CHAPTER 1

INTRODUCTION

This chapter provides an introduction to the research work carried out in this thesis. The chapter begins with the background information of multimedia data and video processing, computer vision systems, data fusion and human gait analysis on which aim, problem statement and motivation of the research were developed. It concludes with the list of contributions of the researcher’s and a brief outline of the rest of the thesis.

1.1. BACKGROUND

Multimedia is a term that collectively describes a variety of media content available in different combinations of text, speech, and audio, still images, video, animation, graphics, and 3D models. The technological advances in the recent years have enabled wide access to multimedia content and many research studies were dedicated to performing automated computational tasks for a wide spectrum of applications such as surveillance, crime investigation, fashion and designing, traditional aerospace, publishing and advertising, medical applications, virtual reality applications to name a few. The volume of multimedia information has become so huge now that the improvement in various tasks of representation, analyzing, searching and retrieving process has become the need of the hour. Among all the available types of media, the video is one of the prominent forms, widely used for analyzing multimedia content and refers to the computerized understanding of the semantics of a multimedia document.
Several types of videos can be generated for capturing real-time moments that can be recorded by various recording devices. However, even the most suitable types of devices used for acquiring images and videos have to deal with two important problems - sensory gap and semantic gap. The sensory gap is the difference between the real world and its representation. “The sensory gap is the gap between the object in the world and the information in a (computational) description derived from a recording of that scene” [1]. The semantic gap is the difference between the behavior description by human vision and the computational model used by the human activity/behavior analysis systems. “The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation”[1]. Many researchers have proposed to build computational models of the human visual system to make representations as close as possible to the reality. A major development was the framework proposed by David Marr at MIT, [2] that inspired many to solve open problems relating to vision at three distinct, complementary levels of analysis namely computational level, algorithmic/representational level and physical level.

Computer vision is used to perceive electronically visual data in the form of images, text, videos. One of the main goals of researchers working in computer vision has been to enable computers to analyze and interpret images or videos. Computer vision systems include various steps such as acquiring different kinds of images and videos, processing them and extracting the information from them to make meaningful decisions. The step of processing involves several tools and techniques that integrate the content of multimedia to arrive at meaningful interpretations. The relation between multimedia content analysis and computer vision is a well-known challenge which is to be studied in this research work. A crucial challenge in multimedia is to be able to describe and compare multimedia contents in a meaningful representation to be understood by humans.

There are several applications that can be related to multimedia content analysis and computer vision. Some of the potential applications that can be used include:
• **Video surveillance:**
  The real-time media used for alerting during the time of occurrence (in progress) of any kind of offence, using tracking and recognition techniques of the human face and gait, as a sample shown in Fig.1.1.

• **Advanced human-machine interfaces:**
  Future machines to sense independently (without assistance) the surrounding environment. For example applications where detecting human or objects present and interpreting behavior or actions, as a sample shown in Fig.1.2.

• **Motion based diagnosis and identification:**
  Synthesis of human movement based on knowledge of the acquisition of the human body model and human gait along with face detection used as a new biometric feature for personal identification, as a sample shown in Fig.1.3.

![Fig. 1.1 Video Surveillance](image1)

![Fig. 1.2 Advanced human interface](image2)

![Fig. 1.3 Motion based diagnosis](image3)

The application chosen for multimedia content analysis in this research is one which automatically interprets human movement found in videos. There are various applications used to deal with the problems of understanding different kinds of human movements but the goal is to present work is to analyze automatically and transform the input video sequences into semantic interpretation. The recognition of human activities has been
studied in computer vision for quite some time but is far beyond the capabilities of human vision. In the human visual system, when a person moving is observed, the human brain recognizes that person’s action by analyzing the transition of postures adopted or interprets behavior by tracking the person’s transition of postures and noting the intent of the action. This analysis is complex for computer vision systems. Since the human body is non-rigid, deformable and articulated, a person can have a variety of postures over time. The works on human activity analysis using computer vision techniques have not provided satisfactory results yet.

To provide solutions to the challenges of human movement analysis using videos, the paradigm of data fusion is recommended. Multimedia data fusion is a way to integrate multiple media and their associated features to integrate intermediate decisions to perform an analysis task. According to B.V Dasarathy, “Combining Multimedia data fusion is a formal framework in which are expressed means and tools for alliance of data originating from different sources for the exploitation of their synergy in order to obtain information whose quality cannot be achieved otherwise” [3]. Few research contributions on data fusion techniques made use of multisensory environments and multimodal fusion with the aim of fusing and aggregating data obtained from multiple sources. It was noted that combining the information gathered from multiple modalities is a valid approach to increase accuracy [4]. However, there are different types of multimedia fusion which are described later. Multimedia fusion is useful for several multimedia analysis tasks such as detection of humans, event identification, tracking of vehicles, human motion tracking and many other applications.

1.1.1. Overview of Multimedia Data and Video Processing
A variety of media types from multiple sources is an essential feature of modern multimedia content. In general, there are three major types of multimedia data namely images, audio and video. Images are a static form of data, whereas both audio and video have a temporal component. The video is a sequence of moving images, displayed at a rate of at least 25 or 30 frames per second (fps) to provide the illusion of motion [5]. Videos are captured and displayed in the Red, Green and Blue (RGB) format. However they are often converted to an intermediate representation for efficient processing such as normalized RGB, Hue-saturation-intensity (HIS) and other formats. The exchange of digital video
between different products, devices and applications are required to conform to standards such as VGA, XGA, WXGA and so on [6]. A multimodal form of multimedia data comprises of both visual and audio features. Basically, the Multimedia content is characterized by the following properties [7]:

- **Diversity:**
  Different media are usually captured in different formats with different rates. For example, a video may be captured at a frame rate that could be different from the rate at which audio samples are obtained, or even two video sources could have different frame rates. Thus, the fusion process needs to address this asynchrony to accomplish the better task.

- **Temporality:**
  Different multimedia data types - video, audio and animation sequences have temporal requirements, which have implications on their storage, manipulation and presentation. The problems become acute when various data types from possibly distinct sources must be presented.

- **Spatiality:**
  Formats - images, graphics and video data have spatial constraints in terms of their content. Usually, individual objects in an image or a video frame have some spatial relationships between them. Relationships usually produce some constraints when searching for objects.

- **Need for Storage Space and Fast Transmission:**
  Large volumes of data also describe the multimedia information. A 5-minute sequence of the same image at 10-15 frames per second requires about 12000 Mbytes of storage. A movie of two hours would require large volumes of data storage.

- **Sensory:**
  The sensory gap between the multimedia object in the world and the information in a description derived from a recording of that scene is also to be considered.

- **Semantic:**
  The semantic gap is the lack of coincidence between the information that one can extract from visual data and the interpretation that the same data has for a user in a given situation.
1.1.2. Overview of Typical Computer Vision Systems

Computer vision deals with extracting meaningful descriptions of physical objects from images using effective methods for automating. Various tasks such as identifying a signature, locating faces or recognizing objects in a scene, restoring images are considered to be within the scope of computer vision. Humans can perform these tasks effortlessly, but developing a system to perform them could be a difficult process. For any given problem, computer vision systems are built to address the question of how to represent visual information. Usually, for a given process, the whole task must be divided into easier-to-solve stages and find the most suitable techniques in order to process the captured images. For each concrete problem, it is necessary to select specific algorithms and techniques.

Typical functions that are found in vision based processes correspond to:

- **Acquisition:**
  Initially, almost all solving of computer vision applications begin with the acquisition of images or videos using various cameras and digitization of the images and are made available for processing.

- **Preprocessing:**
  The next step is to modify the image or video frames to improve it in some way, usually with the aim of assuring that the images satisfy certain convections like noise removal, uniformity of or normalizing images for further manipulation.

- **Segmentation/Detection:**
  Following the previous step, dividing the image in a set of objects as foreground and background are performed for the purpose of segmentation or for detecting specific objects.

- **Feature extraction:**
  Feature extraction is a process used to locate the objects with interest points, lines or edges and extract descriptors of the regions of interest of the image. Some descriptors include shape descriptors, color, histograms and so on.

- **Classification:**
  Using the features extracted in the previous step, classification of objects according to the relevant applications is performed such as detection of persons, object detection and so on.

- **Interpretation:**
A semantic description of the image or video frames, in terms of the application, taking into account the results of previous stages is carried out. Each one of these steps depends highly on the tangible information we look for in sequence of images. This means that a processing technique which works fine with a concrete set of images for a certain problem probably will not perform well for another problem. Thus, for each different problem, a different solution has to be found. It is a challenge finding the best techniques for each step in the process that can give the best possible solution. Of course, not every step has to appear in a real process. Sometimes, situations may be met, in which a certain step is not needed. And it is also true, that these steps are not always so clearly separate. It may occur, for instance, that we have to apply different preprocessing techniques before segmenting or that one or more of these steps are fused into one.

1.1.3. Overview of Data Fusion

The integration of data from various sources is called data fusion. Data fusion provides a theoretical, computational and implementation framework for combining data and knowledge from different sources with the aim of maximizing the useful information obtained from multiple modalities; useful information can be extracted with proper analysis. The well-known definition of data fusion was provided by Hall and Llinas [8], “data fusion techniques combine data from multiple sensors and related information from associated databases to achieve improved accuracy and more specific inferences than could be achieved by the use of a single sensor alone”.

The fusion can also be performed using three abstraction approaches: Early fusion approach, the features extracted from input data are combined initially and represented using compression techniques and later used as input to perform further analysis task. Late fusion, involves learning of combination rules across using a certain amount of data with the ground truth as a validation set. A hybrid fusion is used to utilize the advantages of both early and late fusion strategies [9]. Apart from the frameworks, there are data fusion techniques namely, 1. Data association (Correlation), 2. State estimation (Prediction) and 3. Decision fusion [10].
1.1.4. Overview of Human Gait Analysis

Among different types of human movement, gait is one of the forms of movements. The appearance of gait in video sequences is spatio-temporal \[11\] as seen in the Fig. 1.4

![Fig. 1.4 Poses of a complete human walk action](image)

Gait recognition and analysis is used to signify how an individual can be recognized by the way of walking style and gait detection is the recognition of different types of human locomotion such as running, limping, hopping, etc. Gait detection tasks are closely related to vision methods for detection, tracking and recognition of human motion analysis in general. The task of human gait analysis can be broken into three distinct phases: the first phase comprises of moving human detection and tracking; second phase consists of optimal feature extraction for next stage and final phase consists of semantic interpretation using effective classifiers. Fig. 1.5 gives a general overview of the process.

![Fig. 1.5 Phases of human gait analysis](image)

In many applications, Gait analysis or “human walking motion analysis” aim at detecting moving humans in the video frames. Every action has a particular pattern and corresponds to a reaction to any situation encountered. One of the most challenging tasks is to handle the highly imbalanced distribution between normal and anomalous walking event. Normal actions of gait include walking, running, jumping bending and so on, that are cyclic where as an anomaly can give a cue that an action has deviated from normal action. For instance, a person who is walking in normal pace bends and falls suddenly or changes the pace of regular walking and starts running. Nevertheless, human actions are situation dependent. In a shopping mall running is anomalous where walking is normal while in a sports event
standing still is anomalous while running is normal. Hence, there is a need to have a cohesive conception to distinguish anomalous from normal action pattern.

1.2. AIM OF THE RESEARCH WORK

The aim of this research work is to

- Conduct a detailed investigation of currently available tools and techniques of data fusion and computer vision to be used on multimedia content.
- Propose a framework and assign relevant tasks to the respective levels of the framework.
- Perform representation of multimedia content (videos) of human walking motion (or gait) and
- Perform analysis of multimedia content (videos) for correlation and interpretation of human walking motion (or gait).

1.3. THE PROBLEM STATEMENT

The video is a prominent form of multimedia; this research work attempts to address following questions relating to the analysis of human gait found in video sequences.

1. How does data fusion address the problem of analyzing human walking motion?
2. How to build a representation of moving humans found in the video sequences?
3. How to perform correlation analysis and observe the impact?
4. What optimal method can be used for features extraction from the representation chosen to proceed to interpret gait poses?
5. What approaches can be used to perform automatic detection of
   i. Normal gait poses- walking, running, jumping, bending and so on?
   ii. Transitions- noticing changes in stances of human walking?
1.4. MOTIVATION

From the view point of data fusion, this research work is motivated by the observation that all living organisms have the capability to use multiple senses to learn and analyze an environment and perform a decision task instantly to recognize actions. But, the main drawback of the visual sensory of humans is the limited range of visual perceptions. It is not possible for humans to continuously monitor a situation and assess the intent of actions. Usually, it leads to compromises in the human brain or miss out important information. Whereas, automatic systems, when programmed correctly, can work 24 hours a day and 7 days a week allowing accurate detections.

From the view point of computer vision, the techniques and algorithms used to analyze multimedia content in the form of videos relating to human action recognition are of two types, namely,

1. Marker based approaches, where wearable sensors are attached to the body to measure body movements. This approach can be used only for a certain period of time and cannot be used for continuous recording. Therefore, for certain applications like video surveillance, patient monitoring applications marker based are not useful.
2. Markerless based systems are those where fixed or in-motion cameras are used to record movements of humans using techniques of computer vision. However, the algorithms and techniques of computer vision are yet to improve the performance of analyzing human motion found in videos.

In the recent years, there has been considerable interest in understanding behavior in an environment. Although there has been a vast amount of research done by human activity, many issues are still open and deserve further research. Therefore, the proposed research work is chosen to do an elaborate study on markerless based gait analysis with a promising strategy to integrate different techniques of data fusion and computer vision to enhance the performance of the tasks associated with analyzing human walking motion.
1.5. CONTRIBUTIONS OF THE RESEARCH WORK

The main contributions of this research work are:

- Firstly, the data fusion paradigm is proposed to be used for multimedia content analysis. In the traditional methods such as JDL classification and Dasarathy’s classification of data fusion, a list of features is extracted from multiple media sources and fused to obtain results. In this work, a theoretical framework known as Tri-level Unified Framework is developed based on the level of abstractions at the data level, feature descriptor level and decision levels to address the problem of human gait analysis. In the proposed unified framework, data from a single source is filtered and managed by integrating methods to improve the performance of the gait analysis. All the tasks related to human gait analysis are aligned to the levels defined by the Tri-level Unified Framework. Also, in the traditional works, it is noticed that the performance of representation, analysis and accurate interpretation lacked theories to check on optimality but the unified framework helps in contributing to the optimality of results.

- At data level, the first level of the Tri-level Unified Framework, the task is minimal representation. The original information is reduced in size. In the human gait analysis, for minimal representation we consider a moving human figure in a video sequence, to extract the silhouette and track all the silhouettes present in the frames of the video sequence. To achieve minimal representation, a simple and effective method for silhouette extraction and tracking was developed. Similar observations were made on a variety of algorithms for extracting and tracking of humans walking found in videos. The proposed silhouette extraction method is found to be effective in reducing the size of the storage.

- At feature descriptor level, the second level of the Tri-level Unified Framework, various tasks were defined such as spatial and temporal alignment, correlation analysis and building of a feature descriptor. In the human gait analysis, for the silhouettes extracted in the previous data level, a common representation in a spatial form, known as Aligned Representation is developed. Correlation analysis was performed on the
silhouettes extracted using different methods to study the impact on the overheads and processing time. For building a feature descriptor an Interest based Integrated PbHOG and the 2D-SIFT algorithm was developed to extract minimum key points which remained stable over all actions or poses.

- At Decision level, the third level of the Tri-level Unified Framework features or decisions is fused together such that they referred to the same object or phenomena. In the human gait analysis, meaningful interpretation of detecting normal gait poses is achieved by supervised classification using Support Vector Machine (SVM) and features were analyzed using Extreme Learning Machine (ELM) algorithm in single hidden layer feed forward neural network training for semantic interpretation of noticing changes of poses during walking.

1.6. OUTLINE OF THE THESIS

The remaining chapters in the thesis are organized as follows

**Chapter 2** presents a literature review on various classification categories of data fusion, techniques of data fusion – correlation and prediction, an overview of human motion analysis with respect methodologies, major tasks of human motion analysis - human detection, human motion tracking, feature extraction and pose analysis of human walking (gait analysis). It also reviews selected benchmark datasets available and concludes with remarks.

**Chapter 3** provided details of proposed Tri-level Unified Framework and proposed human gait analysis tasks aligned to the framework.

**Chapter 4** presents an investigation of modified existing methods and proposed novel method for silhouette extraction as the task of minimal representation relating to Data Level of the Tri-level Unified Framework. It also presents evaluations performed on the methods investigated.
Chapter 5 starts with common representation scheme, known as Aligned Representation for silhouettes extracted from the previous chapter, followed by correlation analysis for reduced processing time and overheads. It introduces an integrated method for extraction of features using integration of silhouette and contour based features. These tasks are associated with the Feature Descriptor level of the Tri-level Unified Framework.

Chapter 6 elaborates on the tasks related to decision level of the Unified framework. The analysis of human gait poses is categorized as

i) Detection of normal gait poses - walking, running, jumping and bending using supervised multi-class SVM classifier, and

ii) Interpretation of transitions - automatic interpretation of transitions in poses of human walking using feed forward networks of ANN.

Chapter 7 concludes the work presented in this thesis with a conclusion based on contributions and observations made during of the research work from the earlier chapters. Recommendations are made for future investigations into different tasks associated to human actions using the Tri-level Unified Framework and techniques in a broader perspective.