CHAPTER 3

PROPOSED METHODOLOGY

The aim of this research work is to investigate various techniques of data fusion and computer vision used in the analysis and representation of multimedia content (videos) for the purpose of correlation and prediction of human walking motion (or gait). After the extensive literature review presented in chapter two, this chapter gives a detailed explanation of the proposed work carried out during in this research. The scope of this work has three implications namely –

i) Since the size and target of semantics found in multimedia documents such as scene, objects, people, events, sounds, texts and so on are usually very large, therefore, in this research work more emphasis will be put on algorithms and methods relevant to human walking movements (gait) found in multimedia content particularly videos of single person and single camera.

ii) More focus will be given to the novel framework designed to solve applications through the levels of abstraction which is demonstrated by human walking movement in this work.

iii) The work of analyzing and predicting the types of human gait poses as normal and to distinguish normal gait poses from the gait poses at the point of transition are dealt with respect to the proposed framework defined by data fusion abstraction levels and not using to any existing model.
3.1. DETAILS OF THE PROPOSED METHODOLOGY

There are various representational methodologies and frameworks that have been proposed for analysis and representation of multimedia content (videos). But the proposed methodology used in this work is experimental and uses two step approach, namely,

a. Design of Tri-level Unified Framework and  
b. Analysis of Human Gaits in video sequences

The sequence of tasks used for analyzing human gait is aligned to the levels defined by the Tri-level Unified Framework. Fig. 3.3 gives an overview of the proposed work.

3.1.1. Design of Tri-level Unified Framework

The first task in this research work is to design a theoretical framework which employs data fusion. Data fusion plays a central role in many applications as it has the advantage of combining data to estimate or predict states of observations. According to Steinberg, E., et al. [99] the process of data fusion is defined as “a process of combining data or information to estimate or predict entity states”. Data fusion can occur at different abstraction levels of information. Initially, data fusion techniques have been used in multisensory environments with the aim of fusing data from different sensors. However, the proposed Tri-level Unified Framework describes the abstract levels for data description of the multimedia content analysis (videos). The visual content found in the video sequences are modelled as a hierarchy of abstractions [9] at the three levels namely, 1. Data level, 2. Feature descriptor level and 3. Decision level. Therefore, a unified framework (Fig. 3.1) is proposed to solve the tasks relating to analysis and representation of multimedia content (videos) for analysing human walking motion (or gait).

The Fig. 3.1 illustrates the following:

2. **Data Fusion:** In the data fusion we fuse together the information contained in the multiple forms at various levels depending on the complexity of the data at data level, feature level and/or decision level.

   a. Data level – At this level the different forms of multimedia content can be combined and compression can be made to reduce the size of the storage. After the fusion further processing can be done for analysis.

   b. Feature Descriptor level – At this level, a common representation can be evolved for the compressed data reduced at data level so that the input for further processing “speak a common language”. This involves several processing tasks such as spatial and temporal alignments, feature extraction, semantic and radiometric alignments.

   c. Decision level- this level includes the common representation and decision labelling

   For example, to analyse and interpret human walking present in a video sequence, the Data level extracts silhouettes for minimal representation, i.e., reduce the size of the storage. The Feature Descriptor chooses silhouettes and rearranges them into a common representation and builds a feature descriptor. The Decision level uses the feature descriptor to analyse for a meaningful interpretation of walking.

3. **Compressed format:** The direct input of video stream is considered to be large in size. Therefore, different methods can be used to combine and compress the data to obtain reduced storage size by defining a minimum representation scheme.

4. **Common Representation:** The principal functions in the common representational are

   a. Spatial Alignment - The input frames/segments are spatially aligned with the same geometric base. Without a common geometric base, any information derived from a given input image cannot be associated with other spatial information. The accurate spatial alignment of the input images/ sequence of frames is therefore a necessary condition for multimedia fusion.
b. Temporal Alignment - The spatially aligned input images/ scenes/ frames/audio segments are temporally aligned to a common time. This step is only required if the input images/ scenes/audio are changing or evolving in time. In this case, the accurate temporal alignment is a necessary condition for multiple data fusion.

c. Feature Extraction - Characteristic features are extracted from the spatially and temporally aligned input images/ frames. The output is one or more feature maps for each input image/ frame/segment of audio.

d. Radiometric Calibration - The spatially, temporally and semantically aligned input images/ videos/text and feature maps are converted to a common measurement scale. This process is known as radiometric calibration.

Fig. 3.1 Proposed Tri-level Unified Framework
5. **Decision labelling**: Pixels in each spatially and temporally aligned input image/frame or feature map are labelled according to a given set of criteria. The output is a set of decisions.

6. **Semantic Equivalence**: In order for the input images/frames/segments of audio, features or decision to be fused together they must refer to the same object or phenomena. The process of linking the different inputs to a common object or phenomena is known as semantic equivalence.

The proposed Tri-level Unified Framework is used as a baseline approach for systematically dividing the tasks associated with human gait analysis into three levels - Data level, Feature Descriptor level and Decision level.

### 3.1.2. Analysis of Human Gaits

Gait recognition is the term typically used in the computer community to refer to the automatic extraction of visual cues that characterize the motion of a walking person in video [100]. Gait recognition methods can be broadly divided into two groups, model based and silhouette based approaches. Model based [101] approaches recover explicit features describing gait dynamics, such as stride dimensions and the kinematics of joint angles. The silhouette approach [102] characterizes body movement by the statistics of the patterns produced by walking. These patterns capture both the static and dynamic properties of body shape.

The human gait analysis is carried in three phases that are outlined in Fig. 3.2. as 1. Human detection and tracking, 2. Feature extraction and 3. Types of gait pose analysis. Initially, a simple and effective method is developed to extract the moving silhouettes. Features are rearranged by aligned representation in one dimensional vector followed by the correlation analysis and finally interpretation of poses as normal and interpretation of transitions are performed. The tasks of gait analysis are aligned to the hierarchy of abstractions as proposed by the unified framework.
Fig. 3.2 Proposed phases of tasks in the analysis of human gaits

1. **Phase I - Human Detection and Tracking:**
   During the task of detecting and segmenting humans found in video frames, most of the challenges arise from three principle sources – variation in illumination, a variety of poses and presence of noise. For static backgrounds, Background subtraction method is used. The moving regions are detected by taking the difference between the incoming frames and the reference background modelled frame in a pixel wise fashion. This method is simple but is sensitive to illumination changes, noises, dynamic scenes. The robustness of this technique depends mostly on the background model chosen to be as a reference frame. Generally, all background subtraction algorithms use two steps - 1. Background modeling and 2. Motion segmentation or foreground extraction. Many types of research have been reported, however, there is still need to improve and develop new effective approaches. Therefore, several different background subtraction algorithms that allow human detection/segmentation implemented with a variation from the original usage and a new algorithm is proposed. Therefore, in chapter 4, different techniques that allow human detection/segmentation have been investigated and a new algorithm Recurrent blockwise GMM background modeling on RGB color space with tracking (RBGMM) is developed and compared with the other investigated methods.

2. **Phase II: Representation and Correlation Analysis:**
   While extracting silhouettes, the shapes of the silhouettes change over time. Therefore, the method common coordinate representation is used to represent the extracted silhouettes in a single dimension vector in a common coordinate system. Later, correlation analysis of the silhouettes across the frames is performed. The goal of correlation analysis in this work
is to establish a set of observations or measurements that are generated by the same
detected human silhouette over time and analyze the overall performance. The details of
correlation analysis are described in chapter 5.

3. Phase II - Feature Extraction:
Feature descriptor is one which provides a set of optimized features extracted from the
previous stage of silhouette representation. Feature extraction is a process of describing a
set of useful measurements of the extracted silhouettes that can be used for the further
process in analysis tasks. There are several types of silhouette-based features, however, in
chapter 5 fusion of two methods, probability (Pb) of boundary based histogram of oriented
gradients (HOG) descriptors (PbHOG) [103] and the scale-invariant feature transform
(SIFT) [104], called Integrated feature descriptor is proposed to extract minimum key
points features.

4. Phase III - Types of Gait Pose Analysis:
The final step in the human gait analysis is to feed the feature extracted attributes
computed in previous stages into a classifier to infer desired knowledge about the gait
pose. Two types of tasks are experimented using neural networks in chapter 6.

Type 1: Interpretation of normal gait poses - walk, run, jump, bend using SVM classifier
and
Type 2: Interpretation of transitions - automatic detection of a transition of poses during
walking using feed forward networks of ANN.

In this step of human gait analysis, a novel approach for detecting the occurrence of
transition is used. Instead of modeling anomalous activities and attempting to recognize
them, normal walking patterns are modeled and the when there is a transition from the
normal activity, they are marked and classified as transitions if they do not fit the modeled
activity. This approach allows flexibility. Walking action’s temporal information always
varies from other gait poses. Accordingly they are calculated and are arranged as different
gait poses.
The phases of work for analysis of human gaits are aligned to the unified framework as shown in Table 3.1. The overall proposed work is implemented in two stages - Training stage and Testing stage.

1. **Training Stage:**
The training stage is used to train program about the different types of gait poses including normal as well as transitional poses. The training stage always starts before testing stage in human action analysis. It consists of several processes: reading a training video, computing the common coordinate representation of silhouettes, computing features based the extracted silhouettes, preparing feature vector and saving the feature vector for further processing tasks. All these steps are repeated for each sample of training videos found in the datasets chosen. All processes of training stage are repeated for all training samples in the dataset. At the end of training stage, the feature vectors are made available.

2. **Testing Stage:**
The testing stage is a program that interprets the normal gait poses happening in the video and also detects transitions, i.e., changes of poses from one pose to another. The testing stage consists of several processes such as reading the testing video, computing the feature vector for the testing video, followed by action classification based on training feature vectors and identifying action happening in the testing video. Until this point, these steps are common to the steps of training stage and have to be exactly in the same manner in every detail, for the comparison of the classification algorithm is performed successfully. Later the process of classification feature vector for testing video is applied using one of action classification methods. The proposed classifier algorithms are SVM and single hidden layer feed forward neural networks.

The training mode starts by reading training videos and classifies them into respective classes. The testing mode starts by reading the query video and later the normal gait poses and transitions are matched to their respective classes and thus the ends the process, after all, the video sequences are processed.
Fig. 3.3 Overview of the proposed research work
3.2. COMPARISON OF THE PROPOSED FRAMEWORK WITH EXISTING DATA FUSION MODELS

With reference to the detailed study of the literature the unified proposed framework and human gait analysis can be compared with the existing data fusion models as presented in Table 3.1:

<table>
<thead>
<tr>
<th>JDL Fusion Model</th>
<th>Dasarathy’s Classification</th>
<th>Luo, et al., Levels of abstraction</th>
<th>Proposed Tri-level Unified Framework</th>
<th>Proposed analysis of Human Gaits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0 – Source Preprocessing. Information extraction process to reduce the amount of data.</td>
<td>Data in – data out (DAI-DAO)</td>
<td>Early fusion: Features extracted from input data are combined, to produce new raw data</td>
<td>Data Level - Information extraction process for minimal representation.</td>
<td>Phase I - Extraction of moving human silhouettes.</td>
</tr>
<tr>
<td>Level 1 – Object refinement: Tasks such as spatio-temporal alignment, correlation, object classification and identification are performed. From processed data of the previous level, consistent refinement is done.</td>
<td>Data in – feature out (DAI-FEO)</td>
<td>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.</td>
<td>Feature Descriptor Level- Common representation for spatial, temporal, radiometric calibration, semantic equivalence and building of feature descriptor</td>
<td>Phase II - Common coordinate representation, Correlation analysis and Feature extraction</td>
</tr>
<tr>
<td>Level 2 – Situation assessment: Higher level inference such as likely situations between the observed events and obtained data are performed.</td>
<td>Feature in – feature out (FEI-FEO)</td>
<td>Hybrid fusion: A hybrid fusion is used to utilize the advantages of both early and late fusion strategies.</td>
<td>Decision Level- decision labeling, decisions based on semantic equivalence</td>
<td>Phase III- Gait Pose analysis: a) Interpretation of normal gait poses and b) Interpretation of transitions</td>
</tr>
<tr>
<td>Level 3 – Threat Assessment: Evaluation and prediction of logical outcome are derived.</td>
<td>Feature in – decision out (FEI-DEO)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Level 4 – Process refinement: Improves level 0 to level 3.</td>
<td>Decision in-decision out (DEI-DEO)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1 Comparison of JDL Fusion Model, Dasarathy’s Classification, Level of abstraction, Proposed Tri-level Unified Framework and analysis of Human Gaits