to perform the analysis of human walking motion is carried out in different phases starting from extracting human body shape silhouettes to pose analysis. There are several applications relating to human motion analysis. In general, human motion analysis follows an outline with a major focus on three major tasks: human detection, human motion tracking and analysis of motion for behavior understanding [15].

Initially, human motion is detected, and the human body shape silhouettes are extracted from the videos, later tracking is performed to find the movement of the change in actions, and finally human poses are estimated and analyzed from human body silhouettes. When an action is performed by a human, a sequence of poses is generated. Out of thee poses generated, there can be cases where a pose is deviating from the normal sequence and observations can be noted for change in stances of postures while walking, running and bending or other poses.
2.2.1. Human Motion Methodologies
This section gives an overview of the approaches used for analyzing the human walking action. The methodologies covered in Table 2.2, describe various types of representation of human body based on shape models like stick figures, two dimensional contours, volumetric models, kinematic constraints and appearance.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Methodologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Rius, et al. [16]</td>
<td>Learn various 3D human postures from training sequences, use matching algorithm based on dynamic programming to map different postures and different motion cycles.</td>
</tr>
<tr>
<td>F. Korc, et al. [17]</td>
<td>Use 2D articulated models in single view sequences for detecting and tracking of Humans in three steps, (1) Detecting human candidates, (2) Validating the model of a human and (3) Tracking of the model in consequent frames. The model is developed with a six-link model and an articulated head of frontal view of a person.</td>
</tr>
<tr>
<td>I. Mikic, et al. [18]</td>
<td>Develop an integrated system for automatic acquisition of the human body</td>
</tr>
</tbody>
</table>
model and motion tracking using input data acquired from multiple synchronized video streams. Tracking is performed on the 3D voxel reconstructions computed from the 2D foreground silhouettes; the human body model consists of ellipsoids and cylinders.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Methodology</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. Ramanan and D. A. Forsyth, [19]</td>
<td>Appearance based model is built on people in motion and then clustering of candidate body segments is formulated and then the model is used to find all individuals in each frame.</td>
</tr>
<tr>
<td>H. Ning, et al. [20]</td>
<td>A motion model is built from the semi-automatically acquired training data and motion constraints are computed by analyzing the dependency of joints.Both of them are integrated into a dynamic model in order to reduce the size of the sample set.</td>
</tr>
<tr>
<td>P. Viola, et al. [21]</td>
<td>Combine two methodologies by integrating information of appearance with motion information and detection algorithm along with trained sequences. Both motion and appearance information is used to detect a walking person.</td>
</tr>
<tr>
<td>J. J. Gonzalez, et al. [22]</td>
<td>A robust feature-based tracking by initializing a standard point-wise tracker method and grouping image points undergoing the same rigid motions is built to track human motions of different body parts without articulation.</td>
</tr>
<tr>
<td>A. D. Sappa, et al. [23]</td>
<td>A technique that combines the prior knowledge regarding a person’s motion with human body kinematics constraints is developed. It uses an efficient feature point selection and tracking approach to computing feature points’ trajectories and then 3D motion models associated with each joint are locally obtained by using key frames, meaning frames where both legs are in contact with the floor.</td>
</tr>
<tr>
<td>L. Wang, et al. [24]</td>
<td>Automatic person recognition from body silhouette and gait is performed where background subtraction procedure is combined with a simple correspondence method to segment and track spatial silhouettes of a walking figure. Simple feature selection and parametric Eigen space representation are used to reduce the computational cost during training and recognition.</td>
</tr>
<tr>
<td>J. H. Yoo and M. Nixon, [25]</td>
<td>The extraction of the gait figure is made using 2D stick figures from the body contour by determining the body points using linear regression and motion tracking with topological analysis. Then, the detection of the gait cycle is computed by symmetry analysis, kinematic analysis, and feature extraction to classify the gait patterns.</td>
</tr>
</tbody>
</table>

Table 2.2 Methodologies of human motion activity
2.2.2. Human Detection or Segmentation

A computer vision application where human motion analysis is carried out the initial step is to detect moving human. Detecting moving humans initially provide a focus of attention for future tasks. The objective of segmentation is to obtain relevant information about the human figures located in the frames of human motion. There are various approaches available to detect moving human figures in videos. Commonly used segmentation techniques are given below:

a) Background subtraction technique where the foreground is extracted from the background with a fixed camera at a fixed angle and the background is completely stable.

b) Temporal differencing technique unlike in background subtraction, the background is dynamic. The moving object is detected by taking the difference between two or three consecutive frames and is combined with the figure motion edges [26].

c) Optical flow technique is applied when both camera and humans are in motion, called ego motion [27].

In video analysis, for detecting moving humans or objects found in video sequences, background subtraction technique is recommended. The basic idea of background subtraction is the technique of subtracting the current frame from a reference frame which has been generated from a static background through an episode of time. Other challenges such as illumination changes, dynamic background, occlusion and removal of shadows are usually considered as sub problems to address the problem of locating humans and extracting human figures in video sequences. Broadly, two approaches can be used to segment video images to extract moving humans, Global- top down approach and Local- bottom up approach. In a Global approach, the region of interest (ROI) is taken as a whole frame, and the moving person is localized using background subtraction methods. However, accurate localization, background subtraction, viewpoints, noises, shadows and occlusions are to be handled carefully to get an effective performance. Where as in Local approach- relevant region of interest (ROI) is created initially by detecting spatial or/and temporal interest points around the moving humans, and local patches are created. After which calculations around the points and patches are combined into a final representation.
This approach is less sensitive to noise, partial occlusion and sometimes will not require background subtraction. However, care is to be taken on how to extract the relevant ROI.

2.2.2.1. Global-Top Down Approach

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

According to [28], a background subtraction must adapt to gradual or fast illumination changes (changing the time of day, clouds, etc.), motion changes (camera oscillations), high frequency background objects. In the global – top down approach, A) Background modeling techniques, B) Foreground extraction methods and C) Morphological operations are found.

A) Background Modeling

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

i) Non-Recursive Techniques

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

- Frame Differencing:

Here, the background model is the frame at time t-1, i.e., the previous frame is modeled, and the incoming frame is subtracted with a threshold.

\[ |I_t - I_{t-1}| > T \]  

(2.1)
Where $I_t$ is the intensity of frame $t$ and $T$ is a fixed threshold. This technique is sensitive to a threshold value. Only threshold value can influence which does not give an accurate result and is found to be sensitive to noise. The advantages include high adaptability, less memory space and less computational load. Another main disadvantage is that when human is found to be stationary even for a very short time in the subsequent frames, the human figure becomes background [30].

- **Median Filtering:**
The estimated background value of each pixel in the background model is calculated as the median of that pixel in all frames chosen in the buffer. This is another effective not requiring high computational load but disadvantage is the requirement of $(N \times \text{frame size})$ [29].

- **Linear Predictive Filter:**
The background estimate is computed by applying a linear predictive filter on the pixels of the frames in the buffer. The filter coefficients are estimated depending on the sample covariance at each frame time [30].

ii) **Recursive Techniques**
In these techniques buffer of frames is not used instead the background image is updated recursively. The advantage is only one frame will be updated every time a new frame is received. Another advantage is that these techniques can handle multi-modal backgrounds. However, they are computationally complex and sensitive to illumination changes [31].

- **Running Average:**
A simple background modeling algorithm is computed as

$$B_{t+1} = a C_t + (1-a) B_t$$  \hspace{1cm} (2.2)

Where $B$ stands for background and $C_t$ is the current frame. 'a' is defined as the learning rate with a typical value of 0.05 [32].
• **Approximated Median Filtering:**
The technique was developed by McFarlane and Schofield [33]. Initially, a background estimate is taken and when a pixel in the current frame has a gray value which is larger than the pixel in the background estimate, it is incremented by one. On the other hand, when the value of a pixel in the current frame has a value lower than the background estimate, the pixel in the background estimate is decremented by one. When applying this function to the background model, the model converges to an estimate where half the input pixels are greater than the background and the other half are less than the background model. This technique gives a satisfactory result but is slow to adapt to the big changes in the real background.

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

$$B(t) = A(t)B(t-1)+K(t)[z(t) – H(t)A(t)B(t-1)]$$ (2.3)

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

• **Mixture of Gaussians:**
A Mixture of Gaussians uses 3 to 5 Gaussian distributions simultaneously. Each pixel is modeled by a mixture of Gaussians that will sum to form a probability distribution function $f(x_t)$ [34].

$$f(x_t) = \sum_{i=1}^{k} \omega_{i,t} \cdot \eta(x_t - \mu_{i,t}, \Sigma_{i,t})$$ (2.4)
Where, \( f \) is the probability distribution function, ‘x’, a certain pixel value at time ‘t’ and ‘k’ is a value set between 3 and 5. \( \omega_{i,t} \) is an estimate of the weight of the \( i^{th} \) Gaussian in the mixture and \( \mu_{i,t} \) is the mean value of the \( i^{th} \) Gaussian in the mixture. \( \eta \) is the mean of Gaussian function estimate value in next frame. This algorithm is simple and of low complexity. This algorithm works well for stable scenes and segmentation becomes difficult when both background and foreground is complex.

B) Foreground Extraction

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

- **Shape-based Representation:**
  Foreground extraction can be represented in the different description of shapes such as point, box, silhouette and blob. Collins, L., et al. [35] represented moving objects as blobs. Kuno and Wantanbe [36] use shape parameters of human silhouettes. Representations such as cylinders, cones, ellipsoids and super-quadrics are found [37]. Silhouettes and contours are other representations which are commonly in use. Contour representation defines the boundary of an object. The region inside the contour is called the silhouette of the object. Silhouette and contour representations are suitable for complex non-rigid shapes [38]. Skeleton representation is used as shape representation for both articulated and rigid motions. The skeleton is extracted by applying medial axis transform to the human silhouette [39]. Surface representations such as generic mesh models are described in [40].

- **Kinematic Structure:**
  The kinematic structure comprises of a fixed number of joints with specific degrees-of-freedom. Kinematic initialization is carried such that the lengths of the limb are considered
for the purposes analyzing human walking motion. Further, many approaches are used to build trajectories by placing 3D markers, finding skeletal symmetry of the postures and anthropometric constraints between the ratio of limb lengths are used to allow estimation of the kinematic structure for an unknown scale factor [41].

• **The 3-Dimensional Representation:**
  
  Human figure representation in 2-dimensional can be extended to 3-dimensional representation. Weinland, D., et al. [42] combined silhouettes from multiple cameras into a 3D Voxel model. Multiple images over time can be stacked to form a D-dimensional space-time volume.

• **Appearance:**
  
  Directly persons appearing in the video sequences are observed and statistical models are developed like mixture of Gaussian [43], histograms of color, texture combination [44]. Texture maps are derived in [45]. Approaches such SVM, AdaBoost were successful to learn body parts available during training [46]. Lim, H., et al. [47] address the problem of changing appearance of humans due to motion by mapping pixels inside a bounding box [48].

C) **Morphological Operations**

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

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

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

2.2.2.2. Local-Bottom Up Approach

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

A) Interest point detection
   B) Computations on the localized regions-
      - Smoothing and dimensionality reduction,
      - Representations,
      - Correlations.

A) Interest Point Detection:
Interest point detectors are used to find regions of interest in images. Some of the early and successful methods are Harris interest point detector [49], KLT detector and SIFT detector. Laptev and Lindeberg [50] used Harris corner detection along with 3D for both spatial and temporal domains. Kadir and Brady [51] included methods to calibrate camera motions along with locating 2D salient points and then converting to 3D by computing entropy within each cuboid generated and centres with local maximum energy are selected as salient points. Dollar, P., et al. [52] applied Gabor filtering along spatial and temporal domains and the interest points were calculated using neighborhood points with local minima. Willems, G., et al. [53] computed salient points with the help of a 3D Hessian matrix. Wang, H., et al., [54] used dense sampling for interest point detection along space and time dimensions.

B) Computations on the Localized Regions
Once the regions of interest are computed, the following computations can be carried out.
• **Smoothing, Dimensionality Reduction**

Space and time patches representations can be further processed for smoothing and reducing of state-space search. Schuldt, C., *et al.* [55] calculate patches of the region of interest and compute normalized derivatives in space and time. Niebles, C., *et al.* [56] applied smoothing and dimensionality reduction using Principal component analysis (PCA). Jhuang, H., *et al.* [57] use several phases of computations starting with applying Gabor filters to dense flow vectors, next phase is followed by a low max operation then global max operation is applied and finally matching is performed to obtain a final set of patches. Comparing patches are usually carried out using codebook where clustering patches and selecting the close related patches as codewords.

• **Representations**

Grids can be used to bin patches spatially or temporally. Ikizler and Duygulu [58] used spatial grid approach to sample oriented rectangular patches which bin into a grid. Nowozin, S., *et al.* [59] used temporal instead of a spatial grid. PCA reduced and extracted interest points are mapped on to codebook indices. Laptev and Perez [60] use spatio-temporal grid representation and Bregonzio, M., *et al.* [61] do not use any local image descriptor rather they look at all interest points within cells of the spatio-temporal grid with scales.

• **Correlations**

Another important task is to exploit correlations between the interest point patches to select relevant image descriptors for proceeding to next level of understanding images. Scovanner, P., *et al.* [62] construct a co-occurrence matrix of features is constructed and iteratively merged until all are above a specified threshold. Several other researchers have contributed using supervised, semi-supervised and unsupervised approaches. Supervised learning methods require a large number of samples. Adaptive boosting and Support Vector Machines are commonly used.

### 2.2.3. Tracking Human Motion

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

The most widely used mathematical tools for tracking are - Kalman Filter [63], the Condensation algorithm [64], Dynamic Bayesian Network [65]. Tracking strategies can be divided into various categories according to the suitability of different applications of human walking:

1. Tracking of human body parts such as hand, face, and leg [66]
2. Tracking of the whole body
3. Total number of views - Single-view[67], Multiple-view [68], and Omni-directional view [69]
4. The 2-Dimensional and 3-Dimensional Tracking
5. Tracking environment-indoors and outdoors
6. The number of humans to be tracked - single human, multiple humans and human groups.
7. The camera’s state - moving and stationary [17].

Yilmaz, A., et al. has described three types of tracking representations in [70]. Point correspondence, primitive geometric models and Contour evolutions which are relevant to human walking motion tracking also.

2.2.4. Feature Extraction

The feature is defined as a function of one or more measurements, each of which specifies some quantifiable property of an object, and is computed such that it quantifies some significant characteristics of the object. We classify the various features currently employed as follows:

A) General features

Application features such as color, texture, and shape. According to the abstraction level, they can be further divided into:
- Pixel-level features: Features calculated at each pixel, e.g. color, location.
- Local features: Features calculated over the results of subdivision of the image band on image segmentation or edge detection.
- Global features: Features calculated over the entire image or just regular sub-area of an image.

B) Domain-specific Features

Application dependent features such as human faces, fingerprints, and conceptual features. These features are often a synthesis of low-level features for a specific domain.

On the other hand, all features can be coarsely classified into low-level features and high-level features. Low-level features can be extracted directly from the original images, whereas high-level feature extraction must be based on low-level features [71].

2.2.5. Pose Analysis of Human Walking Motion

In computer vision tasks, Pose analysis is considered to be the final task in accomplishing an understanding of human walking motion. Reliable pose analysis of human walking motion can be performed using gait recognition approaches as reported in the literature. The pose analysis of human walking motion can be processed in two steps:

1) Pose estimation and
2) Action recognition.

A) Pose Estimation

The purpose of pose estimation is to estimate the type of articulation and movement a pose is taken during walking. Poses can be constructed by supervised or unsupervised learning. Supervised learning is used where the unknown poses are compared with predefined poses already recorded in the training whereas for unknown poses patterns are not available beforehand, self-organising and self-learning techniques are used.

Moesland, T. B., et al. [72] separated pose estimation algorithms into three categories: Model-free, where no explicit priori models and direct mapping of 2-dimensional sequences to 3-dimensional poses are performed. Indirect model, a priori model in pose
estimation as a reference to help in the interpretation of measured data. Direct model, use an explicit 3-dimensional geometric representation of human shape and kinematic structure and use an analysis-by-synthesis methodology to optimize similarity between estimated and observed images.

Recent contributions include Kinect pose estimation system [73] based on 3D scene information, and is the only currently available system to our knowledge that reaches satisfactory pose estimation performance in real-world scenarios. Images are depth based, and Kinect system enables easy data acquisition for new pose estimation approaches.

Singh, V. K., et al. [74,75] describe an approach incorporating contextual knowledge into the pose estimation, about human interaction, and significantly improve pose estimation performance based on a combination of Pictorial Structures and a pose estimation approach by Deva Ramanan [76].

Johnson and Everingham [77] propose a new, large scale data acquisition method and they introduce Pictorial Structures in an extended framework called Clustered Pictorial Structure Models. It can deal with extremely articulated poses and reaches very high pose estimation performance. Another useful contribution is by Wang and Koller [78] where, high level information from Pictorial Structure matching can be combined with low level per pixel segmentation information. An energy function is assembled that combines these different levels of information. Wang and Koller apply the relaxed dual composition to include infeasible energy functions in the optimization process. Their concept combines multiple methods and improves their common results.

B) Action Recognition
Pose estimation is followed by Action recognition to identify and investigate actions. There are goals like distinguishing regular walk from irregular walking movements, determining different types of walking; detecting sudden falls and other applications can be represented. Holistic approaches attempt to recognize a person, find gender or perform simple actions like walking or running. With respect to specific tasks relating to particular body parts it is more advantageous to consider relating body parts instead of taking into account the entire body, according to survey paper of Moeslund, T. B., et al. in the
literature, context of actions were classified by researches as following, Nagel [79] suggested a hierarchy of change, event, verb, episode, and history. Bobbick [80] proposed different levels of abstraction – movement, activity and action. Moeslund, T. B., et al. [74] used action hierarchy as - action primitives, actions, and activities.

2.2.6. Gait Recognition Approaches
Existing methods for gait recognition can be divided into two main categories.

- Shape-based/ model-free/ appearance-based/ motion-based/ or silhouette based
- Model-based or structure from motion.

Table 2.3 gives the details of the existing Gait recognition approaches found in the literature.

<table>
<thead>
<tr>
<th>Shape-based/ model-free/ appearance-based/ motion-based/ or silhouette based approach</th>
<th>Model-based or structure from motion approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation between probe silhouette image sequence and those in data set [81]</td>
<td>2-D stick model [82]</td>
</tr>
<tr>
<td>Using low level image features, identifying spatial and temporal variances [83]</td>
<td>Fit models to extract features that describe dynamics of gait [84]</td>
</tr>
<tr>
<td>HMM based feature representing silhouette width distribution [85]</td>
<td>Articulated model [86]</td>
</tr>
<tr>
<td>Using IR images [87]</td>
<td>Link feature trajectories [88]</td>
</tr>
</tbody>
</table>

Table 2.3 Existing Gait recognition approaches

2.3. COMMON DATASETS

In the recent years, many public video datasets are available for performing various research tasks such as segmentation, tracking, recognition, motion analysis and so on. The advantages are that; they save time and resources that there is no need to record new video sequences. A suitable dataset helps in performing tasks to be more reliable and the use of same datasets facilitates comparison of different approaches and gives an insight of various
methods and algorithms. The following Table 2.4 gives the list of the common public data sets which are useful for analyzing human walking motion in the video sequences.

Walking is one of the actions found in all the datasets mentioned in Table 2.4. The ground truth and various walking styles, running, bending actions are also included for the purpose of focusing on the different practical type of scenes, the number of actions. KTH, Weizmann datasets were recorded to study new algorithms to improve performance human actions of which one set is only for human walking movement.

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

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

<table>
<thead>
<tr>
<th>Type of Recording environment, camera movement</th>
<th>Datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indoor and outdoor, simple and static background. Static camera movement</td>
<td>KTH [89], Weizmann [90]</td>
</tr>
<tr>
<td>Indoor and outdoor, complex without static background and illumination conditions are not controlled. Static camera movement</td>
<td>CAVIAR [91], ETISEO [92], CASIA Gait recognition A [93], CASIA Gait recognition B [95]</td>
</tr>
<tr>
<td>Indoor, multi-view recording. Static camera movement</td>
<td>MuHAVi [94]</td>
</tr>
<tr>
<td>Outdoor, multi-view recording. Static camera movement</td>
<td>CASIA [95]</td>
</tr>
<tr>
<td>Online Repository. Several camera movements</td>
<td>VISOR [95], VIRAT [96]</td>
</tr>
<tr>
<td>Web Videos. Several camera movements</td>
<td>HMDB51[97]</td>
</tr>
</tbody>
</table>

Table 2.4 Datasets for action recognition-Human walking motion
CASIA action, CASIA Gait recognition A, CASIA Gait recognition B are three different datasets developed to research biometrics and intelligent surveillance consisting of various sequences of human activities of walking, running, bending and other gait related actions.

MuHAVi was developed for the purpose of evaluating silhouette-based human action methods. It provides a realistic challenge for illumination changes and segmentation.

VISOR, VIRAT are two large datasets collected in natural scenes showing people in various viewpoints, resolutions and background clutter.

HMDB51, videos were collected from various sources from movies and actions are categorised into five types: general facial actions, facial actions with object manipulation, general body movements, body movements with object interaction and body movements for human interaction.

2.4. RELATION BETWEEN DATA FUSION AND HUMAN MOVEMENT ANALYSIS

The relation between data fusion and human movement analysis is the paradigm that is chosen for the purpose of handling the tasks associated with multimedia content (videos) in this research. Hence, this section elaborates on human movement analysis and data fusion models available in the literature and compares with the tasks related to the proposed work.

2.4.1. JDL Model and Human Movement Analysis Tasks

In order to carry out human movement analysis data fusion plays a contributory role in improving the detection of motion, tracking, and description of behaviors. This section analyzes human movements from the viewpoint of the JDL process model.

- At JDL Level 0: Image and video processing are employed to enhance specific properties of the input video sequences.
- At JDL Level 1: Contains most of the works defined for human movement analysis such as gestures, actions and even group activities analysis.
• At JDL Level 2: Works are demonstrated on interactions referring to human-human or human-object.
• At JDL Level 3: Human movements correspond to the prediction of future actions that a person is going to do.
• The Level 4: JDL process model has not been exploited from a human movement perspective.

2.4.2. Dasarathy’s Input-Output Model and Human Movement Analysis Tasks

Human Action Recognition methods employing multiple cameras are defined at FEI-FEO, FEI-DEO and DEI-DEO levels.

• FEI-FEO data fusion level: Diverse methods have been defined at the FEI-FEO data fusion level to combine the information obtained from multiple cameras. Different strategies have been defined to divide this works into three different categories:

Methods projecting 2D features to 3D: A popular approach is to recover the 3D shape projecting 2D silhouettes and recovering the visual hull. Visual hull reconstruction requires accurate silhouette segmentation at the different available views.

Methods combining features in a subspace: Methods to compute features for the 2D views available and combine them employing some simple scheme. The averaging of the multiple features representing pose, global and local motion has been proposed improving the results with respect to other alternatives. A joint Bag-of-Words histogram might be constructed with the local feature descriptors obtained for each one of the Views;

The last class of methods is based on computing a measurement of the quality of each view available. A first approach to the selection of the best view is made estimating the orientation of the human with respect to the camera and measurement based on the properties of the silhouette has been proposed.

• FEI-DEO data fusion level: The next category of works examined employing multiple views of the scene for the recognition of human actions are those defined at the FEI-DEO level. This works to model the existing correlations among the multiple observations in the
structure of the classifier employed for the prediction of the actions. The concatenation of the input features is the most straightforward procedure to perform the fusion. The Fused HMM proposes to model correlations among observations coupling the values of the hidden state chains of parallel HMMs defined for each view.

- **DEI-DEO data fusion level:** The last category of works employing multiple views performs the fusion at the DEI-DEO level, combining the outputs of action classifiers applied to each one of the camera views. Majority voting has been the most common technique for the fusion of decisions. A weighted voting strategy has been proposed in, correcting each vote according to the value of the observed feature. Cilla, R., et al. have proposed to learn an error model to weight the predictions made from the different cameras, improving the overall result [98].

### 2.5. REMARKS

This Chapter provided an extensive literature review on data fusion frameworks, data fusion techniques of correlation and prediction, human motion (gait) analysis and the relation between data fusion and human movement analysis. For human gait analysis, data fusion was implemented with a fusion of data from multiple cameras, sensors or multiple views. But, the visual content of videos recorded by a single camera, itself has multiple data such as visual features (color, texture, shape), motion features (direction, temporal information, motion patterns, variation), metadata and other information which can be combined to perform meaningful interpretations, for which, the work is yet to be implemented. Following points were noted from the review, to provide motivation for the proposed work presented in the thesis.

- Exclusively work was accomplished for human gait analysis; Limited work is available where human gait analysis referred to and model or framework.
- works on correlation were scanty
- Prediction was exclusively handled by behavior models
- Paradigm of data fusion applied very less for human motion analysis