1
Introduction

During the last decades, a significant development in the information technology sector has improved our lives in an unprecedented way. An enormous amount of information generation has posed significant challenges in the storage and transmission of the data. In recent years, high-resolution images and videos are captured at an uncontrolled rate. For example, today, with the advent of internet technology and ease in the availability of handheld mobile phones, it has helped people to share this multimedia information to the broader audience. Also, different web applications like Facebook, Skype, WhatsApp, WeChat, Viber, Messenger allow users to interact with their loved ones through live-video streaming. Moreover, many companies are working on creating a better real-time user experience for video calls, surveillance analysis, video compression, and Spatio-temporal data compression in general. Not only this, but the storage of human genome data is also posing a challenge for efficient storage considering its strategic importance. Thus, it is the need of an hour, that one should be able to compress these data in a very efficient way for better storage and faster transmission.

An increase in the popularity of high-end tablets, handsets, and smartphones for internet usages motivated Cisco to predict the future of video traffic [Index, 2015]. The report says that video data will have more than 75% share in overall bandwidth usages by 2020. The shift towards high-definition video access will be more prevalent, and this will dramatically further increase Internet traffic in the near future. To fulfill such demands, efficient video compression methods will be highly required. [Marpe *et al.*, 2006] explained the framework for H.264, which is currently one of the most popular video compression standard used for recording, storage, and transmission of video contents. However, High-Efficiency Video Coding (HEVC) [Sullivan *et al.*, 2012] is the potential successor of H.264 as it provides about 50% more compression performance for a similar visual quality at the expense of about 300% computational overheads. In these video compression standards, the motion estimation process is one of the most crucial and time-consuming components. However, much research has been done to improve computational complexity at the expense of the loss in the performance of the matching of blocks. A computationally efficient motion search algorithms are highly desired for real-time video analysis and compression. For example, surveillance video analysis demands highly efficient motion search algorithms for faster and higher compression. Surveillance video is one of the industry's most widely used technologies and was expected to generate nearly 16 billion dollars of global revenue in 2016. In recent times, human skeleton action movements are also recorded for security purposes. Hence, this massive amount of skeleton information would pose a challenge for efficient storage.

In the present doctoral research work, efforts are made to develop fast and efficient motion estimation schemes that achieve higher compression performance as well as higher visual quality at a lower computational complexity. Furthermore, a novel approach for efficient motion estimation scheme is explored for surveillance videos, which contain significant static regions. Moreover, the increasing importance of skeleton information in surveillance big data features analysis demands substantial storage space. The efforts are also made for the development of an effective and efficient solution for the storage of these skeleton information.

The following topics are covered in this chapter. A brief introduction to data compression is provided in the beginning, followed by special emphasis on video compression. The background to the research problem is explained with the primary focus on the need, motivation, and relevance of the

Thesis work, followed by the Thesis problem statements and research objective to address some of the gaps in the existing research. Next, the main Thesis contributions are explained in brief, followed by a chapter-wise Thesis outline.

1.1 BRIEF INTRODUCTION TO DATA COMPRESSION

So what is compression of data, and why do we need it? Most of us have heard of JPEG and MPEG, standards of image, video, and audio representation. Such standards use data compression algorithms to reduce the number of bits required to represent an image or a sequence of videos or audio. Briefly, compression of data is the art or science of compact representation of information. By identifying and using structures that exist in the data, we create these compact representations. Data compression is a reduction in the total number of bits needed to represent data. In doing so, data compression can dramatically save storage space, speed up data transmission, and it would, in turn, decrease costs for data storage hardware and network bandwidth. In general, the data compression methods can be broadly categorized into two types: (1) lossless compression, and (2) lossy compression. Many types of data contain specific statistical redundancies. These redundancies can be effectively exploited for the lossless data compression. The core principle of the lossless compression is to minimize the storage capacity required to represent the original input data without losing any information. Hence, lossless compression is also termed as a reversible process. On the other hand, lossy data compression permanently removes some inherent redundancies that are unimportant or imperceptible. To this end, the entire focus is on achieving a better trade-off between preserving information and reducing storage capacity requirements. In practice, only a small loss in information could provide a significant reduction in storage space requirements. It is to bring to readers kind notice that, this low loss in information is permanent. Hence, lossy compression is also termed as an irreversible process. Both the lossless and lossy compression schemes are further elaborated in Chapter 2. This compression is used only in the applications where perfect reconstruction of the original data is not of critical importance. In such cases, our target is to find any possible repetitive patterns which can help to identify redundant information in the given data. These data redundancies could lead to better compression.

1.1.1 Data Redundancy

In literature, mostly, the data is a sequence of numbers that represent samples of a continuous variable. The continuous variable might belong to various combinations of popular domains such as space, time, or frequency. In general, data compression is achieved by removing the redundancies inherent in the input data. The redundancies can be broadly classified into three categories: (1) statistical redundancy, (2) coding redundancy, and (3) psycho-visual redundancy.

Statistical redundancy can be further classified into three categories based on the domains listed above, such as spatial redundancy, temporal redundancy, and spectral redundancy. The spatial redundancy exists due to the presence of a strong correlation between adjacent samples. On the other hand, temporal redundancy exists due to the strong correlation between samples captured within a small time interval. Moreover, spectral redundancy exists due to the considerable amount of correlation between samples obtained within a short wavelength interval. On the other hand, the coding redundancy is associated with the representation of the information in an effective way. Finally, psycho-visual redundancy exists due to the inability of human perception to slight variations in the original data. The existence of these redundancies in the data largely depends on the content of data. The data can be classified into different types based on the content.

1.1.2 Types of Data

Currently, there are various types of data used in the literature to represent information. For example, various types of data include, text, audio, images, videos, medical data, point cloud data, skeleton data, genomics data, astronomical data, hyper-spectral data, radar data, light-field data, satellite data, military data, weather data, seismic data, and financial data, among others. Although information theory concepts are equally valid on all types of data, a data compression algorithm designed for one kind of data might not work as suitable for different kinds of data. In this Thesis, our primary focus is on compression of some of these Spatio-temporal data. Furthermore, different Spatio-temporal data, like video data and human skeleton-point data, are investigated in this Thesis work.

An image is a 2-dimensional (2D) intensity pattern. A k-bit resolution image is represented by an intensity value $I = \{0, V_{max}\}$, where $V_{max} = 2^k - 1$. In general, the intensity value in the grayscale image is represented by an 8-bit resolution. However, color images contain three spectral components representing red, blue, and green color images, respectively. Hence, the color image requires 24-bits in total to represent a single pixel intensity value. Image compression is an extensively studied area, where discrete-cosine-transform (DCT)-based lossy compression technique helps to considerably compress an image without any significant loss in the information. [Wallace, 1992] presented a DCT-based image compression algorithm as a part of the Joint Photographic Experts Group (JPEG). Till today, JPEG is a widely adopted image compression algorithm since it was introduced in 1992. JPEG typically achieves about 10:1 compression with little perceptible loss in image quality.

A video is a 2D intensity pattern that changes with time. In a video, images are displayed at a constant speed to create an illusion of motion pictures. The sampling period T in the time domain is judiciously chosen to create this illusion without dramatically increasing the storage and transmission costs. The sampling period T is associated with the frame rate of $1/T$, where frame rate refers to the number of frames per second. To this end, various frame rates from 30 Hz to 120 Hz are available for video display devices. Although higher frame rates are studied in the literature, it is out of the scope of our discussion. For illustration purposes, consider a colored video with a frame size as 1920×1080 $(1080p)$ and a frame rate as 120 Hz. This video would need $3 \times (1920 \times 1080) \times 120 \times 8 = 5971968000$ bits per second (bps) for storage and transmission. The above data rate requirement is huge and demands compression for efficient storage and real-time transmission.

1.2 VIDEO COMPRESSION

Undoubtedly, the video signal contains significant redundancies in both space and time. The temporal redundancies are mainly due to the presence of a strong correlation between successive frames. Not only this, spatial redundancies exist due to the strong correlation between adjacent pixels within the frame. Video compression works on the principle of removing these redundancies using different coding techniques. Traditionally video compression techniques may exploit only temporal or both spatial and temporal correlations to achieve the desired compression ratio. The data compression ratio (CR) is defined as the ratio between the uncompressed file size and compressed file size.

In literature, the spatial redundancy can be exploited based on the strong correlations between adjacent pixels within the frame using intra-frame coding techniques. Intra-frame coding techniques are similar to JPEG. Moreover, the temporal redundancy can be exploited based on the strong correlations between successive frames using inter-frame coding techniques. A most straightforward inter-frame coding method is to use differential coding for consecutive frames, where the difference between the current frame and reference frame is encoded to reconstruct the current frame at the receiver. It is a well-known fact that using only intra-frame coding techniques can provide modest CR. However, the exploitation of both spatial and temporal redundancies using intra-frame and inter-frame coding can provide significant improvement in CR for efficient storage and compression. This hybrid approach is the core principle used in state-of-the-art video compression standards such as MPEG-4 [Wang *et al.*, 2016], H.264 [Marpe *et al.*, 2006], and HEVC [Sullivan *et al.*, 2012]. In these standards, the motion estimation plays a vital role in inter-frame coding. The video compression and motion estimation methods are further elaborated in Chapter 2.

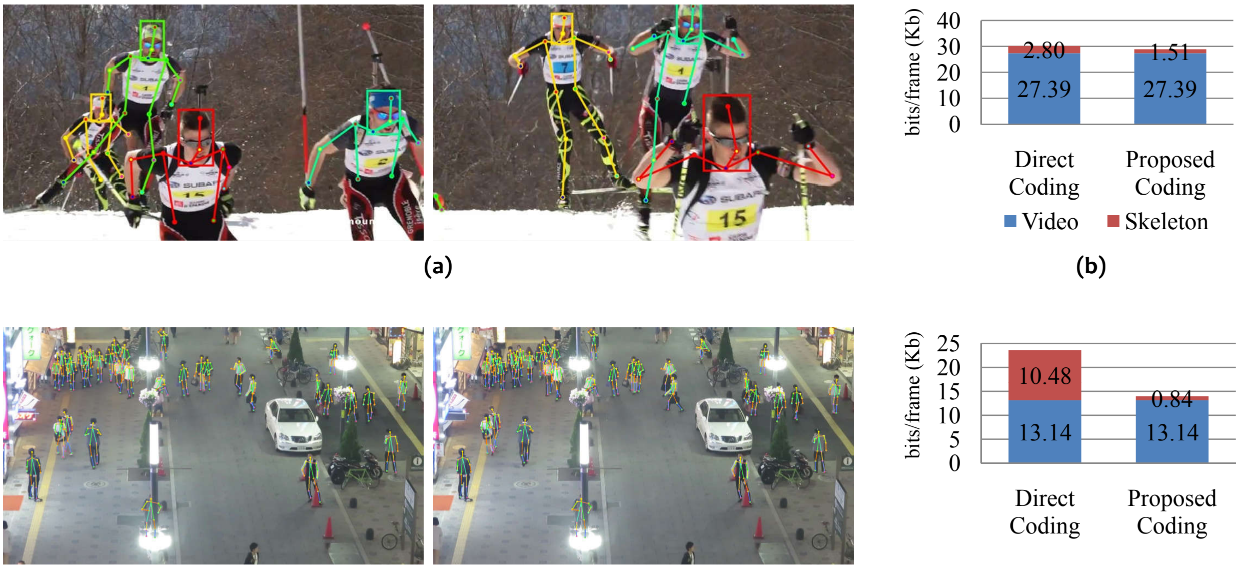
1.3 BACKGROUND AND MOTIVATION

Motion estimation, which used to reduce temporal redundancies through successive frame matching, plays a vital role in different video analysis applications. The applications include moving object detection, traffic movement tracking, human-computer interaction (HCI), hand posture analysis, cinematography, robotic heart surgery, studying plant root growth, temporal interpolation, Spatio-temporal filtering, and video compression methods, etc. 3-D relative motion of space-lander can also be estimated accurately by using onboard navigation camera. Moreover, motion estimation is also finding its use in the computation of motion fields for biophysical analysis of cellular processes. In general, motion estimation and optical flow terms are used interchangeably [Bruhn and Weickert, 2005]. Optical flow methods are used for ultimate motion estimation, which provides the highest accuracy but compromises efficiency. [Seyid *et al.*, 2016] implemented the optical flow algorithms in hardware for real-time motion estimation. In motion estimation, it is desirable to obtain true motion results in an efficient way for real-time analysis. For this, various GPUs, FPGAs, and VLSI architectures are employed for motion estimation [Loukil *et al.*, 2004; Botella *et al.*, 2012]. The well-known application of real-time motion analysis lies in the surveillance videos.

Today, surveillance cameras play an important role in home-care, public safety and security, traffic management, and business enhancement. The increase in a significant market for surveillance videos has brought a great challenge for the storage and maintenance of thousands of Terabytes data that would be produced per minute. Thus for long-time archival and real-time monitoring of surveillance videos, there is a great need for fast and efficient coding methods. On the basis of these challenges and requirements, this work focuses on the computationally efficient surveillance video coding methods.

In the traditional video coding framework, the motion estimation component plays a crucial role in reducing the temporal redundancy. Hence, for better motion estimation, computationally expensive methods are employed. For example, in the Full Search (FS) motion estimation method, the best matching block corresponding to the block in the current frame is matched with the all candidate blocks present in the entire search window in the reference frame [Lin and Tai, 1997]. Although this process provided the best matching performance, it has severe computational disadvantages. Thus to increase the speed of the block matching process, several fast and efficient search algorithms are proposed in the literature. Although these search algorithms are fast, they provide suboptimal matching performance. Hence, there is a need to develop highly efficient search algorithms that not only reduce computational complexity but also do not compromise in matching performance. This problem is more prevalent in surveillance videos, which require real-time processing. Surveillance videos mainly contain static regions, and the remaining active region typically consists of human action movements.

The skeleton information plays an important role in various human action recognition, event detection, surveillance feature analysis, and health-care monitoring applications. This information is very critical for improved performance and accuracy. For example, the skeleton-based video modelling methods have intensively used skeleton information for various action recognition tasks [Jiang *et al.*, 2015; Li *et al.*, 2018b; Fang *et al.*, 2017a,b; Wang and Wang, 2017; Li *et al.*, 2018a; Ke *et al.*, 2017; Tang *et al.*, 2018; Zhang *et al.*, 2019]. In order to save computation complexity at the receiver side, many emerging edge-computing applications tend to extract skeleton data at the sensor or transmitter side and directly transmit the extracted skeleton data together with the original video data to the receiver side. Thus, it becomes a new but non-trivial problem to encode skeleton data efficiently. The higher accuracy demands original and reliable skeleton information. With the significant increase in the skeleton data,



(a) (b)

(c) (d)

Figure 1.1 : (a) Illustration of typical Posetrack skeleton sequence, (b) bit-rate for video and skeleton content in Posetrack, (c) typical surveillance skeleton sequence, (d) bit-rate for video and skeleton content in surveillance. The skeleton sequences are compressed by the traditional fixed-length direct coding approach (left bins) and our proposed skeleton coding approach (right bins). (Best viewed in color)

the storage of skeletons in the original form imposes space constraints. It is desired to perform lossless compression of skeleton sequences to preserve the naturalness.

Basically, skeleton sequences are represented by a sequence of skeletons, as shown in Figure 1.1 (a) and 1.1 (c). Each skeleton typically consists of 15 body joints. In this study, we consider a total of fifteen ordered body joints, namely: neck, nose, head-top, left-shoulder, left-elbow, left-wrist, right-shoulder, right-elbow, right-wrist, left-hip, left-knee, left-ankle, right-hip, right-knee, and right-ankle. Nowadays, with the significant advancement and development in image and video compression techniques such as H.264 and HEVC, the image and video data size can be greatly reduced [Sullivan *et al.*, 2012; Shen *et al.*, 2015; Jamali and Coulombe, 2019]. Comparatively, the compression problem of semantic data in videos, such as skeleton sequence data studied in this Thesis, is mostly neglected. In practice, since many videos include a large number of people and have rich skeleton information, if we only compress video data while not compressing this rich skeleton information, it will occupy a non-negligible large portion in the final encoded bit-stream. For example, in Figure 1.1, we have two skeleton sequences, each containing 4 and 35 people respectively. If we only compress the video data while not compressing these skeleton sequences, the skeleton data will take about 10% and 50% in the final bit-stream respectively (left bins in Figure 1.1 (b) and 1.1 (d)). However, if we compress the skeleton information in a lossless manner by our approach, the bit-requirement can be successfully reduced by about 70% (right bins in Figure 1.1 (b) and 1.1 (d)). Therefore, it is of utmost importance to develop novel encoding methods to handle the huge amount of skeleton data. This background forms the basis for the Thesis work.

1.4 RESEARCH OBJECTIVES OF THE THESIS

The facts mentioned above indicate the importance of efficient motion search algorithms and skeleton information in surveillance videos. The primary concern with the previous motion search techniques lies in the fact that the fast algorithms are sub-optimal due to the notable compromise in matching accuracy in the pursuit of reduction in motion search complexity. This problem is of utmost importance for motion search algorithms used in real-time surveillance video analysis. Moreover, considering the strategic importance of the human skeleton information in the surveillance video scenario, the storage space requirements are naturally high.

This Thesis aims to address the problems mentioned above to not only fill the research gap in the existing works but also to develop the novel mechanisms for efficient motion search and compression of Spatio-temporal skeleton sequences. The main research objectives of the Thesis are listed below.

- **To develop efficient and effective motion search algorithms**: The fact that motion estimation is a time-consuming process has attracted many researchers to improve the computational complexity but at the expense of matching accuracy. To this end, our objective is to study the effect of various block matching techniques, search patterns, search path, search region, and pixel sub-sampling to achieve a trade-off between computational complexity and matching accuracy.

- **To develop computationally efficient motion search algorithms to exploit special characteristics of the surveillance videos**: The key characteristic of surveillance video lies in the typical high proportion of the static region as compared to the active region containing moving objects. Our objective is to develop effective motion search mechanisms for each region to achieve a trade-off between computational complexity and matching accuracy.

- **To develop effective and efficient storage mechanism for Spatio-temporal skeleton sequences**: In practice, since many surveillance videos contain a large number of humans and have rich skeleton information, which demands special attention and novel storage solutions. Our objective is to efficiently compress skeleton data while maintaining exactly the same skeleton quality as the original ones.

1.5 CONTRIBUTIONS OF THE THESIS

The objectives of the Thesis mentioned in the earlier discussion are achieved through the following contributions addressing each task:

- **Efficient direction-oriented motion search algorithm for block motion estimation**: The motion estimation is the most time-consuming component in the video compression methods. Although fast motion search algorithms are presented in the literature, they suffer from sub-optimal block matching performance. For this issue, a novel block matching algorithm is presented in the Thesis. In this regard, novel direction-oriented search patterns are developed to not only exploit directional motion characteristics of the videos, but they also helped in faster search convergence. The computational cost can be further reduced by our novel adaptive threshold-based pixel sub-sampled structures.

- **Efficient motion search for surveillance videos**: Reduction in computational cost in the motion estimation process for the real-time surveillance video analysis is a pressing task. The characteristics of surveillance video, such as a significant proportion of static regions, can address the job mentioned above. A novel three-level block classification mechanism is proposed for effective motion search strategies. The highly efficient no motion search strategy is employed for static regions. On the other hand, novel search patterns are proposed to address motion search strategies at active and boundary regions. Our novel region-based pixel sub-sampled structures can further reduce the computational cost.

- **Adaptive compression scheme for Spatio-temporal skeleton sequences**: The importance of skeleton information in surveillance video analysis demands significant storage space. The Spatio-temporal characteristics and dependencies between the skeleton sequences can effectively improve the compression performance. To this end, a novel skeleton information prediction scheme is presented in this Thesis. The skeleton prediction modes are created specifically to exploit various spatial and temporal correlations among skeleton sequences. Not only this, but we also focused on addressing remaining redundancies, if any, by special coding schemes. By doing so, we have not only reduced the storage space requirement but also maintained the same skeleton quality as the original ones.

List of Publications

International Journals:

1. T. S. Shinde, and A. K. Tiwari, "Efficient Direction-Oriented Search Algorithm for Block Motion Estimation", IET Image Processing, 12.9 (2018): 1557-1566.

2. T.S. Shinde, A. K. Tiwari, W. Lin, and L. Shen. "Background foreground boundary aware efficient motion search for surveillance videos." Signal Processing: Image Communication (2020): 115775.

3. W. Lin, T. S. Shinde, W. Dai, M. Liu, X. He, A. K. Tiwari, and H. Xiong. "Adaptive Lossless Compression of Skeleton Sequences." Signal Processing: Image Communication (2019): 115659.

International Conferences:

1. T. S. Shinde, A. K. Tiwari, and W. Lin. "Low-complexity Adaptive Switched Prediction based Lossless Compression of Time-lapse Hyperspectral Image Data." In 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), IEEE, 2019.

2. T. S. Shinde, and A. K. Tiwari. "Pruning SIFT & SURF for Efficient Clustering of Near-duplicate Images." In 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3132-3136, IEEE, 2019.

1.6 THESIS OUTLINE

The remainder of this Thesis is organized as follows:

Chapter 2 surveys the related work in the field of video compression, motion estimation, motion search in surveillance videos, and skeleton sequence coding.
Chapter 3 describes an efficient direction-oriented motion search algorithm.
Chapter 4 presents an efficient motion search algorithm for surveillance videos.
Chapter 5 illustrates an adaptive algorithm for lossless compression of skeleton sequences.
Chapter 6 concludes the Thesis and discusses important areas for future work.

...