Summer 7-31-2013

Cross-layer Optimized Wireless Video Surveillance

Yun Ye
University of Nebraska-Lincoln, yye@huskers.unl.edu

CROSS-LAYER OPTIMIZED WIRELESS VIDEO SURVEILLANCE

by

Yun Ye

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Doctor of Philosophy

Major: Engineering

Under the Supervision of Professor Song Ci

Lincoln, Nebraska

July, 2013

CROSS-LAYER OPTIMIZED WIRELESS VIDEO SURVEILLANCE

Yun Ye, Ph.D.

University of Nebraska, 2013

Adviser: Song Ci

A wireless video surveillance system contains three major components, the video capture and preprocessing, the video compression and transmission over wireless sensor networks (WSNs), and the video analysis at the receiving end. The coordination of different components is important for improving the end-to-end video quality, especially under the communication resource constraint. Cross-layer control proves to be an efficient measure for optimal system configuration. In this dissertation, we address the problem of implementing cross-layer optimization in the wireless video surveillance system.

The thesis work is based on three research projects. In the first project, a single PTU (pan-tilt-unit) camera is used for video object tracking. The problem studied is how to improve the quality of the received video by jointly considering the coding and transmission process. The cross-layer controller determines the optimal coding and transmission parameters, according to the dynamic channel condition and the transmission delay. Multiple error concealment strategies are developed utilizing the special property of the PTU camera motion.

In the second project, the binocular PTU camera is adopted for video object tracking. The presented work studied the fast disparity estimation algorithm and the 3D video

transcoding over the WSN for real-time applications. The disparity/depth information is estimated in a coarse-to-fine manner using both local and global methods. The transcoding is coordinated by the cross-layer controller based on the channel condition and the data rate constraint, in order to achieve the best view synthesis quality.

The third project is applied for multi-camera motion capture in remote healthcare monitoring. The challenge is the resource allocation for multiple video sequences. The presented cross-layer design incorporates the delay sensitive, content-aware video coding and transmission, and the adaptive video coding and transmission to ensure the optimal and balanced quality for the multi-view videos.

In these projects, interdisciplinary study is conducted to synergize the surveillance system under the cross-layer optimization framework. Experimental results demonstrate the efficiency of the proposed schemes. The challenges of cross-layer design in existing wireless video surveillance systems are also analyzed to enlighten the future work.

Acknowledgments

Like sand through the hourglass, so are the days of my PhD study in Omaha. These days build the most cherished memory in my life, not only because of the knowledge and skills I gain, but also because of the life I experience.

The opportunity of this experience is largely attributed to my advisor. Hadn't I met him in Shanghai, I would not have embarked on this wonderful journey. Besides the academic guidance, his life philosophy is always an inspiration for us, especially when we are having difficult times dealing with the career or family issues.

I would like to thank the faculty members in our department. Without their care and advice, I could not have gone through these years smoothly in a foreign country. The help from my colleagues and my friends is also the reason I am still sound and healthy even I am accomplishing the degree.

I owe my thanks to my collaborators in Chinese Academy of Sciences, and in Northwestern University, for their valuable insights and potent supports.

My gratitude goes to my family, the one place I can resort to no matter what.

To all the people who help me through this journey.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Research Background

Video surveillance over wireless sensor networks (WSNs) has been widely adopted in various cyber-physical systems including traffic analysis, healthcare, public safety, wildlife tracking and environment/weather monitoring. The unwired node connection facility in WSNs comes with some typical problems in data transmission. Among them are line-of-sight obstruction, signal attenuation and interference, data security, and channel bandwidth or power constraint. A vast amount of research work has been presented to tackle these problems, and many have been successfully applied in practice and have become industrial standards. However, for video surveillance applications, especially those with real-time demands, the processing and transmission process at each wireless node with a large amount of video data is still challenging.

1.1.1 Video Surveillance System

The development of the video surveillance system has experienced three generations – the analog video surveillance system, the digital video surveillance system, and the network video surveillance system [1]. The analog video surveillance system adopts analog video capture devices known as the Closed Circuit Television (CCTV). The video signal is transmitted via dedicated wired cables to the monitors, as shown in Figure 1.1. The advantages of this kind of system are the signal safety and cost efficiency, making it still popular in many small-scale surveillance areas like a housing estate and a factory district. In the analog system, the video signal is directly captured, transmitted and stored

without A/D conversion. The displayed video usually maintains high quality with little delay. However, when the transmission distance is getting longer, the signal is subject to deterioration due to attenuation, delay and interference. Applying amplifier helps to elongate the distance but also brings in more noises. Moreover, the system lacks the flexibility of wiring and camera rearranging. The storage of large amount of video data and the magnetic storage device's vulnerability to deformation are also among the limitations of the analog system. Therefore, its application is confined to small surveillance areas.

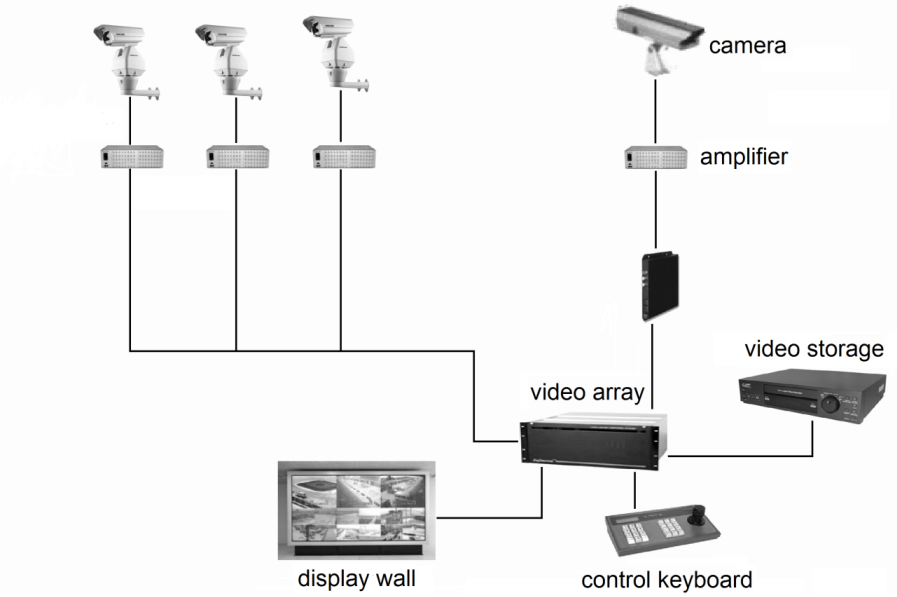


Figure 1.1 : An analog video surveillance system [1].

The digital video surveillance system is widely adopted after the commercialization of digital video recorder (DVR) in 1990s. DVR converts analog signals into digital signals and records the video on hard disk with much larger storage capacity than the magnetic cassette tape. The DVR incorporates the functions of A/D conversion, video

codec, video storage, network transmission, and remote control to support some intelligent services including virtual array switch, networking, and software development for image processing such as encryption, indexing and backup. The digital video surveillance system is deployed based on the DVR and network infrastructure, as displayed in Figure 1.2. Compared to the analog system, it can ensure the video quality for long-distance transmission utilizing the modern digital signal communication technologies. As the computer processor unit's fast growing computing ability enables various video analysis tasks to be implemented in the system for more advanced surveillance applications, e.g. data indexing and event detection, the digital system is rapidly developed, especially in large-scale, intelligent video surveillance applications.

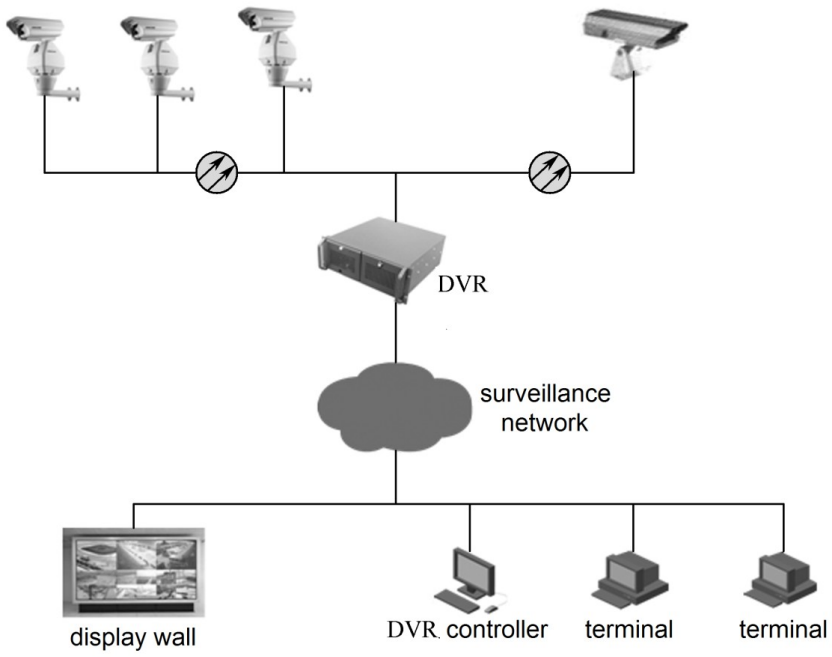


Figure 1.2 : A digital video surveillance system [1].

The popularity of broadband IP internet creates a new era for video surveillance, known as the network video surveillance, or IP video surveillance. This kind of system relies on the technical advances of multidisciplinary fields including image/video codec, microchip, pattern recognition, software control, and mostly the digital communication over the high speed internet. Depending on the networking and intelligent video analysis technologies, the surveillance system is capable of providing more complex services. For example, the authorized user can access the video via the browser at any client node in the network, including the mobile wireless network, and the alarm signal can be automatically generated based on the event detection results. Compared to the conventional surveillance system, the network video surveillance has more flexibility for data access, content display, video distribution, network extension, and protocol adaptability for video streaming. The properties of these three kinds of surveillance systems are summarized in Table 1.1.

Table 1.1 Comparison of three video surveillance systems

	Analog Video Surveillance System	Digital Video Surveillance System	Network Video Surveillance System
Coverage	small	large	large
Networking	dedicated	compatible with internet	compatible with internet
Management	complex, manual	complex, manual	simple, remote control
Extensibility	limited	limited	multi-level
Video Access	surveillance center	designated PC or surveillance center	anywhere, anytime via network
System Scale	small	small	large
Remote Control	no	poor	good
Data Storage	magnetic cassette tape	disk array or optic disk	disk array
Safety	low	normal	high

According to the characteristics of the dedicated service, the existing research and development work on video surveillance can be classified into three categories – the intelligent video surveillance, the high definition video surveillance, and the wireless video surveillance. The intelligent video surveillance system features the advanced video analysis technologies for object tracking or event detection, usually with a set of distributed, cooperative network cameras. It has been successfully applied for some public services such as the intelligent transportation as demonstrated in Figure 1.3 [1].

Figure 1.3 : Intelligent video surveillance at the traffic station [1].

The high definition video surveillance sees its market in more delicate situations when the high definition images are required to perform the pattern recognition tasks. For instance, the accuracy of the face recognition result is significantly improved with the high definition images, and hence the system well accommodates the safety purpose such as the surveillance at the bank, station and other public areas.

The wireless video surveillance is getting popular along with the development of WSN. The properties of WSN enable the source nodes and the intermediate nodes to be placed at some prohibitive locations, like the wildlife habitat, and the contagious or biochemical district. It is also possible to equip each node with certain mobility. These advantages are incomparable by the wired system when it comes to those special surveillance scenarios. However, the wireless video surveillance inevitably inherits the technical challenges for the WSN. A more prominent problem is how to coordinate different components in the system based on the WSN framework. We will discuss this problem in this thesis work and present the solutions with experimental results from several practical projects.

1.1.2 Wireless Video Surveillance

In current state-of-the-art wireless video surveillance systems, each source node is usually equipped with one or more cameras, a microprocessor and/or the storage unit, a transceiver, and the power supply. The basic functions of each node include video capture, video compression and data transmission. The process of video analysis for different surveillance purpose is performed either by the sender or by the receiver, depending on their computational capability. The remote control unit at the receiver's end can also provide some useful information feedback to the sender in order to better serve the surveillance purpose. The major functional modules of a video surveillance system are illustrated in Figure 1.4.

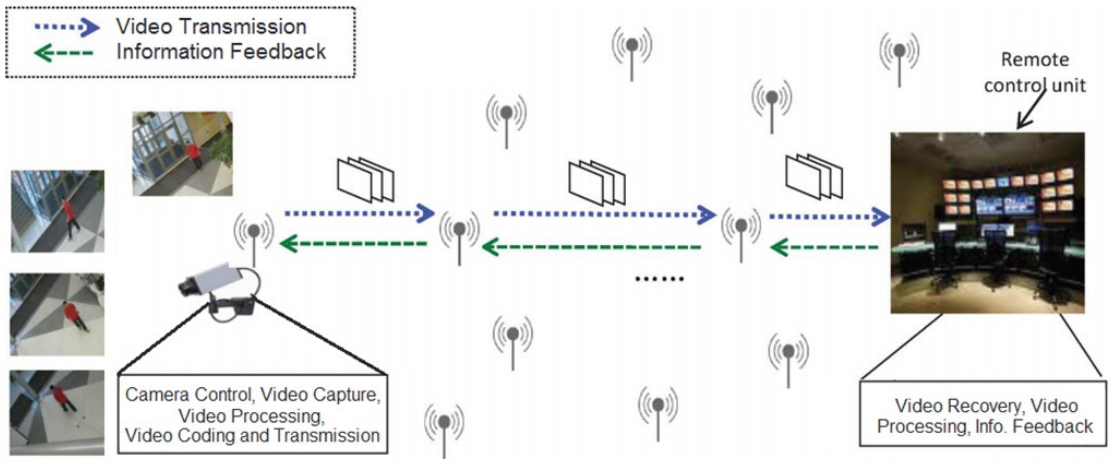


Figure 1.4 : A wireless video surveillance system.

In the U.S.A., the Federal Communication Commission (FCC) is responsible for regulating radio spectrum usage [2]. The most commonly used license-exempt frequency bands in current wireless surveillance systems include 900MHz, 2.4GHz, and 5.8GHz. The 4.9GHz frequency band is reserved for Intelligent Transportation Systems (ITS) for public safety and other municipal services [3]. The specific communication parameters are defined in several groups of standards including IEEE 802.11/WiFi, IEEE 802.16/WiMax, IEEE 802.15.4/ZigBee, etc. The existing WSN technologies are utilized in all kinds of wireless video surveillance applications. One popular application is traffic analysis in intelligent transportation. For example, the traffic signal system deployed by the transportation department in the city of Irving, Texas (Irving, 2004) [4] implemented seventy pan-tilt-zoom (PTZ) CCTV cameras to cover about two hundred intersections. One smart camera capable of video codec and video over IP function was installed at each traffic site together with a radio/antenna unit. The on-site signal is transmitted to the

base stations ringed in a 100 Mbps wireless backbone operating at the licensed frequencies of 18-23 GHz.

The traffic monitoring system developed at the University of Minnesota (UMN, 2005) [5] attached each node with one or more CCTV cameras, a computer for video recording and compression, and an Ethernet radio subscriber unit. Multiple SIF video feeds were recorded and compressed simultaneously at the traffic site, and were transmitted to the base station for real-time observation.

The traffic surveillance system tested at the University of North Texas (UNT, 2011) [6] installed at each of the three campus sites an Axis 213PTZ camera and a radio device. The traffic video was transmitted to the control center through a daisy chain network. The control center was able to remotely adjust the PTZ camera position and the focal length, and to estimate the vehicle speed on a roadway parallel to the image plane.

Video surveillance in other wireless communication applications is also intensively studied. To monitor the lightning stricken forest fire, a remote weather monitoring system (FireWxNet, 2006) initially developed for the fire fighting community in the Bitterroot National Forest in Idaho was introduced in [7]. The system was composed of three sensor networks consisting of a total of thirteen nodes, five wireless access points, and two PTZ web cameras (Sony SNC-RZ30N and Panasonic KX-HCM280). The battery and solar powered webcams had an infra-red night vision feature and could deliver video to designated web servers via a Trango radio unit.

The smart camera network system (SCNS, 2011) used in a railway station was described in [8]. Deployed in a mesh network, each sensor node was able to communicate with others, so that the targeted object was continuously tracked by cameras installed at

different locations of the station. The vision analytic PC associated with each camera processed the captured QVGA video at a speed up to ten frames per second, and transmitted the target information to other nodes. A sensor node equipped with a PTZ camera performed automatic object tracking upon received target information.

An indoor surveillance system tested in a multi-floor department building at the University of Massachusetts-Lowell was introduced in [9]. The moving target was assumed to carry a battery-powered TMote TelosB wireless sensor to signal its information to a nearby wireless router. The router passed its location information to the server and received back the camera control command to enable video capture, compression and transmission.

For surveillance in a wide social area like metropolis, the sensor deployment is more complex. An example is the multi-sensor distributed system developed at Kingston University, named proactive integrated systems for security management by technological institutional and communication assistance (PRISMATICA, 2003) [10]. Both wired and wireless video and audio subsystems were integrated in the centralized network structure. The data processing module at the operation center supported multiple real-time intelligent services, such as overshadowing and congestion detection upon received video.

Another group of research work focused on the power efficiency problem at the sensor node. In the video WSN (Panoptes, 2003) described in [11], each node was built on an Intel StrongARM based Compaq IPAQ PDA platform with an 802.11 card and a Logitech webcam. A central node received data from other client nodes and performed video aggregation to detect unusual events. The energy saving strategy employed by the

client node included data filtering, buffering and adaptively discarding. The power consumption was around 5 Watts when the system fully functioned.

A hybrid-resolution smart camera mote (MeshEye, 2007) was designed to perform stereo vision at the sensor node with low energy cost [12]. A fully integrated Printed Circuit Board (PCB) was built consisting of an Atmel AT91SAM7S family microcontroller with an ARM7TDMI ARM Thumb processor, two kilopixel cameras, a VGA resolution CMOS camera, and a Texas Instruments CC2420 RF transceiver. The location of the targeted object was first estimated from the image data by the two low resolution cameras. Then the high resolution camera marked the position in its image plane and transmitted only the video data inside the target region.

The multiview visual target surveillance system developed at Tsinghua University (Tsinghua, 2009) [13] was capable of target localization through collaboration among sensor nodes. Each sensor node carried an image sensor controlled by a neighboring pyroelectric-infrared sensor. The computing module of each sensor node was constructed on an ARM9 (200 MHz) single-board computer with a normal computing power of 122 mW. It performed the target classification and tracking task, and shared with other nodes the down sampled color data and other information. Different network topologies and collaboration strategies were instrumented to realize automatic, continuous object tracking.

Some technical parameters of these effectively deployed systems are listed in Table 1.2.

Table 1.2 Wireless video surveillance systems

Surveillance System	Surveillance Environment	MAC Protocol	Carrier Frequency	Maximum Throughput	Network Topology	Camera Number	Camera Control	Video Delivery
Panoptes, 2003 [11]	indoor	802.11	2.4 GHz	12 Mbps	star	multiple	fixed	unicast
PRISMATICA, 2003 [10]	indoor/ outdoor	802.11	2.44 GHz	N. A.	star	multiple	fixed	unicast
Irving, 2004 [4]	highway	802.11a 802.16	5.8, 18, 24/23 GHz	20 - 60 Mbps for on-site, 100 Mbps for backbone	star, ring	multiple	PTZ	unicast
UMN, 2005 [5]	highway	802.16	5.4 GHz	3/30 Mbps	star	multiple	fixed	simulcast
FireWxNet, 2006 [7]	wildland	802.11	924Mhz	10 Mbps	tree	multiple	PTZ	unicast
MeshEye, 2007 [12]	indoor/ outdoor	802.15.4	2.4 GHz	11 Kbps	point-to-point	multiple	fixed	simulcast
Tsinghua, 2009 [13]	indoor	802.15.4	900 MHz	19.2 Kbps	star, tree, mesh	multiple	fixed	unicast
UML, 2010 [9]	indoor	802.11b/ g	2.4 GHz	54 Mbps	tree	multiple	fixed	unicast
SCNS, 2011 [8]	railway	802.11g 802.11j	2.4, 5 GHz	20 Mbps	mesh	multiple	fixed, PTZ	unicast
UNT, 2011 [6]	campus	802.11 a/n	5.4, 5.8 GHz	20 Mbps	chain	multiple	PTZ	unicast

In these systems, the wireless communication technologies dedicated for improving the video quality are the key to a successful application. Besides the traditional technical characteristics, the WSN technologies applied for surveillance purpose need to consider the optimal configuration based on the practical system architecture. This new feature requires interdisciplinary study, and has drawn in myriad research endeavor, as the application is becoming increasingly popular.

1.2 Research Motivation

In a wireless video communication system, the limited channel resource is managed through configuring the options at different layers in the network architecture, for example, the coding and error control at the application layer, the congestion control and reliability protocol at the transport layer, the routing at the network layer, the contention

scheme at the MAC (medium access control) layer, and the modulation and coding scheme (MCS) at the physical layer [14]. To jointly implement the configuration procedure, the cross-layer control methodology is developed to optimize the system-level resource allocation [15]. Given the channel state information (CSI), the controller is able to coordinate decision making at different layers in order to maximize the visual quality of the received video. The general optimization framework is formulated as a distortion minimization problem under certain constraints, typically the delivery delay constraint [16-20], and the transmission power constraint [21-24].

While the well-established WSN infrastructure and video coding standards can be utilized in the surveillance system, many new technologies have been proposed to accommodate the special requirements of the surveillance applications, such as the target object tracking, content-aware resource allocation, and the delay or power constrained video transmission. To bring out the optimal configuration, the major components in a surveillance system, including the video capture and preliminary vision tasks, video coding and transmission, and video analysis at the receiving end, have to be jointly considered. Under this circumstance, the cross-layer control proves to be a natural and desirable measure for system-level optimal resource allocation.

In this thesis work, we study the delay sensitive cross-layer control in the surveillance system, based on three practical surveillance applications, including the dynamic video object tracking with a single PTU camera, the binocular video object tracking with fast disparity estimation and the 3D video streaming, and the multi-camera motion estimation for remote healthcare monitoring.

In the first project, a single PTU (pan-tilt-unit) camera is used for video object tracking in an unmanned surveillance environment. The problem studied in this work is how to improve the quality of the received video by jointly considering the coding and transmission process at the source node. The cross-layer controller should be able to estimate the end-to-end video distortion, and to determine the optimal coding and transmission parameters, taking into account the dynamic channel condition and the transmission delay constraint. The error concealment strategies need to be considered for video quality enhancement, and the special property of the PTU camera motion can be utilized to accelerate the processing.

In the second project, the binocular PTU camera is adopted for video object tracking. The presented work studied the fast disparity estimation algorithm and the 3D video transcoding over the WSN for real-time applications. In the surveillance scenario, the disparity/depth information is used to adjust the tracking window size by the MeanShift tracking procedure, so as to improve the accuracy of the tracking results. The transcoding process for the 3D video, including one video sequence and its corresponding disparity data, can be coordinated by the cross-layer controller based on the channel condition and the data rate constraint, in order to achieve the best view synthesis quality at the receiving end with different display dimensions.

The third project is applied for multi-camera motion capture in remote healthcare monitoring, aiming to provide caregivers with timely access to the patient's health status through mobile communication devices. The challenge in this work is the resource allocation problem for the multiple video sequences delivered over the WSN for the 3D joint motion estimation by the receiver, rather than the resource allocation for a single

video in the traditional video surveillance system. The cross-layer control is designed to incorporate the delay sensitive, content-aware video coding and transmission, and the adaptive video coding and transmission procedures to jointly allocate the communication resources for multiple sequences, ensuring the optimal and balanced quality for the multi-view videos.

In these research projects, the essential idea is that the three major components in the surveillance system, namely the video capture, the video compression and transmission, and the video analysis, should be seamlessly cooperating under the cross-layer optimization framework. It is demonstrated through extensive experimental results that the presented cross-layer optimization schemes are preferable for enhancing the performance of a wireless video surveillance system.

1.3 Outline

The remainder of this dissertation is organized as follows.

In Chapter 2, we introduce the cross-layer optimization mechanism in wireless multimedia communications, and the special requirement for implementing cross-layer design in wireless video surveillance.

In Chapter 3, we present a wireless surveillance system for dynamic video object tracking with single PTU camera. The system contains three major modules, PTU camera control for surveillance video capture, cross-layer control for data compression and transmission, and error concealment for video quality enhancement. The system design for data collection and transmission over wireless networks is evaluated with physical surveillance equipments. The camera is capable of automatically following the moving target according to the control information. The target object can be segmented using

background subtraction, based on the special property of the PTU camera movement. The end-to-end distortion estimation in the delay constrained video coding process takes into account the dynamic channel condition and physical layer MCS to determine optimal coding and transmission parameters. Moreover, multiple error resilience and error concealment strategies, including interleaving, boundary match and video up-sampling, are applied utilizing the special property of the PTU camera motion. Experimental results show the efficiency of the surveillance system, and the superiority of the cross-layer optimization scheme, compared to the traditional video delivery scheme.

In Chapter 4, a binocular video object tracking system is designed with runtime 3D video content generation and data streaming over WSN. The purpose of applying two cameras is to generate the disparity/depth information, in order to adjust the tracking window size according to the distance between the target and the camera. Meanwhile, the 3D video content, including one video sequence and the corresponding disparity information, are transmitted for more advanced surveillance applications. To realize the real-time tracking, a fast disparity estimation algorithm is proposed. The disparity estimation process for each stereoscopic pair is formulated as an energy minimization problem. The iterative solution procedure is implemented in a course-to-fine manner. The estimated disparity is used to scale the tracking window by the Mean Shift algorithm, i.e. the size of the tracking area is adjustable according to its inner disparity, and thus the moving object can be better located by the camera. The program maintains the semi-real-time performance and acceptable accuracy as evaluated on a set of standard test data. In our experiment, two PointGrey cameras are controlled through a PTU device. The disparity estimation process on the recorded tracking video (640x480) achieves 6fps on

an ordinary PC (2.66GHz CPU, 4GB RAM). The 3D video content is transmitted over a heterogeneous WSN where the transcoding is required to adjust the frame size according to the terminal display device. A cross-layer optimized transcoding scheme is proposed to select the optimal quantization parameters for re-encoding the 3D video data.

In Chapter 5, a multi-camera motion capture system for healthcare monitoring is presented aiming to provide caregivers with timely access to the patient's health status through mobile communication devices. The major components include video capture, object detection, video coding and transmission, error concealment, and video analysis. In this surveillance system, several novel ideas are developed, including fast object detection, and content-aware and adaptive video coding and transmission. All components are seamlessly integrated in a unified optimization framework dedicated for online data transmission. In the scenario, the subject walked on a treadmill with four tripod cameras capturing the video from different viewpoints. After video compression and transmission over a wireless sensor network, the remote receiver recovered the videos and performed multi-view motion capture for gait analysis. Experimental results show that the presented system design achieves better video quality than traditional video coding and transmission scheme, while the requirement for a low-cost, noninvasive and real-time healthcare monitoring system is accommodated.

The summary of our research contributions and the future research directions are provided in Chapter 6.

Chapter 2

Cross-layer Optimization in Wireless Multimedia Communications

2.1 Introduction

The unstable channel condition and limited resources post great challenges for data communications in wireless sensor networks. This problem is more conspicuous for transmitting the large amount of multimedia data. Different from the traditional OSI communications model with virtually strict boundaries between layers, the cross-layer optimization mechanism enables systematic coordination and jointly decision making for all layers based on the resource constraints, and thus is considered a desirable and promising measure for wireless multimedia communications.

The cross-layer control has been widely adopted in various wireless communication systems [25]. Popular applications include the congestion control in single-path and multi-path routing [26], the adaptive modulation and channel coding for data transmission under delay and error performance constraints [27], and the subcarrier allocation for multiple users in OFDM wireless networks [28]. A common methodology in these cross-layer control schemes is to estimate the end-to-end signal error rate over the wireless channel for the designated resource configuration. These designs mainly focus on lower layers in the communication systems, such as the transmission path selection, channel access multiplexing, automatic repeat request, and the modulation and channel coding. When applied to multimedia communications, the source processing technologies in the application layer provide much more powerful error adaptation capability for the cross-layer controller. Take the video streaming as an example. One representative paradigm is the cross-layer optimized video coding and transmission

scheme [29] that adjusts both the coding parameters (coding mode, quantization parameter) and the transmission parameters (path selection, MCS) according to the video playback delay constraint. This cross-layer control scheme evolves from the joint source channel coding scheme [30, 31] which adjusts the coding parameters based on the varying data rate limit. The presented cross-layer control model is demonstrated in Figure 2.1.

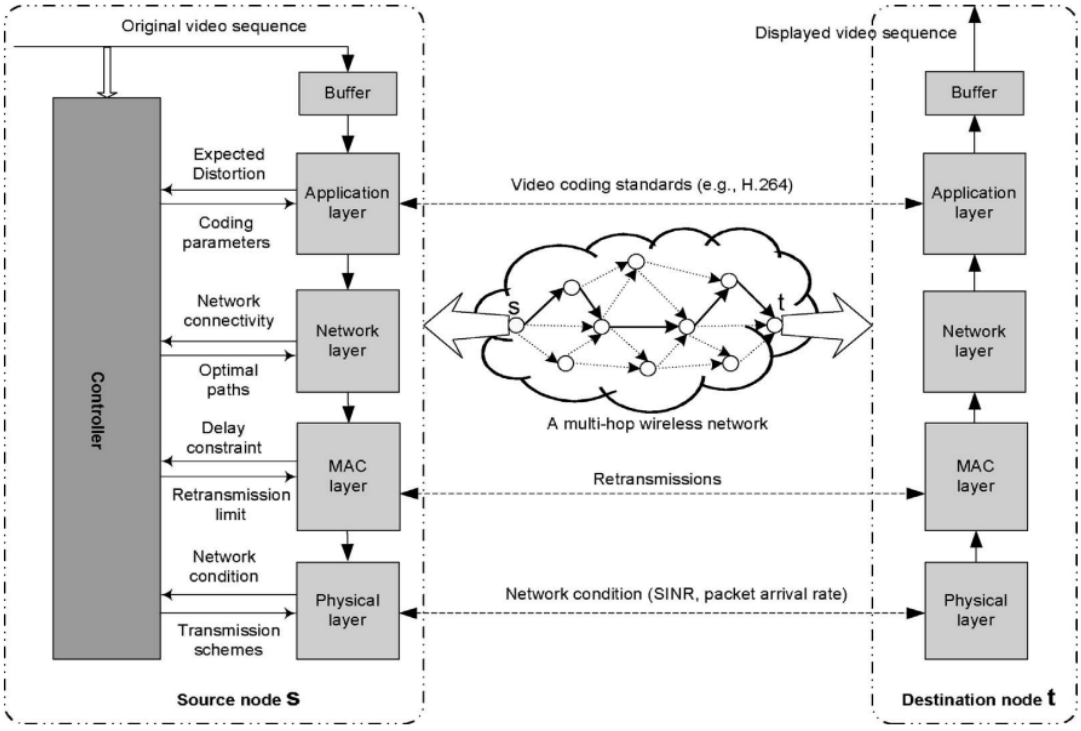


Figure 2.1 : A cross-layer control model for wireless video streaming [29].

The various error resilience and error concealment methods in traditional video source coding can be applied and be considered in the error estimation process by the cross-layer controller. A popular error resilience technique in the video source coding process is multiple description coding (MDC). The main concept of MDC is to create

several independent descriptions that contribute to one or more characteristics of the original signal: spatial or temporal resolution, signal-to-noise ratio (SNR), or frequency content [32]. MDC is considered an efficient measure to counteract bursty packet losses. Its robustness lies in the fact that it is unlikely the portions of the whole set of descriptions corresponding to the same part of the picture all corrupt during the transmission. Each description can be independently decoded and the visual quality is improved when more descriptions are received. The compression efficiency of MDC is affected due to reduced redundancy in each description. The extra overhead is largely ignored when otherwise the complex channel coding schemes or the complex communication protocols have to be applied in the presence of high packet loss.

Other video source coding error resilience measures include the scalable video coding (SVC) [33] and the distributed video coding (DVC) [34]. The error concealment techniques are also considered in the end-to-end distortion estimation process. They are designed to restore the missing packets in the received video, based on the encoding patterns at the transmitter. For example, the slice/frame copy is often adopted to replace the lost packets that are encoded as an entire slice/frame at the source node [30], and the boundary match [35] algorithm is developed for the concealment of the missing blocks that are encoded in an interleaved manner.

Furthermore, the source processing technologies are incorporated with the transmission process in some cross-layer designs to synergize the system performance enhancement. One typical example is to apply the unequal error protection (UEP) mechanism based on the data type. The idea of UEP is to allocate more resources to the parts of the video sequence that have a greater impact on video quality, while spending

fewer resources on parts that are less significant [21]. In the cross-layer optimized video communication system presented in [36], the authors applied the priority-based discarding of video packets. The MAC always drops the packets belonging to the least important SVC layer first to allow more resource consumption by the more important layers in the queues. This way, the MAC ensures that the most important base-layer packets have the highest probability to get transmitted, and thus helping to maintain service continuity for the client. The RoI (region of interest) based wireless video streaming system introduced in [37] adopted multiple error resilience schemes for data protection. The RoI region was obtained through image segmentation, and was assigned more resources than other background areas, including chessboard interleaving rather than slice interleaving, higher degree of forward error correction and automatic repeat request.

Closely related to the practical system architecture, the cross-layer design needs to take into account the specific requirements of different applications, in order to properly configure the parameters. To further describe this mechanism, the rest of this section introduces the formulation of the cross-layer optimization problem, and its implementation in wireless video surveillance.

2.2 Formulation

The cross-layer optimization problem is generally considered the optimal resource allocation process under the system resource constraints [25, 38]. A standard formulation for the cross-layer optimization procedure in the communication system can be expressed as a utility maximization, or equivalently, a distortion minimization problem as follows,

$$min \sum D(\psi_1, \psi_2, \dots, \psi_n, p)$$

$$s.t. \quad \boldsymbol{C}(\psi_1, \psi_2, \dots, \psi_n) \leq \boldsymbol{C}^{max} \tag{2.1}$$

where D is the expected end-to-end data distortion under the system configuration set $\psi_1, \psi_2, \dots, \psi_n$. p is the expected data loss over the WSN given the same configuration. $\boldsymbol{C}(\psi_1, \psi_2, \dots, \psi_n)$ is the vector of corresponding consumed resources, and $\boldsymbol{C}^{max}$ represents the resource constraints. For multimedia data such as the video, the distortion estimation considers both the source coding distortion and the channel distortion. The source coding distortion is caused by data compression, while the channel distortion is due to the packet loss. Based on the application services provided by the system, the data distortion and the resource constraint expressed in Formula (2.1) could have different definitions, and the solution strategy to the optimization problem also varies, as discussed in the following subsections.

2.2.1 Optimization Goal

The expected end-to-end data distortion in Formula (2.1) is application dependent. For most video delivery tasks, the distortion is defined as the difference between the original data and the received data after compression and transmission over the wireless channel, or the utility is measured as the similarity between the original data and the received data, to evaluate the quality of service (QoS), or the quality of experience (QoE) provided by the system. Some commonly applied QoS or QoE metrics include MSE (mean square error), PSNR (peak-signal-to-noise-ratio), mean opinion score (MOS) and SSIM (structural similarity). For other vision tasks, the distortion is defined as the difference between the generated vision product using the original data, and the other generated

vision product using the received contaminated data. For example, in the cross-layer optimized transcoding scheme presented by Liu et al [39], the view synthesis distortion using the transcoded data is considered. The cross-layer control scheme designed by Thakolsri et al [40] distinctively defined the temporal video quality fluctuation as the optimization goal, aiming to provide a smooth viewing experience for the audience.

2.2.2 Optimization Constraint

The resource constraints expressed in Formula (2.1) also depend on the system. A commonly considered constraint for video communications is the transmission delay for real-time applications. To achieve real-time video delivery, the cross-layer optimized video coding and transmission scheme described in [19] considered the physical layer MCS in estimating the dynamic packet loss rate in a Rayleigh fading channel. For video streaming over multi-hop WSNs, the systems demonstrated in [17] and [18] enabled adaptive configuration for both the physical layer MCS and the link layer path selection. The work introduced in [20] incorporates congestion control with link adaptation for real-time video streaming over ad hoc networks.

Power constraint is another consideration for energy efficient mobile devices. In [22], node cooperation is applied to optimally schedule the routing in order to minimize the energy consumption and delay. The cross-layer design presented in [23] jointly configured the physical, MAC, and routing layers to maximize the lifetime of energy-constrained WSNs. The object based video coding and transmission scheme developed by Wang et al [21] performed UEP for the shape data and the texture data in the rate and energy allocation procedure.

Other resource constraints, such as the computation complexity [41], the fairness among multiple users [28], can be conveniently included in the cross-layer optimization framework described in Formula (2.1), according to the requirements of the specific application.

2.2.3 Solution Strategy

The definition of the optimization goal and constraints determines the unique solution strategy for each wireless multimedia communication system. As the implemented vision task gets more complicated, or the posted constraints are more complex, the traditional exhaustive full search method for all possible parameter configurations is time-consuming, and is infeasible for some mobile devices. A popular optimization solution strategy is to utilize the Lagrange relaxation, and to apply dynamic programming to obtain the solution given that the convexity (or concavity for a utility function) of the described problem is guaranteed, and that the distortion estimation is independent between two consecutive recursion steps [18, 21]. When the constraints are complex and thus even the dynamic programming algorithm is time-consuming, the approximate optimization solution strategy based on the relaxation is often applied, such as the fast optimization method for the power-rate-distortion optimized cross-layer control scheme presented in for video streaming over wireless sensor networks [42]. With this kind of approximation scheme, the solution is obtained in an iterative procedure. In current iteration, one part of the parameter set is fixed, and the optimal solution for the rest part is determined, which serves as the fixed part in the next iteration. The procedure terminates until the difference between two consecutive iterations falls below certain threshold, or the maximal iteration number is reached. In the cross-layer optimization strategy for data

communication over the OFDM wireless sensor networks developed by Song et al [28], the optimal resource allocation for multiple users is achieved by applying multiple greedy algorithms for resource allocation, for both concave and nonconcave utility functions.

2.3 Cross-layer Optimization in Wireless Video Surveillance

Most of the cross-layer optimization schemes for wireless multimedia communications are applicable for wireless video surveillance, if the major service provided by the system is the surveillance video delivery. In some scenarios, the primary concern is that the moving target object is of greater interest than the background, and should be given higher priority. When the communication resources are limited, an alternative of heavier compression is to implement unequal error protection for different parts of the video data. Hence the RoI based UEP mechanism is a natural way to optimize resource allocation in a wireless video surveillance system. The system design for surveillance video coding and transmission over wireless sensor and actuator networks (WSANs) proposed in [16] extended the UEP mechanism from source data processing to network management. Each intermediate sensor node in the selected transmission path put the target packets ahead of all the background packets in the queue. Thus the target packet had a lower packet loss rate than the background packet when the sensor node started dropping packets with a higher waiting time than the packet delay limit.

Current video coding standards provide different interfaces for RoI data processing. For example, the object based video representation is supported in the MPEG-4 standard [43]. A contour free object shape coding method compatible with the SPIHT codec [44] was introduced in [45]. In the latest H.264/AVC standard, several tools intended for error resilience like Flexible Macroblock Ordering (FMO) and Arbitrary Slice Ordering (ASO)

can be used to define the RoI region [46]. These interfaces enable convenient incorporation of the object based UEP mechanism in the coding and transmission process in the surveillance system.

In other scenarios, the focus of the service provided by the system could be other issues, such as the security and privacy concerns [47, 48], and other vision applications including super resolution [49], view synthesis [50], and 3D model reconstruction [51]. However, most of these technologies are either based on undistorted video data, or independent of the error control procedure at the transmitter. The impact of video compression on RD performance was considered in several vision applications for optimal source coding decision at the transmitter, including view synthesis [52, 53], object tracking [54], and super resolution [55]. Some JSCC schemes were embedded in the coding structure for optimal resource allocation based on end-to-end distortion estimation [56, 57]. The channel distortion model for more complex vision applications remains a challenging research topic.

2.4 Summary

Cross-layer optimization proves to be an efficient measure for wireless multimedia communications. For wireless surveillance video communications, new challenges are emerging in the process of compressing and transmitting large amount of video data, and in the presence of run time and energy conservation requirements for mobile devices. The requirements and constraints for different surveillance systems can be described under the general cross-layer optimization framework, to facilitate the resource allocation process. Another trend in this field is the 3D signal processing technology in more advanced multiview video surveillance. The wireless communication environment posts

greater difficulty for this kind of applications. How to efficiently estimate the distortion for the dedicated vision task at the receiving end using the compressed and concealed video data remains a research topic worth exploring.

Chapter 3

Video Surveillance with Single PTU Camera

3.1 Introduction

Wireless video sensor networks with active cameras are gaining increasing popularity in various surveillance applications such as intelligent transportation, environmental monitoring, homeland security, construction site monitoring, and public safety [58]. Current work on video surveillance with active cameras mainly focused on automatic camera control algorithms [59, 60]. To extend these algorithms into a wireless surveillance system, the dynamic channel condition and the data delivery delay constraint need to be considered in the data compression and transmission process, in order to meet the visual quality and real-time requirements demanded by the online application.

In block based video compression, the selection of coding mode and corresponding quantization parameter (QP) is an important process for rate-distortion (RD) control. While QP selection under predefined coding mode structure has been extensively studied [61-63], combining mode selection could enhance the coding results for its inherent adaptability [64, 65]. In [64], the intra/inter mode selection process is formulated as a Lagrange cost minimization problem solved by dynamic programming. The authors in [65] further explored the efficiency of multi-resolutional coding by adaptively selecting among intra, inter and down-sampling modes for each macro block (MB). It was reported that better performance is acquired under low data rate constraint [65, 66], since smaller QPs could be chosen for down sampled MBs. These source coding methods merely look into the quantization induced distortion. In packetized video transmission over wireless networks, packet loss is also a major cause for the data distortion observed by the

receiver. Incorporating the packet loss information in the end-to-end distortion estimation process has been proved to be an efficient measure to improve the coding efficiency [30, 31]. The coding method described in [19] further considered the physical layer MCS to estimate the dynamic packet loss rate (PLR) in a time-varying Rayleigh fading channel. Moreover, the video streaming systems discussed in [17] and [18] enabled adaptive configuration for physical layer MCS and link layer path selection through cross-layer control. To enhance the overall system performance, we consider the flexible configuration for both the video coding parameters and the MCS. Intra, inter, down-sampling and skip (packet drop) coding modes are available in the compression process, and the MCS is determined according to the distortion estimation result based on the corresponding PLR and data rate.

Error concealment is another important factor in quality enhancement. In block based codec, interleaving is an effective error resilience technique to avoid losing consecutive MBs in one packet bearing strong spatial or temporal correlation with each other [37, 67, 68]. The proposed interleaving mechanism is capable of grouping the MBs into the designated packet number with controllable decomposition depth. For a lost MB, after deinterleaving, patches from previous frames are compared to its boundary pixels [35]. The property of PTU camera motion is exploited in the patch search process to carry out an efficient implementation. Finally, the down-sampled data is recovered to the normal resolution using the total variation (TV) method [66, 69].

The architecture of the wireless video surveillance system is illustrated in Figure 3.1. The transmitter is in charge of video capture and compression for data transmission. The PTU camera is controlled by both the commands from a remote control unit (RCU) and

the local tracking result. The captured video is recorded by the local tracker and is sent to the data processing center. The data center collects channel state information (CSI) and handles the procedure of video compression and system configuration. Camera parameters, such as the focal length, the camera center, and the pan/tilt angles are transmitted along with the video packets to the RCU. The RCU performs error concealment on received data and provides necessary feedback to the transmitter, including the packet loss information and the camera redirection commands.

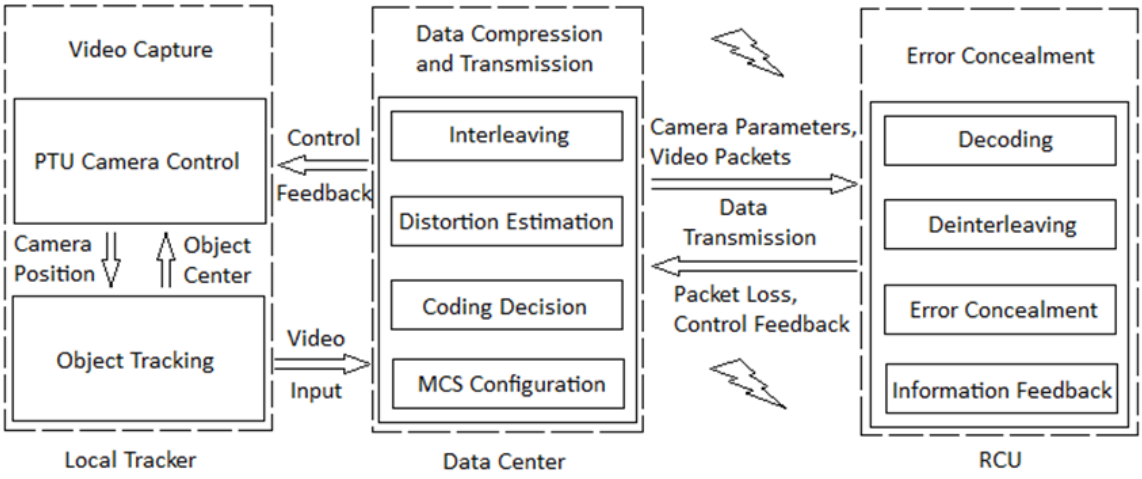


Figure 3.1 : A video surveillance system with single PTU camera.

In the following sections, we will introduce the detail function of each component in the surveillance system, including the property of PTU camera movement and the object tracking algorithm for video capture, the cross-layer optimization procedure for data compression and transmission, and the error concealment strategies for video recovery at the RCU.

3.2 PTU Camera Model

Object detection is a primary task in video surveillance applications. An automatic, intelligent surveillance system should be able to monitor moving objects within certain areas and extract the most important information. Therefore both tracking and detection are critical to the performance of such a system. Furthermore, the method adopted in each system must take into account the specific camera arrangement. For example, if multiple cameras are placed at different locations with static pose, the tracking process should be able to coordinate these camera outputs so that the detection process knows how to locate useful information. If the system has control over camera movement, usually fewer cameras are required, with increased complexity in software algorithms since the tracking process needs to incorporate camera control and detection has to deal with changing background.

A freely moving camera is usually infeasible in an unmanned surveillance environment due to technical or fiscal constraints. A more practical scheme is to use a PTU camera, where the camera projection center is generally unchanged and the retinal plane is capable of angular movement. In this kind of system, the camera control algorithm for tracking process needs to estimate the angular speed/acceleration of the moving object, and background alignment in different video frames is required for motion detection. In the PTU camera tracking algorithm proposed by Petrov et al. [59], a linear feedback controller is applied based on the theory of Lyapunov Stability. The control parameters are updated by object position estimated using Mean Shift [70] algorithm. The unique geometric property of a PTU camera model is that the camera projection center remains unchanged while the pan and tilt angles are controllable, as illustrated in Figure 3.2 (a) [59]. The focus F denotes the projection center. The image

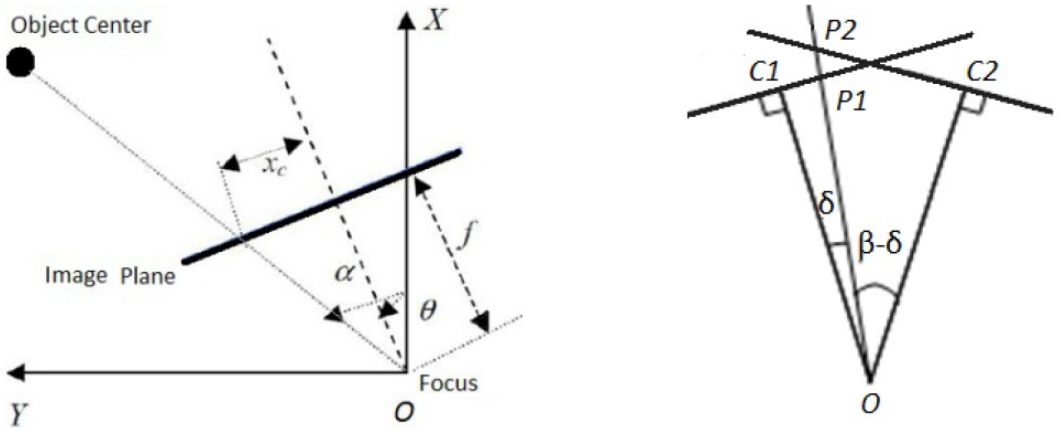
plane is viewed down along its *y* axis, and is projected on the *X*-*Y* world coordinate plane. α is the angle between the object center and the *X* axis, θ is the angle between the image center and the *X* axis, *f* is the focal length, and x_c is the distance between the projected object center and the image center along the *x* axis. Only pan control is displayed in the figure. The algorithm applies to tilt control similarly.

The linear feedback controller aims to minimize x_c and the difference between the estimated object speed and the measured object speed. According to the Theory of Lyapunov Stability, the camera angular speed $w_{\theta k}$, the camera angle θ, and the estimated distance x_c are updated at every time instance:

$$w_{\theta k} = \hat{w}_{\alpha k} + p\, k x_{ck} \tag{3.1}$$

$$\theta_{k+1} = \theta_k + w_{\theta k}\Delta t_k \tag{3.2}$$

$$\hat{x}_{ck+1} = f\, tan(\hat{\alpha}_{k+1} - \theta_{k+1}) \tag{3.3}$$



(a) Object tracking. (b) Corresponding points.

Figure 3.2 : PTU camera model.

where $\widehat{w}_\alpha$ is the estimated object angular speed, and $\hat{\alpha}$ is the estimated object angle. p is a parameter used to control the convergence speed in the dynamic system. And Δt_k is the duration of the k-th time interval for the control parameters update. Different from [59] where the constant time interval is applied, we consider the control delay in the real communication system and adopt noisy interval values in Equation (3.2) to test the performance of our proposed method. The updated camera motion information and the recorded video data are then processed at the data center prior to transmission.

With a fixed focus position, it is observed that the captured video possesses a special correlation between the corresponding background pixels in two consecutive frames. As demonstrated in Figure 3.2 (b), *C1* and *C2* are the image centers (origins of the image coordinates) in previous and current frames, *P1* and *P2* are the image pixels projected on to these two frames from the same static object in the surveillance environment, and β is the pan (or tilt) angle between two camera positions. The correlation of the coordinates of *P1* and *P2* can be expressed as,

$$\begin{aligned} x_2 &= -f \cdot tan(\beta - \delta) \\ &= -f \cdot tan\left(\beta - tan^{-1}\left(\frac{x_1}{f}\right)\right) \end{aligned} \tag{3.4}$$

where x_1 and x_2 are the x (or y) coordinates of *P1* and *P2*, respectively. Even if the object is moving, its corresponding pixel can still be found in a nearby neighborhood according to Equation (3.4) in previous frame, as long as it is not occluded. Thus given the camera parameters and the coordinate of a pixel in current frame, the motion searching process for the pixel's counterpart in previous frames can be accelerated. This property will also be utilized in the error concealment process, as discussed in Section 3.5.

3.3 Video Object Tracking

Object detection with fixed camera often takes advantage of static background. A commonly used technique is background subtraction based on Gaussian Mixture Model (GMM) [71]. This temporal learning process models different conditions of a pixel at certain position as a mixture of Gaussian distributions. Weight, mean and variance values of each Gaussian model can be updated online, and pixels not conforming to any background model are quickly detected. The adaptive learning property makes this technique suitable for real-time applications, and a variety of detection methods are developed combining other spatiotemporal processing [72-74]. For object detection with moving cameras, more factors need to be considered. Mean Shift is efficient in locating object position according to the object's color distribution [75, 76]. However, it can only indicate the approximate position, rather than the shape of the tracked object. Hence further processing is necessary to bring out more accurate information. Among all the features that are useful for shape detection, motion is one of the most essential. The first step for motion detection with a moving camera is background alignment. While traditional optical flow approach is very time-consuming and loses accuracy when the camera has large motions [75], features such as corners and edges [77, 78] are often used to estimate the geometric relation between different video frames. In an affine camera model, this relation can be represented by either the fundamental matrix [79] or the homography [78]. We will show in the next subsection that the fundamental matrix is not suitable for a PTU tracking system, and the multi-layer homography model [80] is unnecessary, because of the special property of the camera movement. Moreover, we propose an object detection method based on the analysis of PTU camera model. In the

tracking process, one PTU camera is used. The linear control algorithm in [59] is used to follow the object's movement. The Mean Shift estimation and the control algorithm work in an interactive way to enhance accuracy and speed up convergence. The moving object can thus always be visible in center area of the image plane. In the detection process, first the background is aligned using the homography computed from RANSAC (Random Sample Consensus) fitting. Then correlation between the video frame and aligned background image is adaptively thresholded according to the distance between each pixel and the estimated object center from Mean Shift. A variational level set method [81] is applied afterwards to contour the outline and remove noises. The experiment on our PTU camera video demonstrates promising results.

3.3.1 Background Alignment

With a moving camera, the object detection procedure has to consider the background difference among the video sequence. Background alignment for motion detection is usually implemented by computing the fundamental matrix [79] or homography [78]. However, the fundamental matrix method fails to handle the situation when the camera projection center is not moving. As shown in Figure 3.3, the 2D points m_1 and m_i projected on two image planes from the same 3D point M are related through the following fundamental matrix *F*:

$$m_1 F m_i = 0 \tag{3.5}$$

When the camera centers C_1 and C_2 overlaps, the computed matrix *F* is singular and it is very sensitive to noises. By contrast, the homography method estimates the direct mapping between corresponding points. And there is no need to worry about the multi-

layer homography problem described in [80] since a point on one projection ray will always be projected to a fixed position in the image plane regardless of the depth of this point. As demonstrated in Figure 3.2 (b), if a point locates on the ray *OP1*, where *O* is the projection center, and *P1* is its projected point in image plane *C1*, its projection on image plane *C2* will be *P1*, no matter how far it is from the projection center.

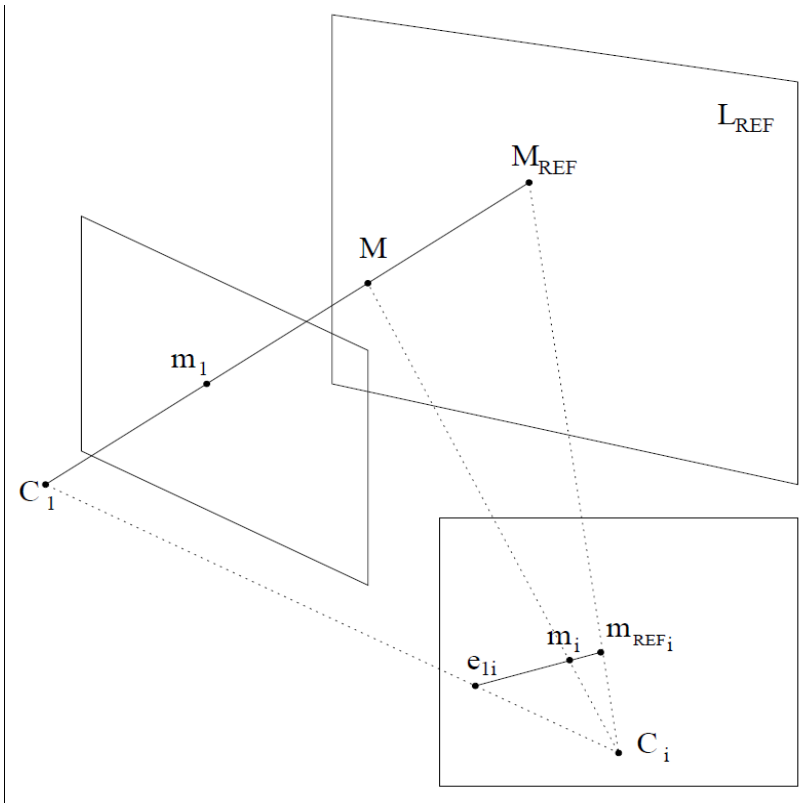


Figure 3.3 : Background matching with fundamental matrix.

Therefore, the homography method is adopted in the system for background alignment, which is then used for object segmentation, as introduced in the next subsection. In the alignment process, the Harris corner finder [87] is applied to detect feature points in both video frame and the background image, as shown in Formula (3.6), where ∇I_{xy} is the gradient of a pixel in image *I*, and ∇I_x , ∇I_y are the projected gradients

in the direction of x and y axis. ε is a small positive number to prevent the evaluation from getting infinitely large in some smooth areas. Figure 3.4 (a) displays the detected feature points in the pre-captured background image.

$$\max\left(\frac{\nabla I_x^2 \cdot \nabla I_y^2 - \nabla I_{xy}^2}{\nabla I_x^2 + \nabla I_y^2 + \varepsilon}\right) \tag{3.6}$$

Corners in one image I_1 find their matches in the other image I_2 with highest correlation.

$$\max\left(\frac{\sum_{\aleph}\left(I_1(x_i) \cdot I_2(x_j)\right)}{\sum_{\aleph} I_1^2(x_i) \cdot \sum_{\aleph} I_1^2(x_j)}\right) \tag{3.7}$$

And mismatched pairs are eliminated by RANSAC fitting, through which the homography is computed:

$$\min(\|Hx' - x\|^2) \tag{3.8}$$

Here H is the homography, and x', x denote the coordinates of a matched pair in two images. The matching results can be observed from Figure 3.4. As seen in the ceiling light area, the mismatched pairs in Figure 3.4 (b) due to similar correlation values are eliminated in Figure 3.4 (c) by the homography method, and the background is correctly aligned with the video frame.

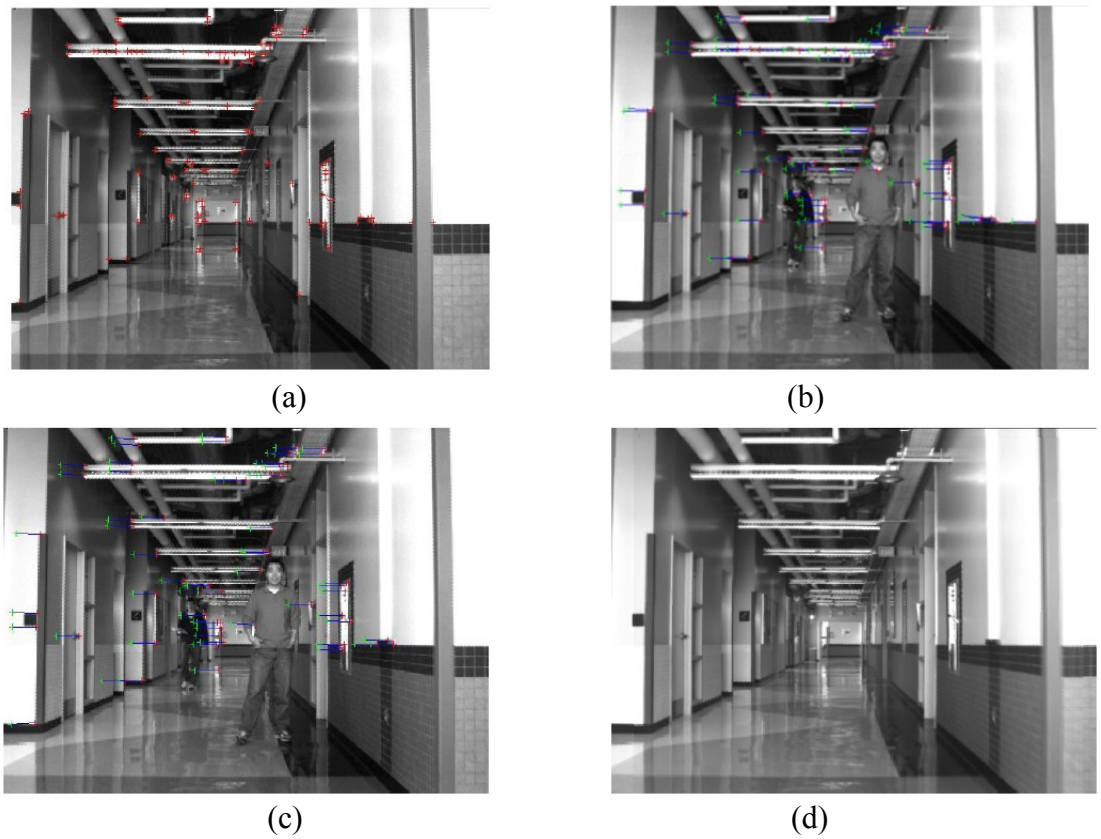


Figure 3.4 : (a) Background image with detected corners; corner pair matching by (b) fundamental matrix method and (c) homography method; (d) aligned background.

3.3.2 Object Segmentation

In the video object tracking procedure, object segmentation is provided for more accurate object detection purpose. The system architecture of the two step object detection method is demonstrated in Figure 3.5. In the tracking process, two modules work interactively. The Mean Shift algorithm provides the measured object position as input for parameters update in PTU camera control, while the updated control parameters provide initial position for Mean Shift at next time instance in order to speed up the convergence. In the detection process, firstly background is aligned with each video frame. A background

image necessary for subtraction can be synthesized either from several pre-captured backgrounds [78], or from some learning methods [82, 83] using other video frames.

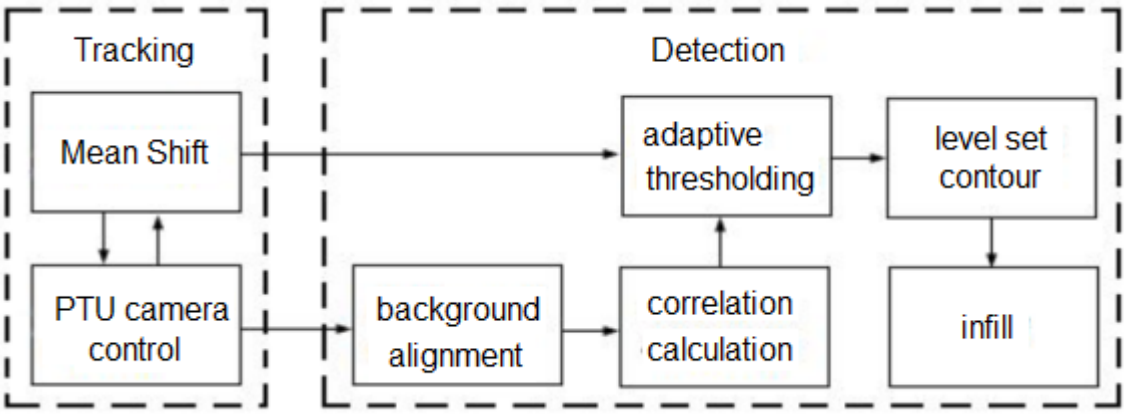


Figure 3.5 : Object detection with segmentation.

After aligning the background using the method introduced in previous subsection, the correlation value of each pixel in video frame with aligned background is calculated and thresholded according to its distance from the object center (x_c, y_c) estimated by the Mean Shift algorithm,

$$T(x,y) = A\, e^{-\frac{(x-x_c)^2+(y-y_c)^2}{B^2}} \tag{3.9}$$

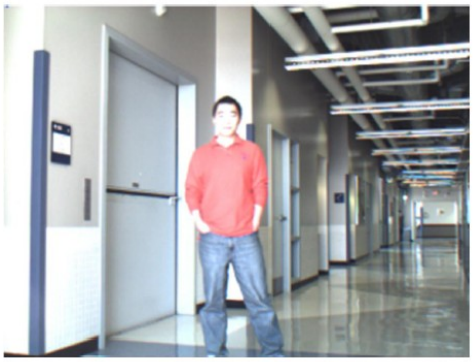
$$I'(x,y) = \begin{cases} I(x,y), & \text{if } \mathrm{Corr(x,y)} < T(x,y) \\ 0, & \text{otherwise} \end{cases} \tag{3.10}$$

The parameter A and B are predefined constants used to adjust the range of the threshold T. By adopting such an adaptive measure, foreground areas around the object center will have higher threshold and can be more easily detected than boundary areas, where the lighting condition is volatile and the subtraction result is prone to suffer. After

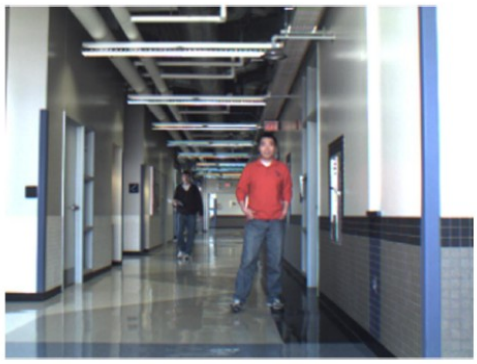
thresholding, the variational level set method proposed by Li *et al.* [81] is applied on the residual image to contour and infill the outline of foreground objects. Noises can be further removed due to the length shrinking property of the evolutional curves. During the implementation of the detection process, a multi-resolution strategy is adopted to reduce the computation.

The experimental results with our PTU (Directed Perception D47) camera tracking system are demonstrated in Figure 3.6. Two different video frames are displayed in Figure 3.6 (b) and (c) to illustrate the tracking result. The walking person dressed in red is always visible in the center area of the image. The background alignment result can be observed in Figure 3.4 from the previous subsection. The correlation threshoding results with constant threshold values are provided in Figure 3.6 for comparison. The threshold value is set to 0.3 (Figure 3.6 (d)) and 0.7 (Figure 3.6 (e)) respectively. The constants defined in Equation (3.9) are selected such that $A = 0.8$, and B equals to one fourth of the image diagonal length. As shown in Figure 3.6 (f) - (g), our adaptive thresholding method is more robust to noises, thus it can facilitate the level set contour process. As can be observed from the outline infill result in Figure 3.6 (h), our proposed detection method can bring out foreground object shape efficiently, even only intensity values of the video data are processed.

(a)



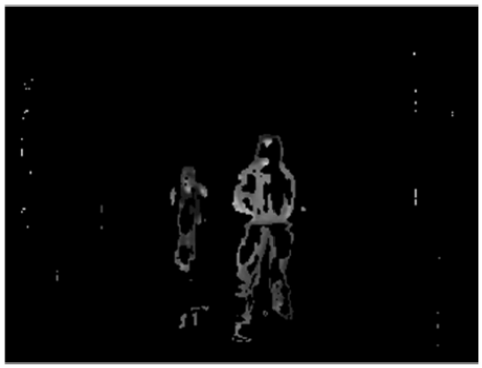
(b)



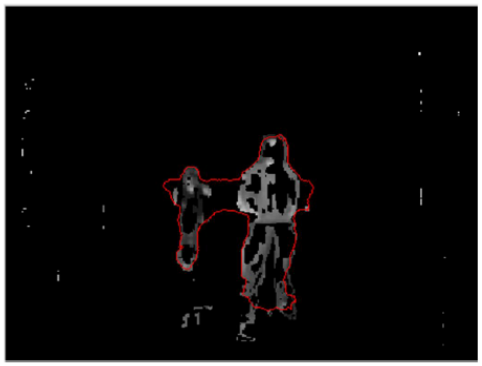
(c)

(d)

(e)



(f)



(g)

(h)

Figure 3.6 : (a) PTU camera; tracking (b) - (c) and detection (d) - (h). Correlation thresholding with (d) - (e) constant and (f) adaptive threshold; (g) level set contour; (h) infill.

The video object detection method presented in this paper utilizes the special geometric property of PTU camera movement. In the tracking process, Mean Shift and the camera control model work interactively to enhance accuracy and achieve real-time performance. This tracking mechanism centers the object in image plane and provides useful information for detection process. In the detection process, foreground areas are efficiently detected through background alignment and correlation thresholding. The adaptive thresholding scheme works successfully given the tracking position information. Experimental results on intensity data are encouraging. The proposed method is suitable for resource limited surveillance applications.

3.4 Cross-layer Control

Video communications over wireless networks face various challenges including power and bandwidth constraint, random time-varying channel fading effect, network heterogeneity, and quality of service requirement [84]. Information from different network layers is necessary to be incorporated in data compression/coding procedure in order to achieve better system performance. Therefore, alongside with the development in source coding strategies [61, 64, 65, 85, 86], increasing dependence on cross-layer optimization is observed in emerging mobile multimedia applications [17-19, 21, 87].

In video compression, coding mode and corresponding quantization parameter (QP) selection is an important process for rate-distortion (RD) control. While QP selection under predefined GOP structure has been extensively studied [18, 19, 61], combining

mode selection could enhance the coding results for its inherent adaptability [64]. Traditional mode selection methods mainly consider RD results obtained in application layer, i.e. given rate or buffer constraint, quantization distortion is minimized by optimal coding decision. A plethora of research work has been done in this area. In [64], the mode selection process is formulated as a Lagrange cost minimization problem solved by dynamic programming. Other information such as local edge or block boundary difference is also utilized to accelerate the coding decision process [86]. The authors in [65] further explored the efficiency of multiresolutional coding by adaptively selecting among intra, inter and down sampling modes for each macro block (MB), and reported that better performance is acquired under low data rate, since smaller QPs could be chosen for down sampled MBs.

These source coding methods merely look into quantization induced distortion. In packetized video transmission over wireless networks, packet loss is also a major cause of distortion at receiver's side. In [30], the 'recursive optimal per-pixel estimate' (ROPE) method is presented for MB coding mode decision. Expected end-to-end distortion is estimated with certain PLR. This statistical model demonstrates a new way to adjust coding decision according to both source coding and channel distortion, whereas the impact of dynamic channel condition on RD results is not considered. Another joint source channel coding (JSCC) method introduced in [31] adopts random intra refreshing. Source coding distortion is modeled as a function of the intra MB refreshing rate, while channel distortion is calculated in a similar recursive fashion as is done in [30]. In this method, channel coding rate and forward error correction (FEC) are considered, yet some

parameters need to be determined beforehand exclusively for each video, and the time varying channel condition is still ignored.

To cope with channel fading, the object based video coding method described in [21] calculates practical channel capacity in a Rayleigh fading channel. Distortion for intra and inter mode coding is estimated separately with resulting PLR. Discriminative coding decision is determined for shape and texture data under delay and transmission power constraints. The work in [19] takes into account physical layer MCS, and adaptively estimates PLR in a Rayleigh fading channel based on convolution coding and BPSK modulation. More flexible MCS configuration is applied in [17, 18] through cross-layer design with channel information feedback. However, these methods provide no specification for online coding mode selection. Exhaustive searching is time consuming; dynamic programming and random intra refreshing prearrange coding decision on consecutive packets/frames, and might not be suitable with online channel information feedback. Thus we propose to incorporate cross-layer design in coding decision. The mode selection process is formulated as a delay constrained distortion minimization problem. Optimal decision is carried out for each packet by a cross-layer controller with adaptive MCS configuration. In our experiment, intra or inter mode is selected for MBs in each packet under proper MCS, and down sampling is considered an alternative for inter coding in low data rate transmission.

3.4.1 Problem Formulation

The object tracking procedure ensures that the moving target is visible in the recorded video. The recorded video is then processed at the data center for compression and transmission over wireless networks. The interleaving step before video encoding divides

the MB data in one frame into a desired number of packets. Details of this interleaving and deinterleaving method will be introduced in Section 3.5. In the surveillance application, the encoded data rate is constrained by a frame delay T_n, and the coding parameters for each frame are selected according to the estimated distortion D:

$$(Q^*, M^*, m^*, r^*) = arg\ min \sum_{k=1}^{K} D_{n,k}(Q, M, m, r)$$
$$s.t. \quad \sum_{k=1}^{K} \frac{L_{n,k}(Q,M,m,r)}{R_n(m,r)} \leq T_n \tag{3.11}$$

Here Q and M represent the QP and the coding mode. (m, r) denotes the physical layer modulation and FEC code pair as defined in IEEE 802.16 [88]. Numerically, m is the modulation order and r is the FEC code rate. $L_{n,k}$ is the encoded data length of the k-th packet in the n-th frame, and R is the data rate limit for each frame calculated from the channel bandwidth W and the SNR γ:

$$R(m \cdot r) = m \cdot r \cdot W \cdot log_2(1 + \gamma) \tag{3.12}$$

The data distortion is estimated with the coding parameters, the dynamic channel condition and the configurable MCS. To better illustrate the distortion estimation process, some assumptions adopted in this work are listed below.

- The channel condition remains time invariant for one frame, but varies from frame to frame. In the simulation, a Rayleigh fading channel is modeled with an exponential distributed SNR:

$$\mathcal{F}(\gamma) = \frac{1}{\bar{\gamma}} e^{-\frac{\gamma}{\bar{\gamma}}} \tag{3.13}$$

Here $\bar{\gamma}$ is the average received SNR. The resulting PLR associated with a specific (m, r) is estimated using the 802.11a WLAN packet loss model described in [87].

- Perfect channel CSI is available to the receiver and is fed back to the transmitter without error and latency. This assumption could be approximately satisfied by using a fast feedback channel with powerful error control information as adopted in [88].
- The RCU provides packet loss information to the transmitter.
- The data center is capable of cross-layer coordination, information collection and system configuration throughout the network architecture.
- A one-hop scenario is assumed in the transmission process, and other communication overhead is ignored.

3.4.2 Distortion Estimation

To determine the optimal coding parameters, data distortion is estimated for each coding mode and its optional QPs. Four coding modes, intra, inter, down-sampling and skip modes, are available to accommodate different channel conditions. The expected end-to-end distortion for one packet contains both the source coding distortion D_s, and the channel distortion D_c.

$$D = D_s + D_c \tag{3.14}$$

D_s is mainly determined by the QP. For the down-sampling coding mode, the corresponding up-sampling method is also considered. The packet loss induced D_c is

estimated in a recursive fashion regarding previous packet loss condition [30, 31]. Specifically, for a pixel i with a value of f in an intra coded packet in the n-th frame, given the PLR p, D is proportional to the distance between f and the concealed value $\tilde{f}$ from a pixel j in previous frame:

$$D_c(n,i) = p(n) \cdot (E\left\{\left[f(n,i) - \tilde{f}(n-1,j)\right]^2\right\} - D_s(n,i)) \tag{3.15}$$

In an inter coded packet, the estimated channel distortion of the pixel is related to the channel distortion of a motion predicted pixel h in previous frame, and the distortion caused by the concealed pixel j (not necessarily the same as h) in previous frame:

$$\begin{aligned} D_c(n,i) = &\left(1 - p(n)\right) \cdot D_c(n-1,h) + p(n) \\ &\cdot \left(E\left\{\left[f(n,i) - \tilde{f}(n-1,j)\right]^2\right\} - D_s(n,i)\right) \end{aligned} \tag{3.16}$$

For a packet coded in the down-sampling mode, the down-sampled data is compressed with intra coding. The expected channel distortion is estimated in a similar fashion as for the intra coding:

$$D_c(n,i) = p(n) \cdot \left(E\left\{\left[f(n,i) - \tilde{f}(n-1,j)\right]^2\right\} - D_s(n,i)\right) \tag{3.17}$$

Under poor channel condition, we consider the skip coding mode a more desirable option than applying large quantization step. When the packet is dropped, the distortion solely depends on the error concealment result:

$$D(n,i) = E\left\{\left[f(n,i) - \tilde{f}(n-1,j)\right]^2\right\} \tag{3.18}$$

The received pixel value $\tilde{f}$ at the receiver's end is estimated with the quantized pixel value $\hat{f}$:

$$E\{\tilde{f}(n,i)\} = (1-p(n)) \cdot \hat{f}(n,i) + p(n) \cdot E\{\tilde{f}(n-1,j)\} \quad (3.19)$$

The recursion step is decided by the data processing center according to the received packet loss feedback from the RCU, and the system's computation capability. If the transmitter receives no loss feedback after $\triangle$ frames, a random packet loss decision will be made by the encoder for that frame based on its estimated PLR.

3.4.3 Parameter Selection

The video sequence is coded in a way that one of the three coding modes, intra, inter and down-sampling (with a down sample rate of two in this work), is assigned for each frame with a uniform QP for all MBs. If the resulted coding rate exceeds the data rate limit in Equation (3.12), skip mode is applied to one or more packets from that frame until the rate or delay constraint is satisfied. The optimal coding parameters and the MCS configuration are selected according to the distortion estimation result by Equations (3.11) to (3.19). Several approximation methods for source coding rate and distortion estimation can also be adopted to reduce the computation overhead [19, 31].

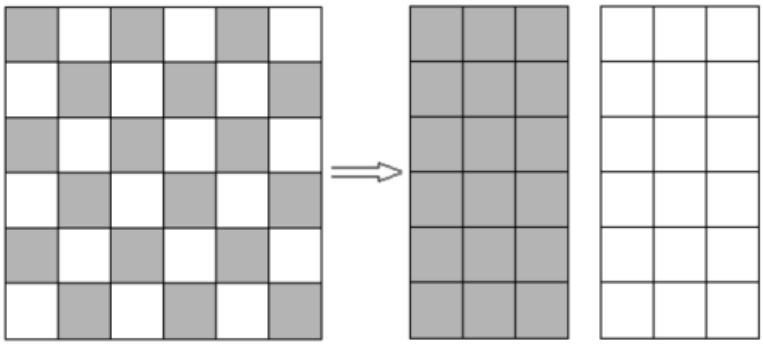
3.5 Error Concealment

The channel distortion estimation process discussed in previous section is related to the error concealment strategies adopted by the receiver. Before encoding, interleaving is implemented to separate spatially neighboring MBs into different packets. At the receiver's end, after decoding and deinterleaving one frame, the lost MBs are replaced

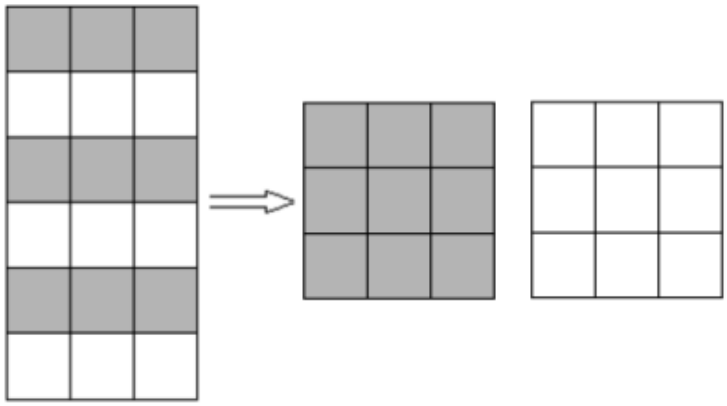
with matched patches from previous frames, and the down sampled data is restored to normal resolution through video up-sampling.

3.5.1 Interleaving

The interleaving process adopted in this work includes chessboard decomposition and row separation, as show in Figure 3.7. The chessboard decomposition separates horizontally or vertically connected MBs into two groups. The row separation in one group further divides the MBs in the odd-numbered and the even-numbered rows into two subgroups. These two steps can be repeatedly operated on one group of MBs until the desired data length is acquired. A successive implementation of these two steps ensures that no MBs in one group are directly connected (horizontally, vertically or diagonally) with each other in the original image. Further, after i times of repeating such implementation, no MBs in a 2^ix2^i neighborhood are in the same group. The receiver performs deinterleaving according to the information of encoder interleaving steps and the packet order. Figure 3.8 (b) shows a received decoded frame (intra coded, Q= 28) with five-step interleaving (resulting in 32 packets) and four packet loss. The luminance component of the original frame is shown in Figure 3.8 (a).



(a) Chessboard decomposition



(b) Row separation

Figure 3.7 : Interleaving.

(a) Original surveillance video
(luminance component)

(b) Packet loss
(4/32 packets lost, Q= 28)

(c) Boundary match
(PSNR= 36.45, SSIM= 0.96)

(d) Video up-sampling
(PSNR= 25.04, SSIM= 0.85)

Figure 3.8 : Error concealment.

3.5.2 Boundary Match

For a lost MB, its available boundary pixels are used to search for a concealment patch in previous frames. As explained by Equation (3.4), the correlation between the coordinates of two corresponding points in consecutive frames is utilized to determine the center of the search area. A patch yielding the smallest sum of absolute difference value in the search area is used to replace the missing MB, followed by a deblocking filter. The concealment result for the frame in Figure 3.8 (b) is displayed in Figure 3.8 (c). The visual quality is measured in PSNR (dB) and SSIM [89].

3.5.3 Video Up-sampling

Upon receiving a down-sampling coded frame, lost data is first recovered with the above described boundary match method. The reference frames are also sampled at the same resolution. Afterwards, the TV up-sampling method [69] is applied to restore the data to the normal resolution. Finally the up sampled data is compared with the reference frames using once again the coordinate correlation property to further refine the visual quality. The concealment result for a down-sampling coded frame in Figure 3.8 (a) can be observed in Figure 3.8 (d), with the same QP, interleaving and packet loss imposed.

3.6 Experimental Results

In the experiment, the PLR estimation method introduced in [87] models packet loss in video transmission over 802.11a WLAN networks, with a fixed packet length 8k bits.

$$p_{8kb}^{M} = \frac{1}{1+e^{\beta^{M}(\chi-\delta^{M})}} \tag{3.20}$$

Here χ is the signal to interference-plus-noise ratio (SINR) in dB. It is considered SNR when user interference noise is ignored. β^M and δ^M are constant parameters associated to each MCS denoted by *M*. Convolution coding is used as the FEC code. For variable packet size *L*, we adopt an approximate PLR calculation as shown in Equation (3.21). Four of the tested MCSs are listed in Table 3.1.

$$p^M = 1 - (1 - p_{8kb}^M)^{L/8000} \tag{3.21}$$

Table 3.1. Tested MCSs

Modulation, code rate	β^M (dB^{-1})	δ^M (dB)
MCS1 (64-QAM, 2/3)	0.625	18.2
MCS2 (16-QAM, 3/4)	0.352	15.1
MCS3 (QPSK, 1/2)	0.461	5.3
MCS4 (BPSK, 1/2)	0.640	2.3

3.6.1 Settings

The physical equipments for surveillance video capture include a PTU device (Directed Perception D47) and a PointGrey camera, both connected to a PC [60]. The video is recorded at 30 fps and is formatted to a 640x480 YUV420 sequence. The PTU camera motion is updated every two frames and is recorded for both the encoder distortion estimation and the decoder error concealment. The proposed video codec is developed based on the H.264/AVC standard [90]. The Y component is processed. The MB size is 16x16 for coding at normal resolution, and 8x8 for down-sampling coding. The available QPs include 16, 20, 28, 32, 34, 36 and 38. The maximum recursion step $\triangle$ in distortion estimation is set to two frames. Four MCS (*m*, *r*) configurations, namely MCS1 (6, 2/3), MCS2 (4, 3/4), MCS3 (2, 1/2), and MCS4 (1, 1/2), and the corresponding PLR models

are chosen from [87], as shown in Table 3.1. The frame delay constraint for video delivery is set to 1/30 second. In the transmission process, the first frame is fully intra coded and is assumed to be correctly received with extra protection (ARQ or stronger FEC). The average SNR $\bar{\gamma}$ in Equation (3.13) is set to 20dB, and the channel bandwidth W is set to 1MHz and 100kHz to simulate different channel conditions.

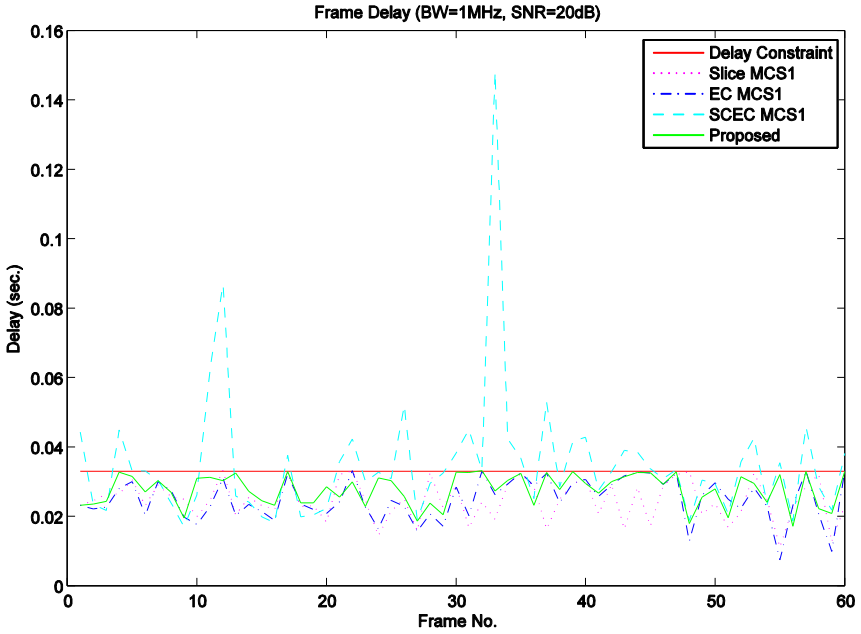
3.6.2 Performance

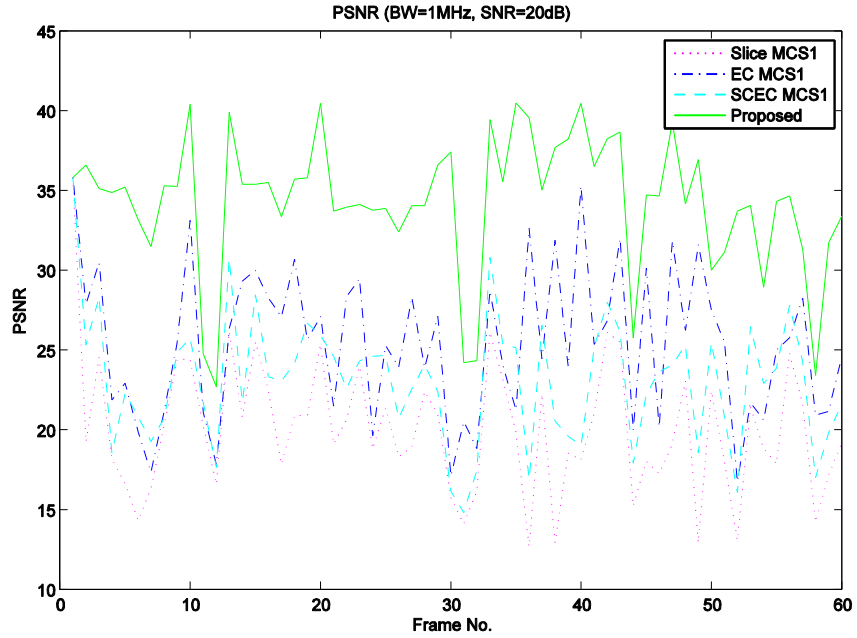
To test the efficiency of the proposed surveillance video coding and transmission scheme, three other types of coding methods with a fixed MCS are considered for performance comparison. The first one is the traditional intra+inter coding method, using slice copy as the error concealment (Slice MCS). The end-to-end distortion estimation process for the intra, inter and skip modes is applied. The second one is the down-sampling included coding method with the error concealment stated in Section 3.5 (EC MCS). The same distortion estimation process is also applied. The third one is similar to the EC MCS method, except that the distortion estimation only considers compression induced source coding distortion under a fixed data rate (SCEC MCS), i.e. using $\bar{\gamma}$ in Equation (3.13).

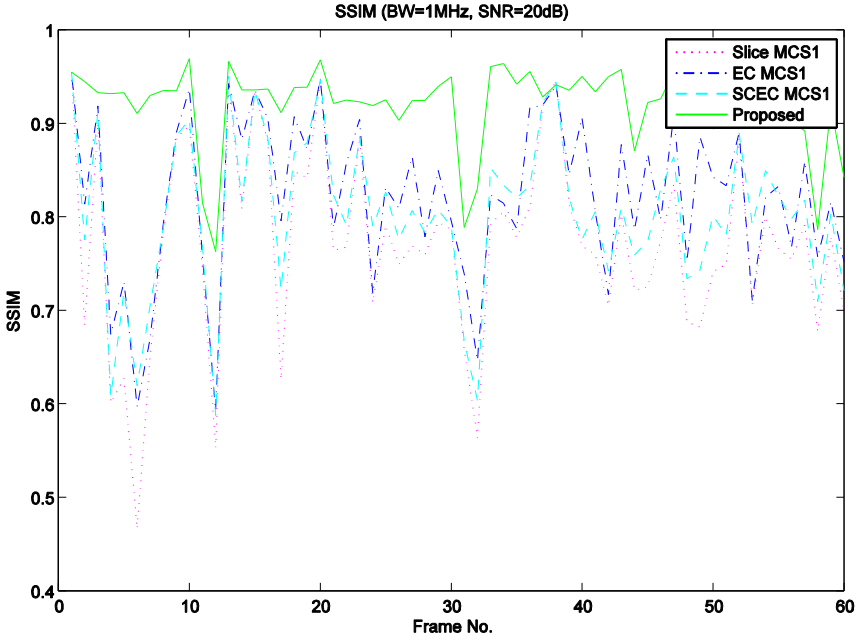
The delay (second), PSNR (dB) and SSIM performance of the three coding methods using MCS1 on sixty recorded frames are demonstrated in Figure 3.9. Under a better channel condition (W = 1MHz), the average PSNR/SSIM for Slice MCS1, EC MCS1, SCEC MCS1, and the proposed scheme is 20.52/0.77, 26.13/0.82, 23.48/0.80, and 34.32/0.92. Under a worse channel condition (W = 100kHz), it is 17.63/0.64, 19.68/0.70, 18.94/0.67, and 23.55/0.80, respectively. In both tests, the proposed scheme achieves the highest video quality. The experiment with the EC MCS1 method yields better video quality than the Slice MCS1 and the SCEC MCS1. This result demonstrates the

superiority of the adaptive coding, and the error concealment strategies. Lacking proper error concealment measures, the video quality with the Slice MCS1 method is the lowest.

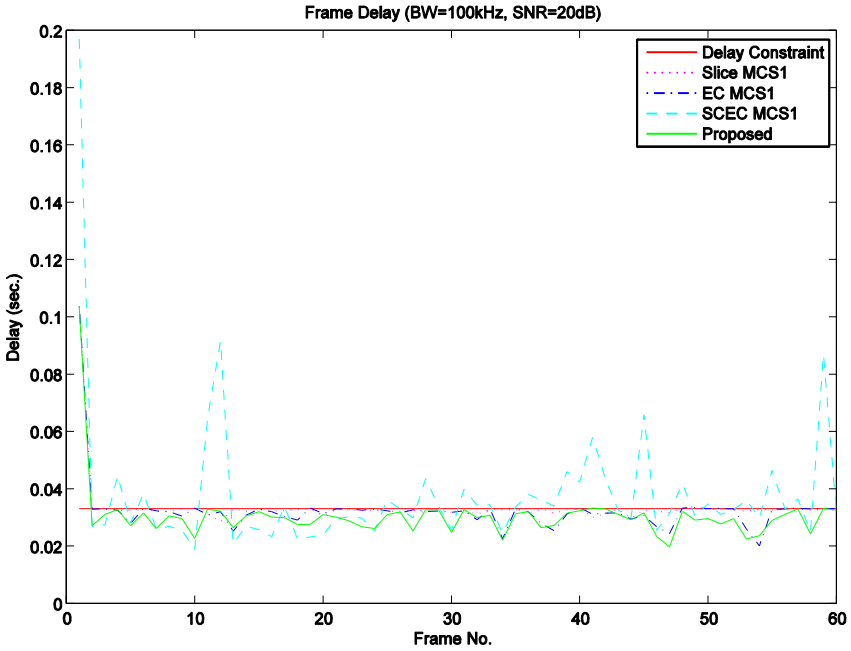
In both tests, the delay performance of the SCEC MCS1 scheme is poor due to the lack of channel information, while other methods provide proper response to the dynamic channel status and confines to the transmission delay constraint.

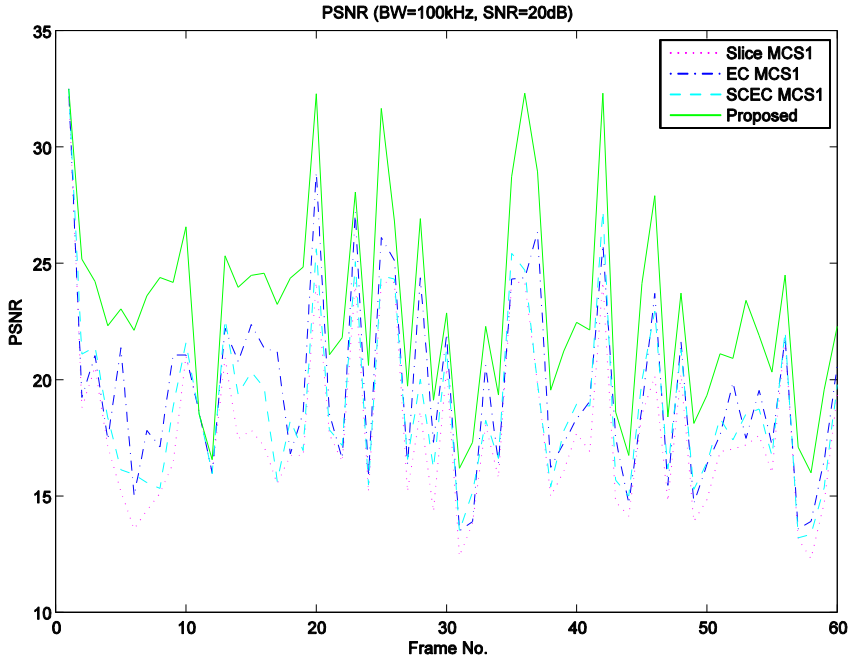


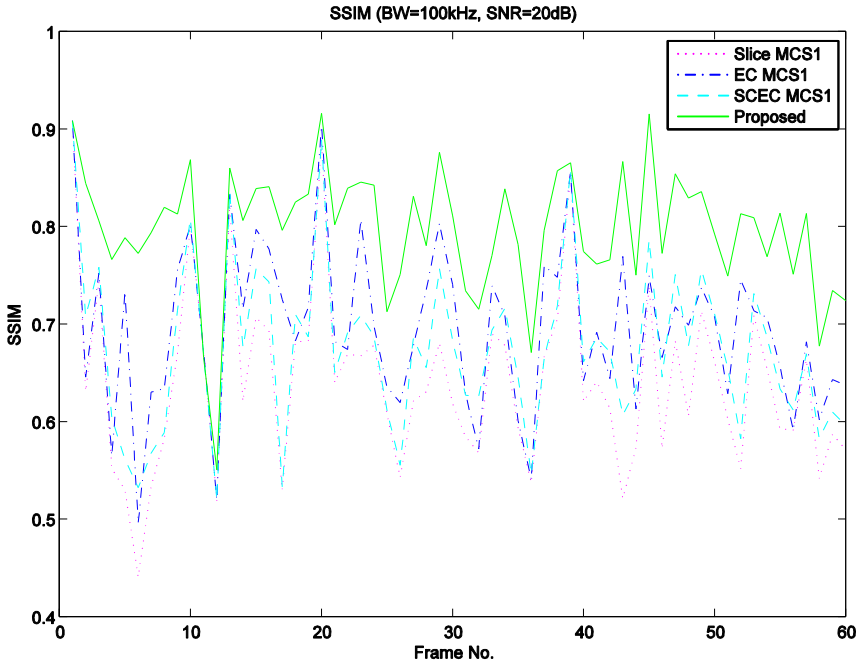




(a) W= 1MHz, $\bar{\gamma}$= 20dB


Frame Delay (BW=100kHz, SNR=20dB)
Delay Constraint
Slice MCS1
EC MCS1
SCEC MCS1
Proposed
Delay (sec.)
Frame No.
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
0.18
0.2
0
10
20
30
40
50
60


PSNR (BW=100kHz, SNR=20dB)
Slice MCS1
EC MCS1
SCEC MCS1
Proposed
PSNR
Frame No.
10
15
20
25
30
35
0
10
20
30
40
50
60



(b) W= 100kHz, $\bar{\gamma}$= 20dB

Figure 3.9 : Delay constrained video delivery.

The comparison of the Slice MCS method with different MCS on the average delay (millisecond, excludes the first frame), PSNR (dB), and SSIM can be observed from Table 3.2. For W= 1MHz, MCS3 is better than others while MCS1 is preferred for W= 100kHz. Thus the adaptability of the proposed scheme is desirable under the dynamic channel condition. The percentage (%) of the intra and inter coded packets for the proposed scheme is provided in Table 3.2. With lower bandwidth, a higher portion (88%) of the packets is encoded in down-sampling mode to accommodate the deteriorated channel condition.

Table 3.2. Performance comparison (Slice MCS vs. Proposed)

		Delay	PSNR	SSIM	Intra	Inter
1MHz	MCS1	25.5	20.52	0.7722	95	5
	MCS2	27.9	24.44	0.8382	100	0
	MCS3	27.6	28.78	0.8940	97	3
	MCS4	31.9	22.76	0.8165	70	30
	Proposed	27.7	34.32	0.9189	70	20
100kHz	MCS1	32.3	17.63	0.6437	83	17
	MCS2	31.7	16.02	0.5813	70	30
	MCS3	18.0	15.35	0.4720	70	30
	MCS4	8.4	13.24	0.4472	100	0
	Proposed	28.8	23.55	0.7967	12	0

3.7 Summary

This Section presented a system design for wireless video surveillance with a single PTU camera, including video capture with automatic camera control, data compression and transmission with cross-layer control, and error concealment. The video coding process is formulated as a delay constrained distortion minimization problem with consideration to configurable system parameters. Multiple error concealment strategies are adopted to counteract packet loss. The property of the PTU camera motion is also exploited to accelerate the concealment process. The experimental results are promising for a real-time surveillance application. Part of this work is published in [60, 62, 63, 91, 92].

Chapter 4

Binocular Video Object Tracking and 3D Video Transcoding

4.1 Introduction

Traditional monocular tracking methods mainly explore the temporal correlation in one video to detect moving areas [59, 75, 93, 94]. With the development of the 3D signal processing technologies, multiview video object tracking is gaining increasing interest. In the tracking algorithm presented in [95], both inter-frame and inter-view correlations are utilized to predict the object's position and speed, using optical flow and disparity estimation. The outdoor tracking algorithm presented in [96] performed ground view alignment using homography; each moving object is detected via background subtraction. These works use fixed cameras and may have limitation in view scope. The post-capture tracking method described in [97] handles the changing camera viewpoint by constructing a panoramic image used for background registration and object detection.

In the video tracking and streaming system introduced in [98], active cameras are adopted for runtime operation. A master camera is manually controlled and other slave cameras will automatically follow. For object tracking with moving cameras, a more practical strategy is to use automatic PTU/PTZ cameras, where the camera projection center is generally unchanged and the retinal plane is capable of angular movement. In this kind of system, the camera control algorithm for the tracking process needs to estimate the angular speed/acceleration of the moving object, and the background alignment in different video frames is required for motion detection. In the PTU camera tracking algorithm proposed by Petrov et al. [59], a linear feedback controller is applied based on the Theory of Lyapunov Stability. The control parameters are updated by object

position estimated using the Mean Shift [93] algorithm. Mean Shift is efficient in locating object position according to the object's color distribution [97], [98]. However, a key problem with this method is the scaling of the tracking area, since the size of the object appears differently as its depth changes. A tentative scheme is suggested in [93] to adjust the tracking region according to the similarity measure. It might have problem if similar color is present around the boundary of the tracking area. The object segmentation based approach is more robust, and time consuming [75]. Yang et al. [94] developed an updating rule for scaling factor by comparing second-order moments between the template and the target, but only small tracking region is tested.

To enable tracking in an unmanned environment, we use a PTU device (Directed Perception D47) to perform camera control. The master camera is able to rotate and its projection center stays unchanged [59]. A slave camera is placed on the flank, and moves along with the master camera, as shown in Figure 4.1. The Mean Shift algorithm [93] is adopted in our project for real-time tracking. With binocular video output, we consider the object depth/disparity as a natural and reliable resource for adjusting the tracking window, since disparity contains object position information and it is necessary for multi-view video streaming [98].

Figure 4.1: Binocular PTU cameras.

During the 3D data streaming process, we consider the requirement of transcoding by different display devices in a heterogeneous wireless network. Since different types of 3D displays can render 3D video, there are many types of formats to realize 3D video [99]. The principle of 3D video is based on the binocular vision fused by the signals of both eyes. Therefore, the simplest format for 3D video is the stereoscopic video which contains two captured views. This kind of format can provide the 3D perception, but it cannot provide the parallax-adjustable 3D effects [100]. Comparably, as an alternative of stereoscopic video, video plus depth representation can provide the parallax-adjustable 3D perception in a limited range. Due to its simplicity of compression, video plus depth based stereoscopic video can be easily used for mobile 3D video applications [101].

Currently, 3D video is mainly aiming at the home application with high-definition (HD) formats. The HD 3D contents are mostly delivered through the ways of terrestrial broadcast, cable, satellite and IPTV [102]. Though HD 3D videos provide the vivid visual

effects, it requires much more transmission bandwidth. With the high-rate Internet, the HD 3D video is also possible to be transmitted. However, with the advancements of mobile communication technologies, Internet networks are mostly the heterogeneous networks which consist of the wired and wireless networks. The heterogeneous networks usually provide the different transmission qualities. Since the HD 3D video has been compressed and distributed for home application, the solution of scalable video coding is not appropriate for the 3D video transmission over the heterogeneous network. To deploy the 3D video service on the mobile devices, the rate reduction and down-sample transcoding must be considered to adapt to the wireless channel and mobile device. Figure 4.2 shows the rate reduction transcoding application for mobile 3D video streaming. After being encoded at the media server, the captured video and generated depth data are firstly streamed through the wire-line Internet. When the receiver is mobile user, the high bit-rate HD 3D video data is transcoded to the low bit-rate mobile 3D video data at the transcoding gateway and is then streamed to the mobile receiver.

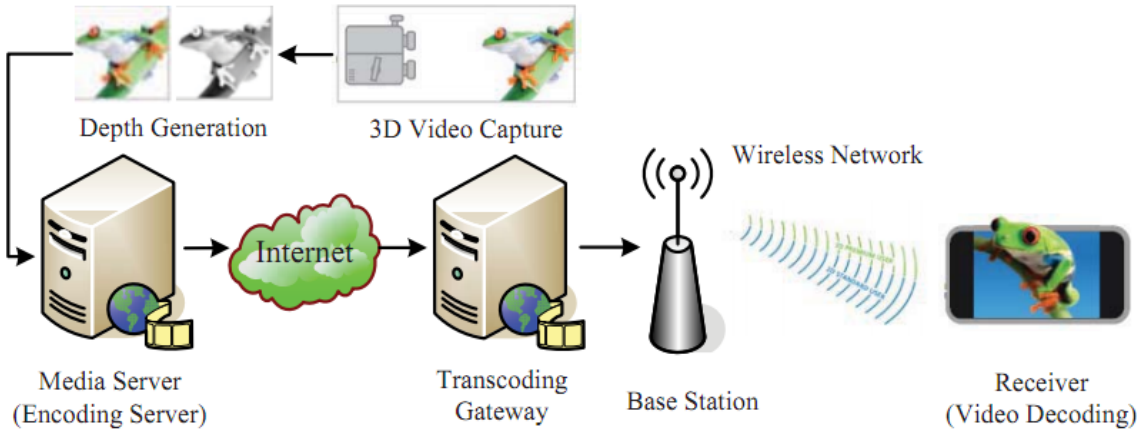


Figure 4.2 : The mobile 3D video transcoding application.

Based on the 3D video content generated from the binocular object tracking procedure, we proposed a dynamic rate allocation scheme for the 3D video transcoding over the wireless channel, through the cross-layer controller. The rest of this section is organized as follows. Section 4.2 introduces the binocular Mean Shift tracking using the disparity information. Section 4.3 describes the fast disparity estimation procedure. Section 4.4 provides the object tracking results using the proposed disparity estimation algorithm. Section 4.5 introduces the cross-layer control procedure for the video/depth 3D data rate allocation according to the dynamic channel condition. Experimental results are provided in Section 4.6, and Section 4.7 draws the conclusions.

4.2 Binocular PTU Camera Tracking

In this subsection, we present a binocular PTU camera video object tracking scheme using the Mean Shift algorithm and the runtime disparity estimation. The proposed method is to accommodate the requirement of 3D content generation and accurate tracking in more advanced video surveillance applications. The disparity estimation process for each stereoscopic pair is formulated as an energy minimization problem. The iterative solution procedure is implemented in a course-to-fine manner. The estimated disparity is used to scale the tracking window by the Mean Shift algorithm, i.e. the size of the tracking area is adjustable according to its inner disparity, and thus the moving object can be better located by the camera. The program maintains the semi-real-time performance and acceptable accuracy as evaluated on a set of standard test data. In our experiment, two PointGrey cameras are controlled through a PTU device. The disparity estimation process on the recorded tracking video (640x480) achieves 6fps on an ordinary PC (2.66GHz CPU, 4GB RAM).

The video object tracking procedure is illustrated in Figure 4.3. The estimated disparity values from the stereoscopic image pair are used to adjust the size of the tracking region. The object center is detected using the Mean Shift algorithm, and it is provided for PTU control [59] on the master camera.

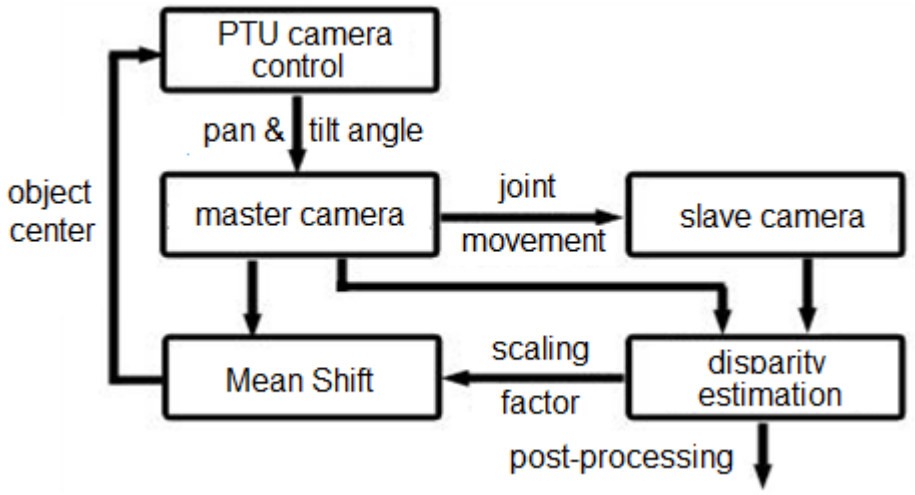


Figure 4.3 : Tracking procedure.

The unique geometric property of a PTU camera model is that the camera projection center remains unchanged while the pan and tilt angles are controllable, as illustrated in Figure 3.2 (a). The focus F denotes the projection center. The image plane is viewed down along its y axis, and is projected on the X-Y world coordinate plane. α is the angle between the object center and the X axis, θ is the angle between the image center and the X axis, f is the focal length, and x_c is the distance between the projected object center and the image center along the x axis. Only pan control is displayed in the figure. The algorithm applies to tilt control similarly.

The linear feedback controller aims to minimize x_c and the difference between the estimated object speed and the measured object speed. According to the Theory of Lyapunov Stability, the camera angular speed $w_{\theta k}$, the camera angle θ, and the estimated distance x_c are updated at every time instance according to Equations (3.1) ~ (3.3).

Once the control parameters are updated, the disparity information for the master camera side is estimated using the method described in Section 4.3. The camera projection matrix used for the stereo rectification is obtained beforehand through chessboard calibration. After the control parameter update, the Mean Shift algorithm is applied to locate the object center in the new frame. The size of the rectangular tracking window is scaled according to

$$l1/l2 = z2/z1 = d1/d2 \tag{4.1}$$

where $l1$, $l2$ denote the edge length of the tracking window at two consecutive updates, z is the depth of the object, and d is the average estimated disparity for the object region, as shown in Figure 4.4.

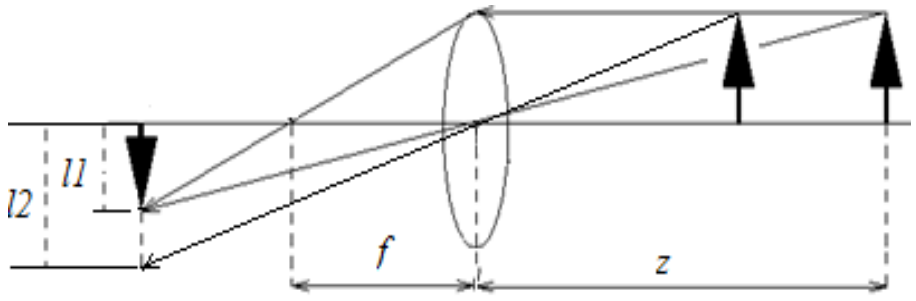


Figure 4.4 : Binocular Mean Shift tracking with window size adjustment.

4.3 Fast Disparity Estimation

The implementation of the disparity estimation process is essential for the real-time object tracking. Stereo matching/disparity estimation has been extensively studied as a fundamental vision task. Popular solutions include the local winner-take-all [103-105] and the global MRF (Markov Random Field) optimization [106-108]. Local methods compare matching cost computed within a neighborhood. They are known for their fast implementation, but have difficulty in dealing with ambiguous or similar textures. Global methods are capable of imposing smoothness constraint, such as graph cut [106] and belief propagation [107]. Their occlusion detection ability is impeded in the presence of curved surfaces and the computation is usually very expensive. Smith et al. [108] proposed to perform graph cut optimization on a sparse graph obtained using maximum spanning tree. Local filtering is applied at the finer grid for further refinement. Although this method better detects non-planar surfaces, the process of spanning tree generation and full image filtering are still costly to implement. For our tracking procedure, disparity estimation has to meet the runtime requirement in order to timely adjust the tracking region. While some real-time approach [103] relies on GPU implementation, Geiger et al. [104] introduced an efficient matching method based on Delaunay triangulation [109]. This method exhibits superior results in less textured areas, and semi-real-time performance is reported. However, the initial supporting points acquired using the local method fall short of resolving spurious matches caused by similar textures, which are commonly encountered in an indoor surveillance environment with reduplicate building structures or wall decorations. This will in turn result in incorrect estimation on the finer grid. Moreover, the disparity values on the finer grid are also estimated using the local

method, thus the estimation changes of the neighbors could not be further utilized to estimate the disparity of the current pixel.

To overcome these shortages, we propose to implement global optimization on the initial supporting points. A best disparity for each initial supporting point is selected from multiple candidates through Iterative Conditional Mode (ICM) [110]. The disparity estimation process on the finer grid is formulated as an energy minimization problem. Both the data consistency term and the smoothness term constrain the iterative solution procedure. The program still maintains semi-real-time performance and acceptable quality.

4.3.1 Problem Formulation

According to Bayes' rule, the process of disparity estimation can be formulated as a MAP-MRF problem [107]. For example, in binocular stereo matching, the estimation on the left image from a stereoscopic pair I_1 (left), I_2 (right), is usually considered as a process of minimizing the following energy function,

$$\begin{aligned} E\big(d(x,y)\big) &= \iint_{\Omega} F(d)\, dx\, dy \\ &= \iint_{\Omega} \big(\lambda_1 \cdot f_{data}(d) + \lambda_2 \cdot f_{smooth}(d)\big)\, dx\, dy \end{aligned} \tag{4.2}$$

where $d(x,y)$ is the estimated disparity at pixel (x,y). λ_1 and λ_2 are the scaling factors. f_{data} and f_{smooth} are distance measures and represent the penalty terms on photo consistency and smoothness. For horizontally rectified images,

$$f_{data}\big(d(x,y)\big) = \frac{1}{2}\sum_{(x_k,y_k)\in\aleph_1((x,y))}\big(I_1(x_k,y_k) - I_2(x_k-d,y_k)\big)^2 \tag{4.3}$$

$$f_{smooth}(d) = \frac{1}{2}|\nabla d|^2 \ , \tag{4.4}$$

where ∇ is the gradient, and $\aleph$ denotes the neighboring pixels. Using the calculus of variations [111], the minimum of Equation (4.2) can be obtained by solving its Euler-Lagrange equation:

$$\frac{\partial F}{\partial d} = \lambda_1 \cdot \sum_{(x_k,y_k)\epsilon\aleph_1((x,y))} I_{2x}(x_k - d, y_k)\big(I_1(x_k, y_k) - I_2(x_k - d, y_k)\big) - \lambda_2 \cdot \Delta d = 0 \quad , \tag{4.5}$$

where I_{2x} is the derivative of the feature response in I_2, and Δ is the Laplacian operator. The solution procedure is implemented in an iterative manner. The initial value of d is essential to the convergence speed. Kosov et al. [112] adopted a multi-grid strategy. The disparity values are estimated at a lower resolution, and are refined at a higher resolution with a feature-adaptive full approximation scheme. The estimation at the coarser grid provides a good initial guess for the iterative refinement at the finer grid, but the computation is still high for the full coarse grid estimation. Geiger et al. [104] perform Delaunay triangulation interpolation on a set of detected feature points to achieve fast implementation. The idea is that given the disparity values at a set of sparse supporting points, triangulation on these points can segment the image into small triangular regions, and the disparity of a point inside each region can be approximated through interpolation by disparity values of its three vertices. This method is very efficient for obtaining initial disparities. A disadvantage is that the erroneous detected vertices using the local method result in false interpolation. Thus we apply the global ICM on the detected supporting points. The iterative solution procedure is performed on the normal grid with the initial

values obtained from the triangulation results. The details of the multi-resolution strategy for disparity estimation are provided in the following subsection.

4.3.2 Multi-resolution Strategy

To accelerate the disparity estimation process, we adopt a coarse-to-fine strategy to reduce the computation overhead. A sparse grid is obtained with robust left-right image feature point correspondence, using both local search for photo consistency, and global ICM match. An initial estimation is provided through performing Delaunay Triangulation interpolation on the sparse grid, and is then refined using local search on the finer grid. To prepare candidate options for the ICM implementation, multiple disparities d_1, d_2, …, d_n, with the highest photo consistency are selected for each textured supporting point on a sparse grid.

$$\begin{aligned} & (d_1, d_2, \dots, d_n) = arg\ min(f_{data}(d)) \\ s.t.\ \ & f_{data}(d_1) \le f_{data}(d_2) \le \cdots \le f_{data}(d_n), \\ & \text{and} \qquad O(d_1) = \text{false} \end{aligned} \tag{4.6}$$

Occlusion is detected with the thresholding method.

$$O(d) = \begin{cases} \text{false}, & \text{if } f_{data} \le T_1 \\ \text{true}, & \text{otherwise} \end{cases} \tag{4.7}$$

Here T_1 is the thresholding parameter. The best option d is chosen from these n ($n = 2$ in our experiment) candidates according to

$$d = arg\ \ min_{(d_1, d_2, \dots, d_n)} F(d). \tag{4.8}$$

If a supporting point could not yield any valid match as described in Formula (4.6), it will be eliminated from the set. After several iterations of ICM operation according to Equation (4.8), most spurious matches due to the presence of similar textures could be corrected, and a sufficient number of supporting points are obtained. The result can be observed from Figure 4.5 for the test data *Aloe* (1282x1110) [113]. The mismatch rate is reduced from 2.92% to 2.46% (two-pixel error threshold on non-occluded areas). And the extra computation time for two ICM iterations is negligible.

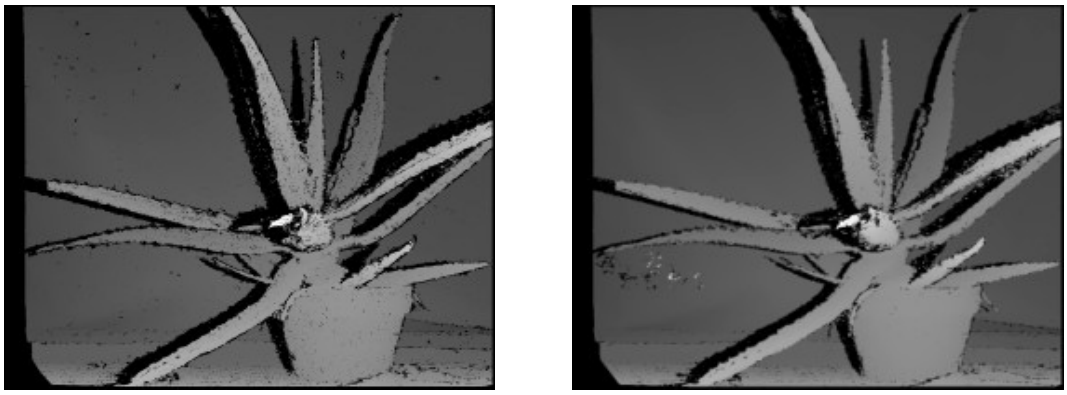


Figure 3.5 : Disparity estimation for Aloe. The results using Geiger et al.'s method [104] with (left) and without (right) ICM operation for selecting the supporting points.

The idea of Delaunay triangulation interpolation is that given disparity values at a set of sparse support points, triangulation on these points can segment the image into small triangular regions, and the disparity of a point inside each region can be approximated through interpolation by disparity values of its three vertexes [104]. As shown in Figure 3.6, given the estimated disparities of the supporting points S_1, S_3, S_4, the triangulated plane $S_1S_3S_4$ provides interpolated disparity values for the pixels inside the plane, such as the one denoted in green. Thus an initial guess for the whole image's disparity data can

be calculated from the sparse set of the supporting points processed by the method described above.

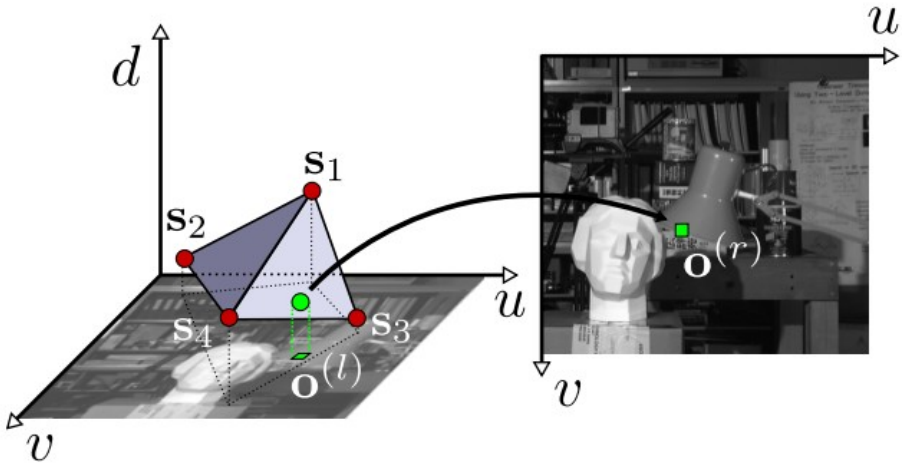


Figure 3.6 : Delaunay Triangulation interpolation [104].

After obtaining the initial disparity values by applying Delaunay triangulation interpolation on the supporting points, the result is further refined using Equation (4.5). The discretization form of the equation is

$$\lambda_1 \cdot \sum_{(x_k,y_k)\epsilon\aleph_1((x,y))} \big(I_2(x_k-d,y_k)-I_2(x_k-1-d,y_k)\big)\big(I_1(x_k,y_k)-I_2(x_k-d,y_k)\big) \\ -\lambda_2 \cdot \textstyle\sum_{d_j\epsilon\aleph_2(d)}\big(d_j-d\big)=0 \tag{4.9}$$

By applying the Gauss-Seidel method [114], above linear equation can be solved iteratively. At the (t+1)-th iteration, the disparity of pixel i is updated as

$$d_i^{t+1}=\frac{\lambda_1}{|\aleph_2|}\cdot \\ \sum_{(x_k,y_k)\epsilon\aleph_1((x,y))}\Big(I_2(x_k-1-d_i^t,y_k)-I_2(x_k-d_i^t,y_k)\Big)\Big(I_1(x_k,y_k)-I_2(x_k-d_i^t,y_k)\Big)$$

$$+\frac{\lambda_2}{|\aleph_2|}\cdot\left(\sum_{d_j\epsilon\aleph_2^{-}(d_i)} d_j^{t+1}+\sum_{d_j\epsilon\aleph_2^{+}(d_i)} d_j^{t}\right) \tag{4.10}$$

where $\aleph_2^{-}(d_i)$, $\aleph_2^{+}(d_i)$ are the neighboring pixels processed before and after pixel *i*. In practical implementation, two situations are considered. In textured areas (detected with Sobel operator), a local search is applied according to the value computed from the second part $\lambda_2 \cdot (*)$. The disparity with the minimum data consistency penalty is used as the update. In non-textured areas, the value of the first part $\lambda_1 \cdot (*)$ from the data consistency penalty is small enough and is hence ignorable. λ_2 is set to 1 to impose the influence from the neighboring pixels. The evolution process stops when the maximum number of iterations is reached, or the change of the disparity values falls below a threshold. The iterative process is bound to converge since $|\aleph_2| > 1$.

4.3.3 3D Content Generation

To verify the efficiency of the proposed disparity estimation method, several stereoscopic image pairs from the Middlebury dataset [113] are tested on the matching accuracy and the processing time. The object tracking results using the disparity-based window scaling Mean Shift algorithm are provided in Section 4.4.

The supporting feature points used for triangulation interpolation are selected from a sparse grid on the tested images. Only intensity data are processed. Two types of grids with different cell size are tested in the experiment, the 8x8 cell size, and the 16x16 cell size. The calculated mismatch rate (M.R.) and the number of selected supporting points are listed in Table 4.1, with two-pixel error threshold on all non-boundary areas. Two iterations are performed on the interpolated initial estimation. The disparity estimation

results can be observed from Figure 4.7. The occluded areas are interpolated using the results from the neighboring pixels.

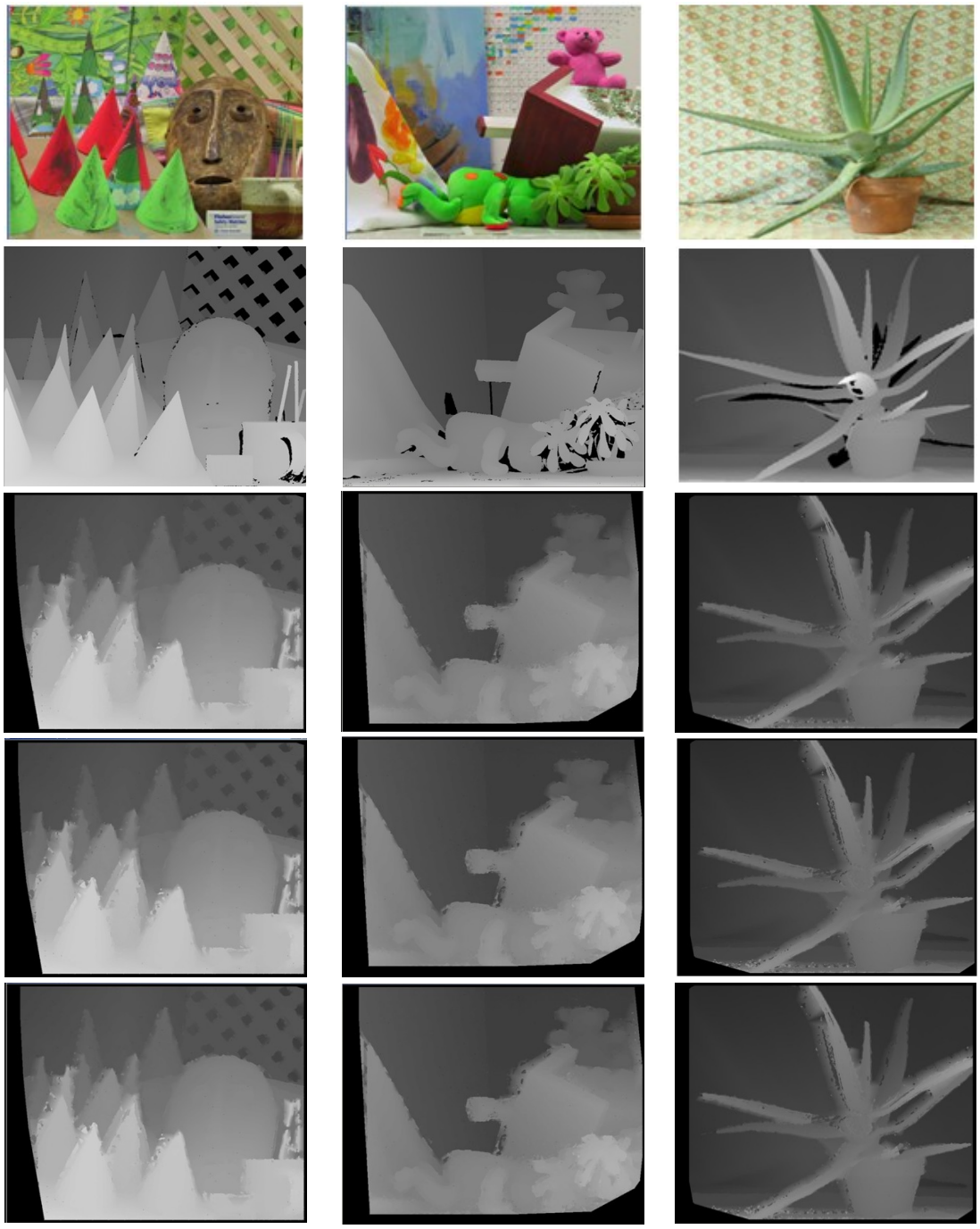
Table 4.1 Mismatch rate (%) and the number of supporting points.

	Cones (900x750)		*Teddy* (900x750)		*Aloe* (1282x1110)	
	M.R.	Points	M.R.	Points	M.R.	Points
8x8	6.2	1497	7.6	1247	9.9	2503
16x16	6.7	541	7.8	405	10.1	864

The processing time on *Cones* for different phases of the estimation process is listed in Table 4.2. The mismatch rate reduction is displayed in Figure 4.8. Most of the disparity change occurs during the first two iterations. Note that the multi-grid algorithm by Kosov et al. takes up 300ms to 1300ms to process the same data at half the resolution (450x375), with similar computational resources, as reported in [113]. Compared to Geiger et al.'s method [104], the extra processing time concerns the iterative evolution at the finer grid. The average processing time is 52ms per iteration, and the average mismatch rate reduction is 0.5%.

Table 4.2 Processing time (ms) for different phases: computing supporting points, triangulation interpolation, 1st iteration, 2nd iteration, 3rd iteration, 4th iteration, and 5th iteration.

	Supp.	Tri.	Iter.1	Iter.2	Iter.3	Iter.4	Iter.5
8x8	489	158	57	52	52	52	52
16x16	383	124	62	53	52	52	53



(a) Cones (900x750) (b) Teddy (900x750) (c) Aloe (1282x1110)

Figure 4.7 : Disparity estimation. From the first row to the last row: the left image, ground truth, initial estimation, 1st iteration, 2nd iteration.

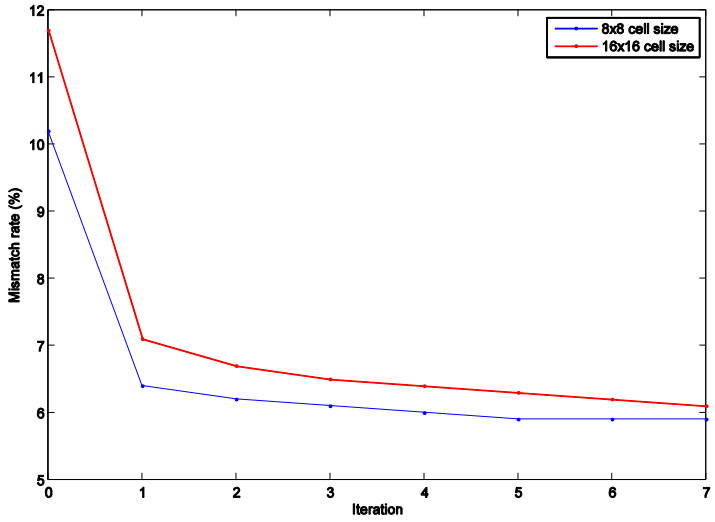
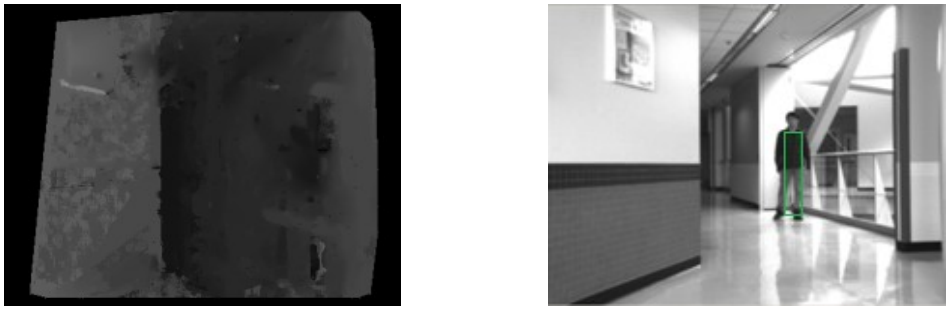


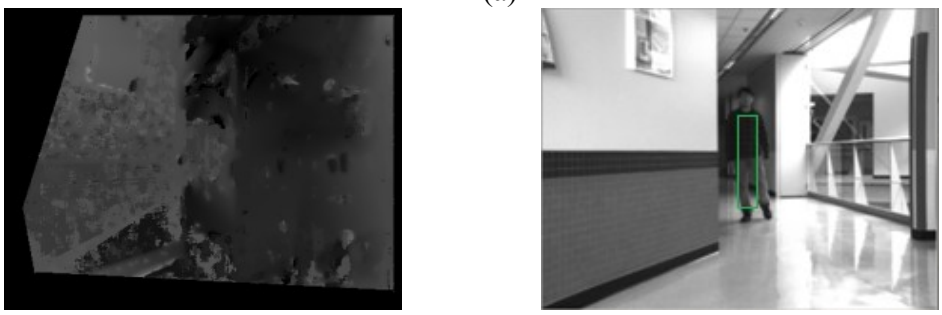
Figure 4.8 : Mismatch rate reduction.

4.4 Object Tracking

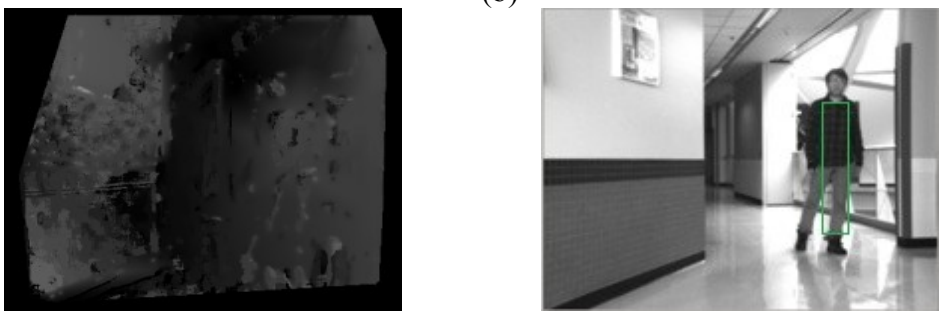
In the object tracking process, two PointGrey Firefly MV CMOS cameras are placed on the PTU device, and are connected to a desktop via a 1394 firewire USB2.0 hub. The 640x480 video is recorded at a frame rate of 15 fps. The average processing time for the disparity estimation is 178 ms per frame. An initial tracking window is obtained through user input. The tracking window is rescaled every 3 frames. The disparity estimation and the tracking results are provided in Figure 4.9. In the tracking process, the object walked along the corridor inside our department building. The environment contains both textured (mosaic tiles) and non-textured (wall, floor, pillar) materials. When applying traditional Mean Shift tracking using fixed window size, the camera lost track of the object easily when the object approached areas with similar colors, such as the situation in Figure 4.9 (b), when the color distribution inside the old tracking window could not fully represent the original target model.



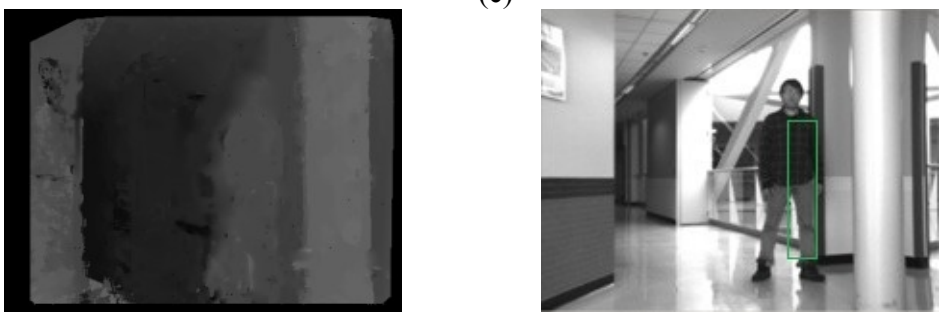
(a)



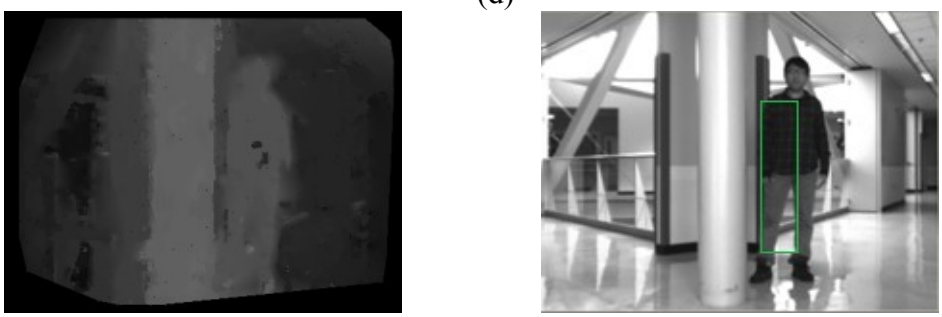
(b)



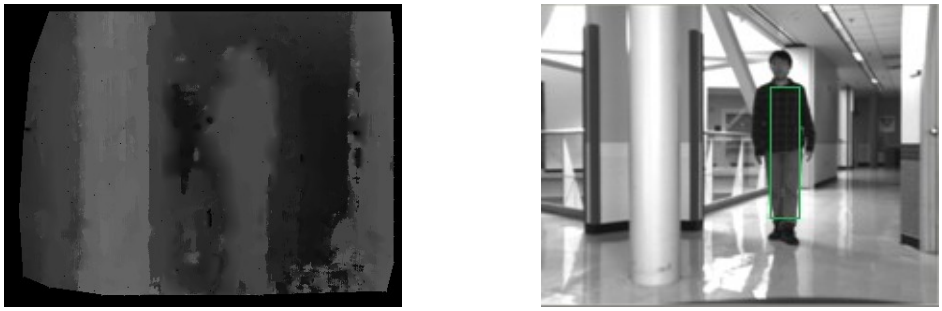
(c)



(d)



(e)



(f)

Figure 4.9 : Video object tracking.

4.5 3D Video Transcoding

The binocular video object tracking procedure generates the video plus depth 3D video data. To deliver the 3D content to different display terminals in a heterogeneous wireless network, we design a cross-layer optimization framework for 3D video transcoding. In the past years, a lot of transcoding works were proposed, but there were little research works on 3D video applications. Liu et al. [115] first proposed an efficient 3D video transcoding scheme for the virtual view. However, it did not aim at the error-resilient transcoding. As we all known, the wireless network is error-prone due to the non-static end-users, and the 3D video transcoding should correspondingly provides a certain degree of the error-resilience ability [116]. The bandwidth resources of wireless channel are limited so that the rate adaptation to the channel is very necessary. Because the video plus depth based 3D video usually generates a pair of views consisting of one captured view and one synthesized virtual view, the virtual view quality has a strong dependency on the depth and video fidelities of the reference view. The video/depth rate allocation [117] can control the fidelities of video and depth of reference views, and correspondingly has an influence on the virtual view quality. The video/depth rate allocation for 3D video coding is originally optimized for the wire-line transmission, and

it does not consider the packet loss effect on the virtual view quality so that it is not appropriate for the new wireless transmission. Hence, the new error-resilient rate allocation between video and depth should be taken into account in the 3D video transcoding. In the transcoding proxy, the original signals cannot be accessed to compute the distortions of video and depth at different bit-rates, and consequently the optimal rate allocation between video and depth cannot be evaluated if not given the original video and depth data. Moreover, if we do the exhaustive seeking of the optimal rate allocation between video and depth at the transcoding proxy, the transcoding proxy will endure such computational complexity as to affect the speed of transcoding. To solve the problem of the absence of original signal, and in the mean time to avoid the additional complexity, a look-up table method is proposed to realize the mobile 3D video transcoding. In this section, we propose an error-resilient transcoding framework for mobile 3D video streaming, which transcodes the HD 3D video stream to the mobile 3D video stream. The proposed transcoding framework builds a rate allocation table at the streaming server side and then transmits it to the transcoding proxy. The encoding server first encodes the video plus depth based HD 3D video, and then down-sample the decoded video and depth. By utilizing the exhaustive full-searching method, the down-sampled video and depth will be re-encoded with different QP pairs for computing the virtual view distortions under different packet loss rates (PLRs) to generate the Rate-QP-PLR table. According to the actual PLR returned from the wireless channel, the transcoding proxy converts the HD 3D video stream to mobile 3D video stream by looking up the Rate-QP-PLR table. Detail implementation of the cross-layer control mechanism is explained in following subsections.

4.5.1 Framework

To provide a rate-adaptive and error-resilient transmission over the heterogeneous network, we propose a 3D video transcoding framework, which consists of the encoding server and the transcoder. The video encoding server and the transcoder are originally independent components of the video streaming system. Currently, we integrate the encoding with transcoding to transfer a part of transcoding computations to the encoding server. Especially, the heavy computation for video/depth rate re-allocation originally needed to be performed at the transcoder can be moved to the high performance encoding server. Moreover, the problem of accessing the original signals at the transcoder also disappears by moving the video/depth rate allocation to the encoding server.

Figure 4.10 shows the flowchart of the transcoding framework. The encoding server compresses the HD 3D contents, and generates the Rate-QP-PLR table for assisting the transcoding. Generally, the mobile display size is small, and the former spatial resolution does not adapt to the mobile device. The video and depth need to be appropriately down-sampled. The server reconstructs the video and depth from the former streams and then down-samples them. After that, the server re-encodes the down-sampled video and depth data multiple times to perform the video/depth rate allocation and generates the Rate-QP-PLR table. The Rate-QP-PLR table contains the specific information of the video QP and depth QP under different rate and PLR levels. Since the packet loss effects have been considered in the video/depth rate allocation, the Rate-QP-PLR table can be used to guide the error-resilient transcoding. Via looking up the Rate-QP-PLR table, the transcoder performs the cross-layer control, collects the PLR and channel information returned from the actual transmission channel, and transcodes the compressed video and depth streams

with the appropriate QP pairs according to the Rate-QP-PLR table. Though the encoding server does not know the actual channel behaviors in advance, the random packet loss previously simulated in the encoding server can reflect the actual influence of the transmission error on the video quality in the statistical sense.

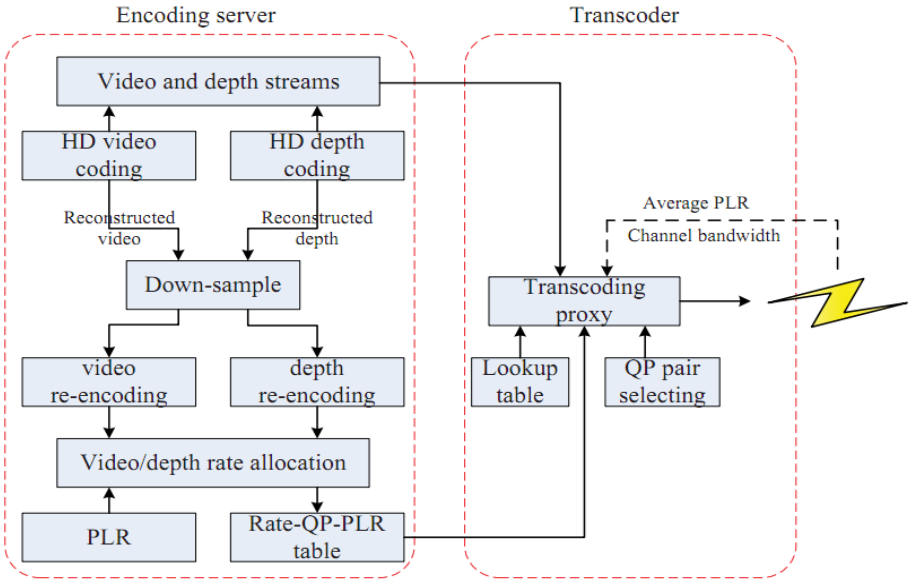


Figure 4.10 : The proposed transcoding framework.

4.5.2 Cross-layer Control for Video/Depth Rate Allocation

During the re-encoding process, the encoding server also considers the effect of different intra refreshing rate (IRR). Intra coding can eliminate the temporal error propagation by restraining packet loss induced prediction drifting error. Cyclic intra refresh coding is to split the image into N equal-sized regions and intra-code all macroblocks (MBs) of one region in every inter-coded frame. Here, 1/N is the IRR. The intra refresh cross N frames will generate the equal error-resilience effect of inserting one intra frame among N frames. Increasing the number of intra-coded MBs in inter-coded frames can strengthen

the error-resilience to packet loss. However, intra-mode coding will consume much more bit-rate than inter-mode coding so that there is a trade-off between IRR and the coding efficiency under given packet loss rate. And also the bit-rate increasing or decreasing will change the packet loss possibility for a given packet under the constant bit error rate (BER). When one slice packs into one package for transmitting, intra refresh coding changes the coding bits for a slice and the estimated PLR will also be changed since it is related to the BER. Hence, the PLR and IRR need to be jointly considered to guarantee the optimal rate distortion performance. For video plus depth based 3D video, Figure 4.11 shows the 3D rate-distortion performances under different PLRs and different IRRs for the *Book_Arrival* sequence.

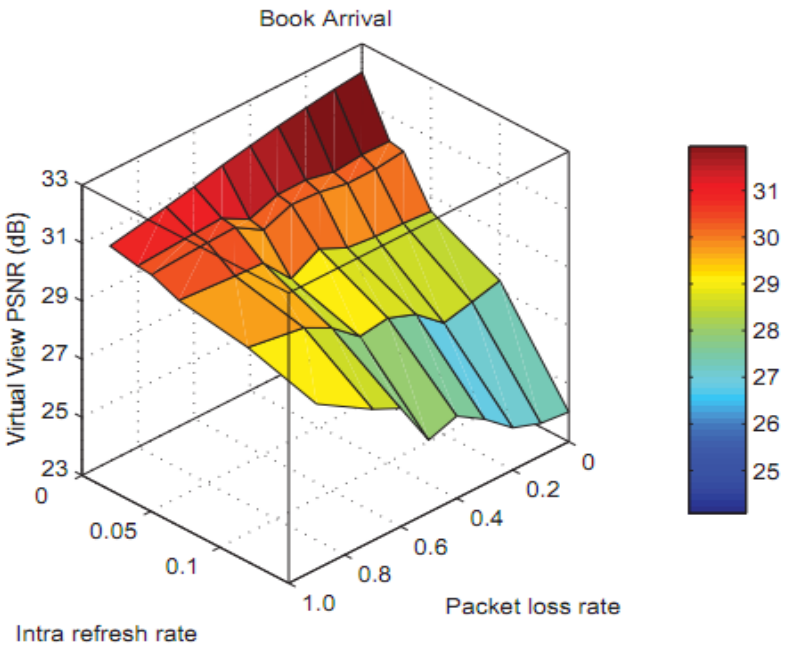


Figure 4.11 : 3D rate-distortion performance at the total rate (video plus depth) of 1Mbps with 600kbps for video and 400kbps for depth.

Since the IRR has a strong relation with the PLR, we preset several levels of the intra refresh to adapt to the actual PLR returned from the channel. The IRR can be adaptively set for the error-resilient video/depth rate allocation. Therefore, the transcoder performs cross-layer control to dynamically select the optimal QP and IRR for the re-encoding process by the encoder server. The formulation of the decision making procedure can be expressed as follows.

$$\left(q_{2,v}^{opt}, q_{2,d}^{opt}, \eta_v^{opt}, \eta_d^{opt}\right) = \underset{q_{2,v}, q_{2,d} \in Q, \eta_v, \eta_d \in \xi}{\arg\min} D_S\left(q_{2,v} | \left(q_{1,v}, \rho_v, \eta_v\right), q_{2,d} | \left(q_{1,d}, \rho_d, \eta_d\right)\right)$$

$$s.t.\ R_v\left(q_{2,v}, \eta_v\right) + R_d\left(q_{2,d}, \eta_d\right) \leq R_c \tag{4.11}$$

The symbols $q_{1,v}$, $q_{1,d}$ and $q_{2,v}$, $q_{2,d}$ denote the QPs applied for encoding the video data and the depth data, before and after the transcoding procedure begins. ρ and η are the corresponding PLR and IRR for current slice. D_S denotes the view synthesis distortion at the receiving end, using the re-encoded video and depth data. R_v and R_d are the resulted data rates, and R_c is the total data rate constraint limited by the channel bandwidth. To seek the optimal QP and IRR pair for the video and depth, the full searching method is used [117]. Via traversing all the candidates of video and depth QP pairs, the optimal QP pair with minimal distortion for one rate constraint is selected, at a specific IRR and the associated PLR and data rate constraint. We can compute the optimal QP pairs corresponding to a series of discrete rate constraints, and then build a rate allocation table. The complete algorithm for building the transcoding rate allocation table is listed below:

Step 1: Encode the video and depth with $q_{1,v}$ and $q_{1,d}$, respectively. Decode the compressed video and depth streams and down-sample the reconstructed video and depth.

Step 2: According to the pre-set total rate (video rate plus depth rate) constraint levels, compute the all possible QP candidates for video and depth; Encode the down-sampled video and depth with all possible QP candidates and record all the encoded streams.

Step 3: Set the possible PLR levels of ρ_v and ρ_d. Through simulating the packet loss behaviors, decode the corrupt streams of different PLR levels with error concealments.

Step 4: Synthesize the virtual views with all possible combinations of video and depth, compute their distortions and then select the optimal QP pairs for video and depth with minimal distortions for different PLR levels.

Step 5: Build the Rate-QP-PLR table for different rate constraints under different PLR combinations of video and depth.

During the transcoding procedure, the cross-layer controller collects the channel PLR information according to different IRR and the data rate constraint, and selects the QP and IRR pair for the video and depth data, based on the minimal estimated view synthesis distortion, as shown in Formula (4.11).

4.6 Experimental Results

This section evaluates the proposed transcoding framework. Since video plus depth is used for mobile 3D video streaming, we have implemented the proposed 3D video transcoding framework based on JM17.1. The packet loss model of SVC/AVC [118] (including random loss and burst loss) and the error concealment method with motion vector prediction are used in the experiment. In the experiments, the GOP size is set to 30 and the IPPP coding structure is used. Currently, the proposed transcoding framework adopts the fixed QP coding for video and depth to guarantee the consistent visual quality in the temporal domain. To illustrate the performance of the proposed framework in

Figure 4.10, the fixed QP transcoding with fixed video/depth rate ratio (5:1) [100] is used as the comparison reference. In the transcoding, rate control can be used to obtain the appropriate bit-rate. However, the rate control often results in the non-consistent visual quality in the temporal domain. Thus, we also adopt the fixed QP for the reference transcoding. Since the virtual view generally does not exist, the virtual view synthesized by the original video and depth is used as the reference signal for computing the virtual view distortion. In the experiment, the HD 3D video sequence of *Poznan_CarPark* (1920x1088) is used for wireless transcoding with target display size of 320x240, and the original video and depth are both encoded with the QP of 32. For *Poznan_CarPark*, the video and depth of view2 are used to synthesize the view3 to generate the video plus depth based stereoscopic 3D video.

Figure 4.12 (a) shows the performance of the proposed transcoding with error-resilient video/depth rate allocation (Transcoding_ER_VDRA) compared with the transcoding with fixed ratio (5:1) video/depth rate allocation (Transcoding_FR_VDRA), when no packet loss occurs. It can be seen that, when no packet loss occurs the proposed Transcoding_ER_VDRA can obtain the better virtual view quality than the Transcoding_FR_VDRA. Currently, the QP for the reference transcoding is carefully selected to satisfy the bandwidth limitation, it means that the QP selection for the reference transcoding is also optimized. Therefore, the performance improvement of the proposed framework over the reference is not very large.

When the video and depth have the different PLRs, for example, 5% for the video and 10% for the depth, the transcoding performance can also be improved by the optimal error-resilient video/depth rate allocation, as Figure 4.12 (b) displays. Figure 4.12 (c) and

(d) show the error resilience performance of the proposed transcoding framework with the increasing PLRs. In these figures, the PLRs for video and depth have equal values.

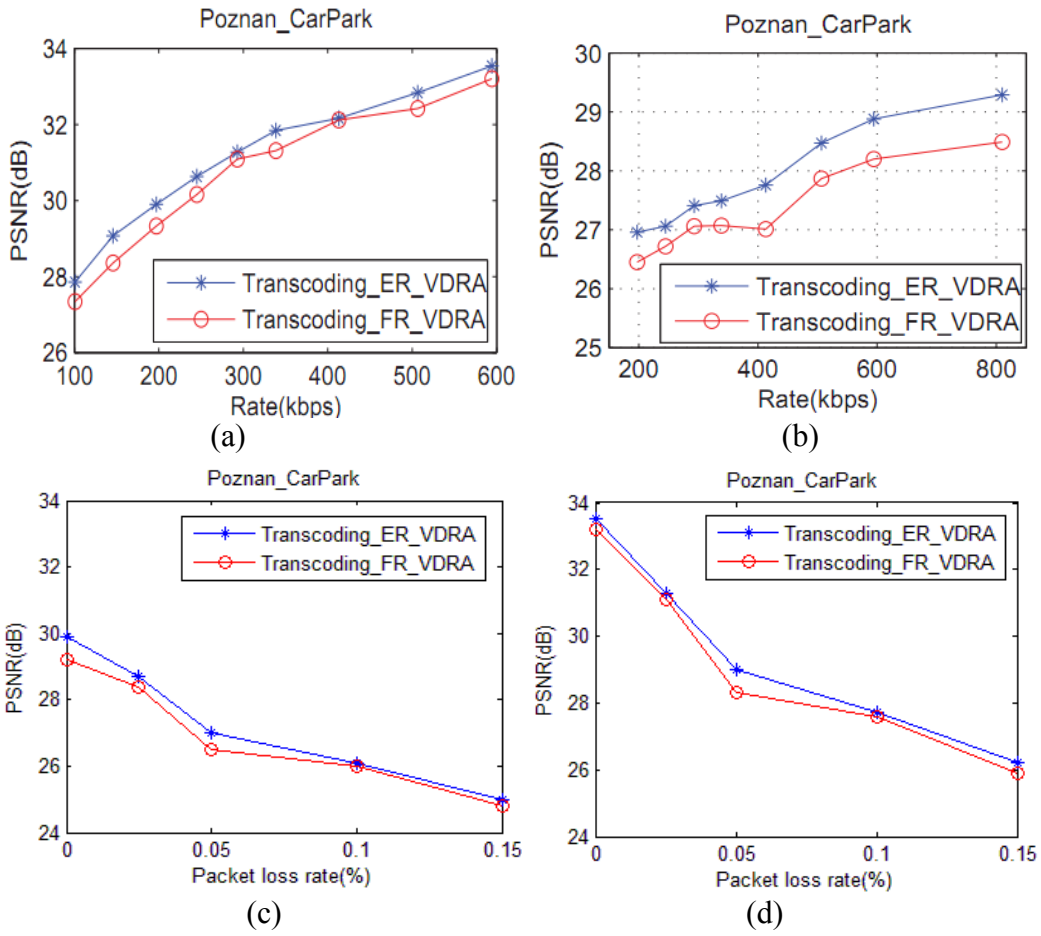
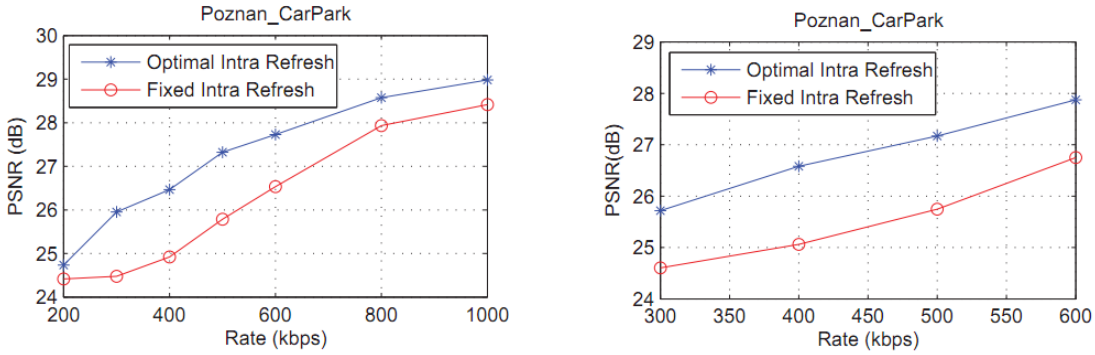


Figure 4.12 : (a) The performance of the proposed Transcoding_ER_VDRA compared with Transcoding FR_VDRA without packet loss; (b) The transcoding performance with PLR of 5% for video and PLR of 10% for depth; (c) The error-resilience performance of the proposed transcoding at total rate of 200kbps; (d) The error-resilience performance of the proposed transcoding at total rate of 600kbps.

To further eliminate the packet loss effect, the IRR is adaptively regulated in the transcoding. Intra refresh coding can result in the new optimal QP combination of video and depth. Figure 4.13 (a) shows the transcoding performance of the proposed transcoding framework with the optimal IRR. It can be seen from Figure 4.13 (a) that the

transcoding with the optimal IRR can promote the transcoding performance when the preset IRR is adapted to the actual PLR.

The wireless channel bandwidth is time-varying. The variable bit-rate (VBR) channel has much more influence on the 3D video than 2D video. To verify the adaptation performance of the proposed transcoding framework over the wireless channel, we also performed the VBR transcoding experiment. Since the original length of the sequence is very short for evaluating the proposed transcoding framework with VBR coding, the sequence is extended to 600 frames with three duplicate 200 frames. In the experiments, three temporal-periods with each period of 200 frames are used, and the Rate-QP-PLR tables for different temporal-periods are built and transmitted to the transcoding proxy. The rates vary from the first period to the third period, and the rate will increase 100kbps for each ordinal period. Figure 4.13 (b) shows the transcoding performance for the VBR channel, and the channel rate is the average rate of the three temporal-periods.



(a) (b)

Figure 4.13 : (a) The transcoding performance of optimal IRR compared with fixed 10% IRR; (b) The transcoding performance of optimal IRR compared with fixed 10% IRR under the variable bit-rate channel.

4.7 Summary

In a multi-camera surveillance environment, online 3D content generation is required for more advanced applications. In this work, we presented a video object tracking scheme incorporating a fast disparity estimation process into the Mean Shift based tracking algorithm, in order to adjust the size of the tracking window, thus the camera can better follow the moving object. Due to the run-time requirement of the tracking application, traditional disparity estimation methods such as graph cut and belief propagation do not suffice. While most of the existing real-time disparity estimation methods rely on GPU implementation, the proposed scheme achieves 6fps on an ordinary PC (2.66GHz CPU, 4GB RAM) on the recorded tracking video (640x480). Its accuracy and runtime performance are evaluated on a set of standard test data, and the comparison with the semi-real-time schemes by Geiger et al. and Kosov et al. is analyzed. Currently the disparity estimation is performed independently on each image pair. The temporal correlation utilizing the camera control parameters and the tracking performance with more complex scenes will be studied in the future work.

For the 3D video delivery over the WSN, we also proposed an efficient transcoding framework for mobile 3D video streaming with optimal video/depth rate allocation. Through building a Rate-QP-PLR table, the proposed framework can select the optimal transcoding QPs for the error-prone transmission to satisfy the channel rate constraint and therefore it can adapt to the error-prone network with the optimal quality for mobile 3D video streaming. In the future work, we shall intend to integrate the proposed framework with the fast transcoding architecture to promote the transcoding speed. Part of this work is published in [39, 119-121].

Chapter 5

Multi-camera Motion Capture for Remote Healthcare Monitoring

5.1 Introduction

Remote healthcare monitoring is becoming increasingly popular due to the advances in multiple disciplines, like sensor circuits, 3D modeling, and wireless communication technologies. One important task in a healthcare monitoring system is to provide a means to monitor walking patterns since it is a necessity for health evaluation of the neuromuscular system. For instance, gait analysis is often used to provide prognostic and diagnostic measures of pathological locomotion bio-rhythms such as Parkinson's disease, diabetic peripheral neuropathy, and Huntington's disease. It is also utilized for the clinical assessment of stroke rehabilitation, prosthetic alignment, and the success of orthopedic interventions such as anterior cruciate ligament reconstruction [122]. However, there are three major issues which prevent existing human gait monitoring systems from being used in the resource-limited environment such as rural clinics: 1) existing human motion capture systems using infrared sensing or other body sensing equipments are expensive. The average cost is around $250,000 which usually is not affordable for small clinics. 2) A motion capture system containing any body attachments, such as reflective or magnetic markers, gyroscopes and accelerometers, will be considered invasive, especially in geriatric attendance. 3) Currently most marker-less motion capture systems are dedicated to offline and error free video transmission in wired networks, such as the system designed by Ballan and Cortelazzo [123] and the one by Hasler et al [124]. When there is interaction between the remote caregiver and the patient involved, e.g. instruction on how to adjust the gait, real-time transmission of the monitoring videos over the WSN

is required. Optimal resource allocation to multiple video sequences is the most challenging problem in this kind of communication system. This issue is of primary concern when the communication resources are constrained.

Based on these considerations, we designed a marker-less motion capture system using multiple off-the-shelf cameras, aiming to provide caregivers with timely access to the patient's health status through mobile communication devices. This research is dedicated to developing a cost efficient remote healthcare monitoring system (through human gait analysis for neuro-health evaluation) at rural clinics in western Nebraska, based on our existing testbed of large-scale wireless multi-hop networks deployed in remote rural areas. The focus of this research is to study how to enhance the end-to-end video quality in an application-centric delay-constrained scenario through a cross-layer design method, by which video content analysis, video encoding/decoding, and video transmission are systematically considered. Therefore, multiple factors in the system level configuration are considered to determine the optimal video encoding and transmission parameters, including unequal error protection (UEP), transmission delay, quality balance, and error concealment.

To describe the function of each component in the multi-camera motion capture system, Section 5.2 describes the system architecture and the formulation of the delay-constrained video coding and transmission problem. The fast object detection algorithm for UEP is introduced in Section 5.3. The content-aware video coding and transmission procedure is described in Section 5.4, and the adaptive video coding and transmission procedure is described in Section 5.5. The error concealment scheme by the receiver is explained in Section 5.6. In Section 5.7, the multi-view motion estimation process is

described. Experimental results are provided in Section 5.8. Section 5.9 draws the conclusions.

5.2 Problem Description

The multi-camera motion capture system aims to provide caregivers with timely access to the patient's health status through mobile communication devices. The major components include video capture, object detection, video coding and transmission, error concealment, and video analysis. In the scenario, the subject walked on a treadmill with four tripod cameras capturing the video from different viewpoints. After video compression and transmission over a wireless sensor network, the remote receiver recovered the videos and performed multi-view motion capture for gait analysis.

5.2.1 System Architecture

The presented motion capture system for remote healthcare monitoring is illustrated in Figure 5.1. The videos showing the subject's walking pattern on a treadmill are recorded by four synchronized and calibrated tripod cameras from different viewpoints, as displayed in Figure 5.2. These videos are processed at the data center, i.e. the computer, where the RoI information is detected, and the parameters for video encoding and transmission are determined through cross-layer control. The multi-view motion estimation process is implemented by the receiver using the recovered videos and the camera calibration parameters [125]. To achieve optimal resource allocation, a content-aware video encoding and transmission procedure is applied by the cross-layer controller; and to ensure real-time video transmission, an adaptive encoding and transmission procedure is also applied concurrently based on the CSI. The number of cameras is

limited for the consideration of cost and processing time. The cameras are sparsely positioned around the treadmill, and little inter-view correlation exists between different videos. Therefore, the four sequences of video packets are simulcast over the WSN.

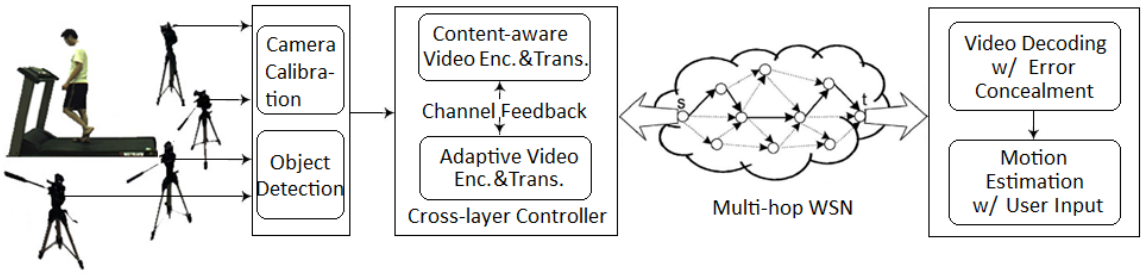


Figure 5.1 : Multi-camera motion capture system over WSN.

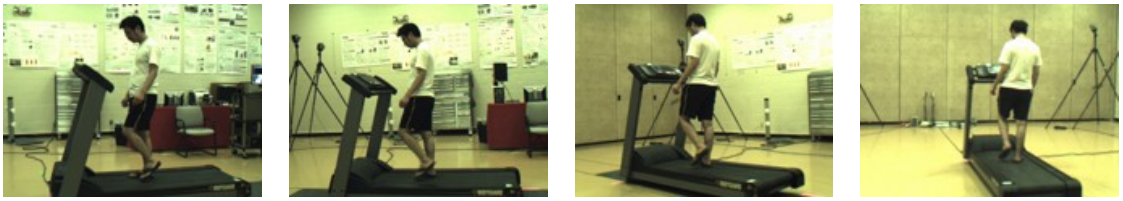


Figure 5.2 : Recorded video frame from four different views.

5.2.2 Formulation

At the cross-layer controller, the video encoding and transmission process is formulated as an end-to-end distortion minimization problem under a frame delay constraint:

$$\{s_{k,n}^{*}, c_{k,n}^{*}\} = arg\ min \sum_{k=1}^{K} \sum_{i=1}^{I} E[D_{k,n,i}]$$

$$s.t. \quad min \quad \max_{k=1,2,\ldots,K} \left(\sum_{i=1}^{I} E[D_{k,n,i}] \right)$$

$$\sum_{k=1}^{K} \sum_{i=1}^{I} E[T_{k,n,i}(s_{k,n}, c_{k,n})] \leq T^{max} \qquad (5.1)$$

Here $E[D]$ is the expected end-to-end distortion of one packet i, K is the number of views, and I is the number of packets in one frame. $\{s_{k,n}, c_{k,n}\}$ denotes the source coding parameter and channel transmission parameter vector for a frame n in view k. $E[T]$ represents the expected transmission time for one packet, and T^{max} is the maximum allowable delay for all the packets in one frame from K views to be transmitted.

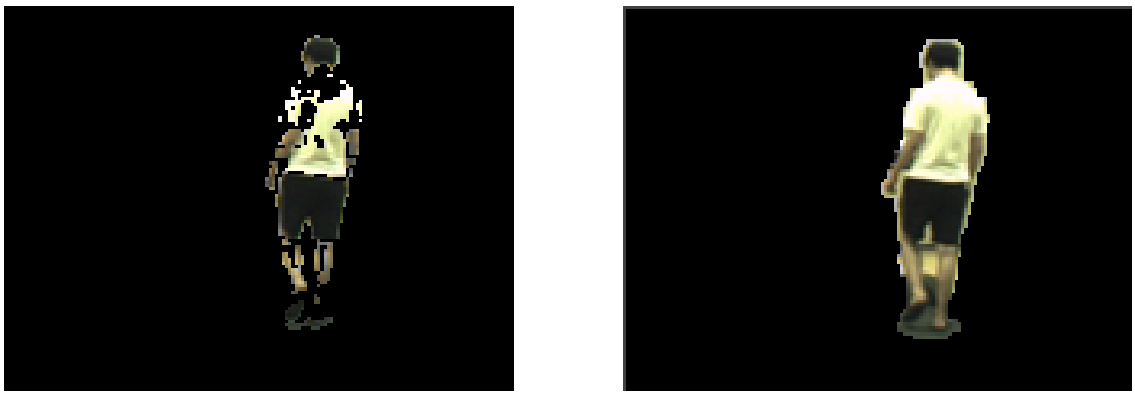
Besides frame delay, another constraint is that, the maximum distortion of all the video frames should be minimized, i.e., the lowest quality is maximized, which also implicates a balanced quality among all the views. This constraint is necessary since the visual quality of each received video is considered to contribute equally to a successful 3D motion estimation process.

According to Formula (5.1), a best parameter vector $\{s^*_{k,n}, c^*_{k,n}\}$ is chosen for a new frame based on multiple factors affecting the expected distortion, including RoI, current channel condition, and previous packet loss information. Details of the solution procedures are explained in following sections.

5.3 Fast Object Detection

Background subtraction using Gaussian Mixture Model (GMM) is a popular video motion detection method known for its change adaptability and noise tolerance. GMM is an online learning process. Each pixel in a new frame is checked against the existing background models until a match is found. A match is defined as the distance between the mean and the pixel value is within 2.5 times the standard deviation [71]. To accelerate the learning process, the background setting without moving objects is recorded at the

beginning of the video, when sufficient data can be acquired to train the background models. Figure 5.3 (a) shows the foreground detection results for one frame in one view.



(a) Background subtraction (b) Anisotropic diffusion

Figure 5.3 : Object detection.

A problem with the temporal GMM based motion detection method is that it fails to detect some foreground regions with similar color to the background. As can be observed from Figure 5.2 and Figure 5.3, part of the body area is missing where the color of the T-shirt is close to the color of the wall. Spatial color correlation can be utilized to solve this problem, such as anisotropic diffusion [126]. Here anisotropic diffusion is applied as a post-processing step to improve the detection result. The motive of the diffusion process is to minimize the difference between neighboring pixels, while the edge property is preserved. The process can be formulated as the following energy minimization problem.

$$min \quad \iint_{\Omega} \|\nabla I\|^2 dxdy \tag{5.2}$$

where Ω is the image domain, and (x, y) is the pixel I's coordinates. The solution can be obtained using the gradient decent [127],

$$\frac{\partial I}{\partial t} = -\lambda \, \Delta I \tag{5.3}$$

and a 4-nearest-neighbors discretization of the diffusion is expressed as

$$I_i^{t+1} = I_i^t + \lambda[c_N \nabla_N I_i^t + c_S \nabla_S I_i^t + c_E \nabla_E I_i^t + c_W \nabla_W I_i^t] \tag{5.4}$$

where I_i^t is the diffusion value at iteration t and at pixel i. λ is a constant between 0 and 1/4. N, S, E, W are subscripts for North, South, East, West. ∇I_i^t denotes the nearest-neighbor difference, and the conduction coefficient c is a kernel function of the Euclidean norm of ∇I_i^t,

$$c_i^t = f(\|\nabla I_i^t\|) \tag{5.5}$$

We design the kernel function as reversely increasing with ∇I_c, the color difference between adjacent pixels,

$$f(\|\nabla I_i^t\|) = \frac{w(\nabla I_c)}{\left(1+(A\|\nabla I_i^t\|)^2\right)\sum_{\aleph} w(\nabla I_c)} \tag{5.6}$$

$$w(\nabla I_c) = e^{-(B\|\nabla I_c\|)^2} \tag{5.7}$$

where A and B are predefined constants controlling the diffusion speed. $\aleph$ denotes the neighboring pixels. The diffusion value is initiated with GMM learning result, i.e., if a pixel i is detected as background, $I_i^0 = 0$; otherwise $I_i^0 = 1$. At the end of each iteration, resulting I_i^{t+1} is thresholded so that pixels with higher I_i^{t+1} value are determined as

foreground. The iteration process is terminated either when the predefined maximum number of iteration is reached, or when the difference of the number of detected foreground between two successive iterations is below certain threshold, whichever comes first. Function (5.6) is a weighted version of the kernel function introduced in [126]. The merit is that if some region is missing, and it has neighboring foreground regions with similar color, its diffusion value will be raised continuously during the iterative diffusion process, making it more likely to be merged with those neighboring foreground regions. The final detection results are displayed in Figure 5.3 (b).

The video object detection algorithm has an efficient implementation. For 300 recorded 640x480 frames from one view, the average processing time is 0.3 second per frame on a 32-bit PC machine with Intel E7300 2.66GHz CPU and 2GB RAM. Compared to traditional video object detection methods, the presented algorithm is 15% faster than the min-cut [128], and 50% faster than the Iterative Conditional Mode (ICM) [110], with similar visual quality. The RoI region is defined as the smallest rectangle containing all the foreground pixels, aligning to the encoder block size. When the computation resource is constrained, only the data from one view is processed, the frames are down sampled (average processing time is 0.02 second per 160x120 frame), and the RoI regions for other views are projected using the camera parameters, and the input of the object's stature [125].

5.4 Content-aware Video Coding and Transmission

The recorded videos endure data compression and transmission before arriving at the receiver. When the communication resources are limited in a WSN, an alternative of heavier compression is to implement unequal error protection (UEP) to impose higher

priority on the parts of the video sequence that have a greater impact on video quality, e.g. the RoI [21, 129]. Thus the distortion estimation process by the cross-layer controller needs to consider the different packet delay for the unequally protected foreground and background packets.

5.4.1 Unequal Error Protection

In the content-aware video coding and transmission procedure, the foreground data and the background data are grouped into different packets. While the sender applies the same compression and transmission parameters to all packets in one frame, the intermediate nodes in the WSN put a foreground packet ahead of all background packets in the queue. When a packet is lost, it will be retransmitted until it is correctly received, or is discarded when the maximum transmission delay T^{max} is exceeded. This UEP mechanism reduces the packet loss possibility of the foreground target packets due to the transmission delay constraint.

5.4.2 End-to-end Packet Delay

As a result of the retransmission mechanism, the packet loss probability over a link between two nodes (u, v) mainly exhibits as the probability of packet drop due to delay deadline expiration when queuing at node u. Based on priority queuing analysis, it can be calculated from the tail distribution of the waiting time [130]:

$$p_{g,u} = Prob\left(E\left[W_{g,(u,v)}\right] + t_{g,u}^{0} > T^{max}\right)$$

$$= \left(\sum_{g=0}^{1} \phi_{g,u} E\left[Z_{g,u}\right]\right) \cdot e^{-\frac{\left(T^{max} - t_{g,u}^{0}\right)\sum_{g=0}^{1} \phi_{g,u} E[Z_{g,u}]}{E\left[W_{g,(u,v)}\right]}} \tag{5.8}$$

$$g = \begin{cases} 0, & if\ it\ is\ a\ foreground\ packet \\ 1, & if\ it\ is\ a\ background\ packet \end{cases} \tag{5.9}$$

where $t_{g,u}^{0}$ is the packet arrival time at node u, and $\phi_{g,u}$ is the average arrival rate of the Poisson input traffic into the queue at node u. $E[W_{g,(u,v)}]$ is the average packet waiting time at the queue of node u, and $E[Z_{g,u}]$ is the average service time at node u, measured as a geometric distribution with the effective transmission rate (goodput), packet length, and packet error and collision rate. Both the goodput and the packet error and collision rate are related to the link SINR (signal to interference and noise ratio) information and the selected modulation and channel coding scheme (MCS) [29]. Accordingly, the end-to-end packet loss rate (PLR) over a selected path P is estimated as

$$p_g = 1 - \prod_{(u,v)\in P}\left(1 - p_{g,u}\right) \tag{5.10}$$

The end-to-end packet delay is estimated as the sum of the packet delay $t_{g,(u,v)}$ over each link (u, v):

$$T_g = \sum_{(u,v)\in P} t_{g,(u,v)} = \sum_{(u,v)\in P}\left\{E\left[Z_{g,u}\right] + E\left[W_{g,(u,v)}\right]\right\} \tag{5.11}$$

The estimated packet loss rate and delay over each path are used by the cross-layer controller for optimal decision of coding and transmission parameters based on Formula (5.1). The solution strategy is summarized in next section.

5.5 Adaptive Video Coding and Transmission

The multiple video sequences are simulcast over a multi-hop WSN. To accommodate the dynamic channel condition, flexible configuration of the video encoding and transmission parameters is enabled, including the selection of quantization parameter (QP), coding mode, MCS, and transmission path, resulting in a configuration quadruple (*Q*, *Mode*, *MCS*, *P*). In literature, how to choose the combination of the parameters for multiple sequences has been studied in various video streaming applications [17, 131]. Without the min-max (quality balance) constraint, the problem expressed in Formula (5.1) resembles the multiple-choice knapsack problem (MCKP) in classical combinatorial optimization [132]. In our application, the resource allocation is constrained by both transmission delay and quality balance. The expected video distortion is estimated with online CSI. And the optimal encoding and transmission parameters are configured by a cross-layer controller based on the distortion estimation results, using a greedy search algorithm.

5.5.1 End-to-end Distortion Estimation

When transmitted over the wireless network, the end-to-end distortion of a video packet includes the source coding distortion D^s and channel distortion D^c. Under a given configuration (*Q*, *Mode*, *MCS*), an optimal path *P* is selected based on the estimated video distortion, using the routing algorithm similar to the work in [29]. According to Equations (5.8) to (5.11), the estimated distortion for a packet π_g is

$$D_g(Q, Mode, MCS, P) = \begin{cases} E\left[\sum_{i\in\pi_g}\left(f_i - \tilde{f}_i\right)^2\right], & if\ \ T_g > T^{max} \\ D_g^s + D_g^c, & else \end{cases} \tag{5.12}$$

$$D_g^s = (1 - p_g) \cdot E\left[\sum_{i \in \pi_g} (f_i - \hat{f}_i)^2\right] \tag{5.13}$$

$$D_g^c = p_g \cdot E\left[\sum_{i \in \pi_g} (\hat{f}_i - \tilde{f}_i)^2\right] \tag{5.14}$$

f denotes the original data. $\hat{f}$ is the encoder recovered data after quantization. $\tilde{f}$ is the concealed data in the presence of packet loss. It is determined based on the receiver's packet loss feedback for previous frames. When the estimated packet delay is larger than the threshold, the concealment result is used to calculate the distortion directly. It is assume that perfect channel CSI is available to the sender without error and latency. This assumption could be approximately satisfied by using a fast feedback channel with powerful error control information as adopted in [88].

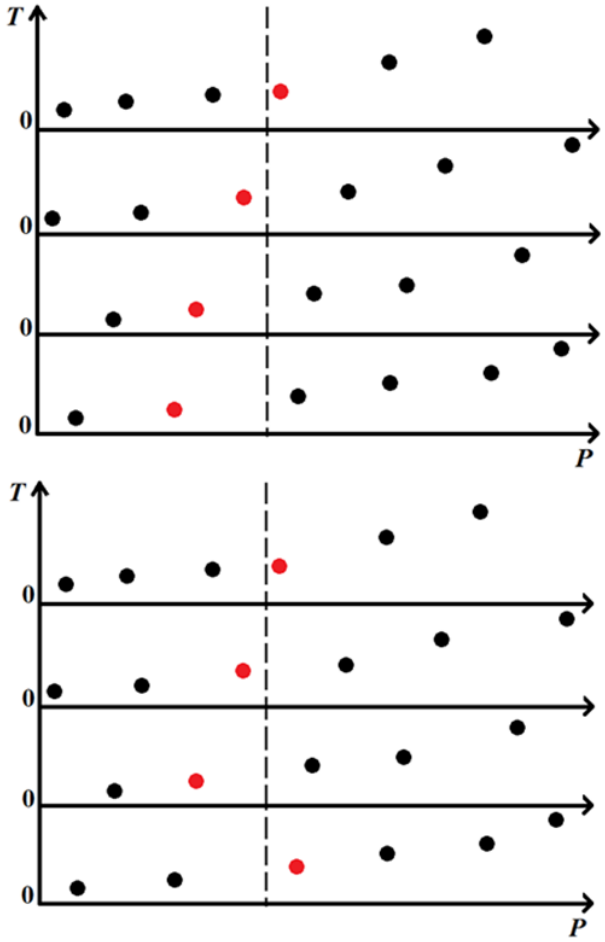
5.5.2 Cross-layer Control

From previous discussion, each configuration quadruple leads to a $\{D, T\}$ pair. It serves as an operation point for parameter selection. For each frame in a single view, the number of operation points is factored by the number of packets and available QPs, coding modes, and MCSs. To reduce the overhead, the packets in one frame share the same configuration. Maximum and minimum QPs for each view are tested under different coding modes and MCSs. The (*Mode*, *MCS*, *P*) configuration with minimum distortion is first selected for current frame in each view. To accommodate the video with the lowest quality, the selected (*MCS**, *P**) with maximum distortion among *K* views is assigned to other views. Then the maximum and minimum QPs are tested again under different coding modes and the assigned (*MCS**, *P**) to choose the optimal coding mode for each of the other views. After the (*Mode**, *MCS**, *P**) parameters are determined for each

view, operation points using different QPs are generated, i.e. the number of operation points for each view is identical to the number of QPs, N_Q. The optimal QP is then chosen for each view according to Formula (5.1). To compare with the MCKP algorithm aiming at maximum sum product [132], the $\{D,T\}$ pair is transformed to $\{P,T\}$. P represents the quality (product), e.g. PSNR. It bears an increasing profile with T (weight). The solution procedure is listed in Figure 5.4 (a). Figure 5.4 (b) illustrates the point selection procedures with four views. The upper half shows selected four operation points (in red color) before step 2.4 proceeds. The lower half shows the selection result after first replacement.

1 Arrange the $\{P, T\}$ operation points $\{P_{jk,k}, T_{jk,k}\}$, $jk = 1, 2,\ldots, N_Q$, $k = 1, 2,\ldots, K$, for each view in an increasing order. Remove the dominated points, i.e. $\{P_{jk,k}, T_{jk,k}\}$ is removed if $T_{jk,k} > T_{jk-1,k}$ and $P_{jk,k} \le P_{jk-1,k}$.
2 Select any one view k. Beginning with the point containing the highest weight that satisfies $T_{jk,k} < T^{max}$, perform the following greedy search:
(2.1) For each view h ($h \neq k$), find the point $\{P_{jh,h}, T_{jh,h}\}$, $P_{jh,h} \le P_{jk,k}$, $P_{jh+1,h} > P_{jk,k}$. If $P_{1,h} > P_{jk,k}$, $jh = 1$.
(2.2) Calculate the total delay $T_s = \sum_{iz\epsilon\{j1,j2,\ldots,jK\}, z\epsilon\{1,2,\ldots,K\}} T_{iz,z}$. If $T_s \le T^{max}$, go to step 2.4.
(2.3) If $jk = 1$, no solution exists. The program terminates. Otherwise set $jk = jk - 1$ and go to step 2.1.
(2.4) Sort the selected points $\{P_{iz,z}, T_{iz,z}\}$ according to increasing P. From the first one, calculate $T_{temp} = T_s - T_{iz,z} + T_{iz+1,z}$. If $T_{temp} < T^{max}$, set $T_s = T_{temp}$, replace $\{P_{iz,z}, T_{iz,z}\}$ with $\{P_{iz+1,z}, T_{iz+1,z}\}$, and repeat step 2.4. Else if $T_{temp} == T^{max}$, output all points, otherwise output current point.
(2.5) Output the selected combination $\{P_{iz,z}, T_{iz,z}\}$ and the corresponding QPs. The program terminates.

(a) Greedy search.



(b) Operation points.

Figure 5.4 : Search for optimal combination of QPs.

5.6 Error Concealment

To counteract packet loss, error resilience and error concealment technologies are adopted to improve the video quality, including interleaving and boundary match. Before video encoding, interleaving is implemented to separate spatially neighboring MBs into different packets. This interleaving mechanism contains two steps, chessboard decomposition and row separation, as show in Figure 5.5. The chessboard decomposition separates horizontally or vertically connected MBs into two groups. The row separation

in one group further divides the MBs in the odd-numbered and the even-numbered rows into two subgroups. These two steps can be repeatedly operated on one group of MBs until the desired data length is acquired. A successive implementation of these two steps ensures that no MBs in one group are directly connected (horizontally, vertically or diagonally) with each other in the original image. Further, after i times of repeating such implementation, no MBs in a 2^ix2^i neighborhood are in the same group. The receiver performs deinterleaving according to the information of encoder interleaving steps and the packet order.

For lost blocks in received video, the decoder performs boundary match [35] to search for similar patches in a spatiotemporal neighborhood. A patch yielding the smallest difference value in the search area is used to replace the missing MB, followed by a deblocking filter.

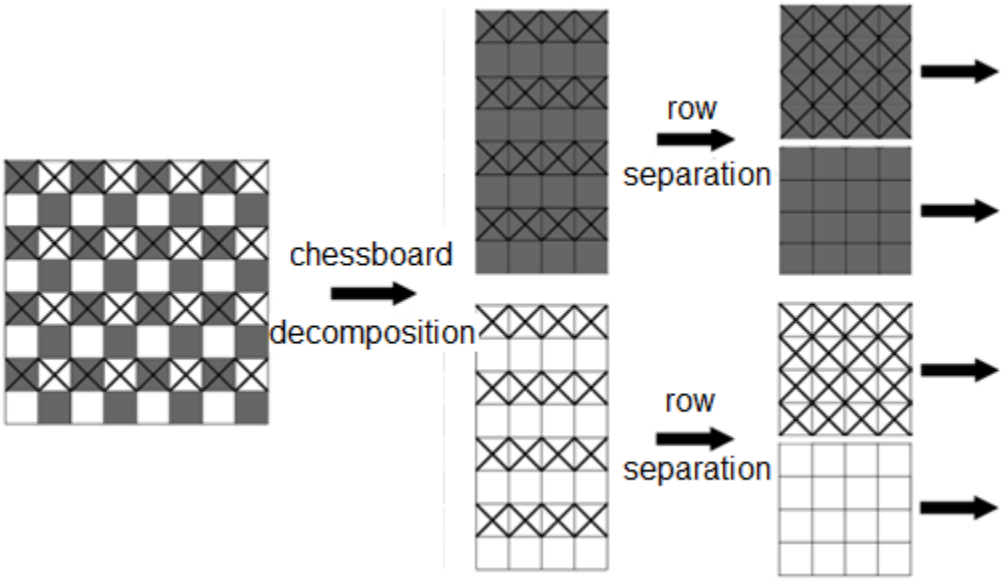


Figure 5.5 : Interleaving.

5.7 Motion Estimation

The recovered video sequences are observed by the receiver. For motion estimation, the 3D positions of the object's joints are reconstructed using triangulation [133] based on the selected 2D coordinates from each view, as shown in Figure 5.6. Specifically, the projection from a point M in 3D world coordinates (X, Y, Z) to a pixel m_i (x, y) on a 2D image plane is

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \backsim PM \quad \text{or} \quad \begin{cases} x = P(1)M/P(3)M \\ y = P(2)M/P(3)M \end{cases} \tag{5.15}$$

$P(i)$ is the i-th row of the camera projection matrix P. Equation (5.15) is equivalent to

$$\begin{bmatrix} P(3)x - P(1) \\ P(3)y - P(2) \end{bmatrix} M = AM = 0 \tag{5.16}$$

For K views, there is a system of equations according to Equation (5.16). The solution for M is obtained by singular value decomposition using the joint matrix [A_1; A_2; ...; A_K].

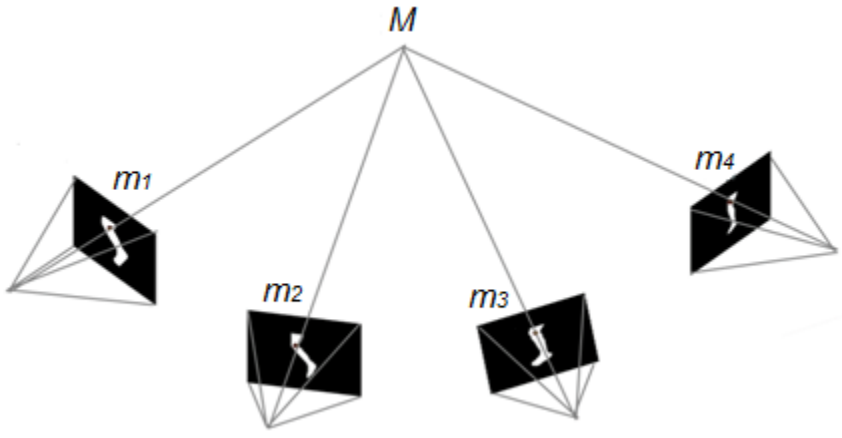


Figure 5.6 : Triangulation.

5.8 Experimental Results

In our experiment, four tripod cameras (PointGrey Firefly MV) are placed around the object for video recording. The image size is 640x480. 100 frames from each view are processed. They are down-sampled to 160x120 to accelerate the computation. The video codec is based on the H.264/AVC standard [90]. The available QP set is {16, 20, 24, 28, 32, 34, 36, 38, 39, 40}. The MCSs include MCS1 (6, 2/3), MCS2 (4, 3/4), MCS3 (2, 1/2), and MCS4 (1, 1/2) with a packet size 1k bytes [87]. A 30-node network with a directed-acyclic-graph (DAG)-modeled connectivity structure and the Rayleigh fading channel [29] is simulated in MATLAB. The packet arrival rate at each node is set to 100 packets/s. To test the system performance under different conditions, the frame delay constraint is set to 15 fps and 30 fps, the average SINR is set to 15dB and 20dB, and the channel bandwidth BW is set to 100kHz and 1MHz.

The content-aware video coding and transmission procedure places higher priority on foreground packets. Under better channel condition (BW = 1MHz, $\overline{\text{SINR}}$= 20dB), the average PSNR for the RoI is 36dB under 15 fps delay constraint, and 32dB under 30 fps, 2-5 dB higher than the traditional coding and transmission scheme without priority, as shown in Figure 5.7.

The adopted error concealment also has significant impact on the visual quality of the received videos. Figure 5.8 (a) shows one recovered frame using the traditional scheme with slice copy as the error concealment measure. Compared to the result in Figure 5.8 (b) obtained with the proposed method, the misplaced ankle could impose considerable error for the 3D motion estimation.

The adaptive coding and transmission procedure provides more accurate rate-distortion control under the dynamic channel condition, as demonstrated in Figure 5.9 (a) and (b). The source coding scheme using a fixed MCS and transmission path is compared with the proposed method, on the average PSNR of four views. The delay constraint is set to 30 fps.

The parameter selection procedure in Section 5.5.2 achieves the min-max requirement as expressed in Formula (5.1). Figure 5.10 lists a set of {P(dB), T(ms)} operation points for one frame from four views. The total weight constraint is 30ms. The selected combination by the MCKP algorithm [132] is {15, 2}, {15, 4}, {28, 10}, {32, 14}. The result with our algorithm is {19, 5}, {23, 8}, {21, 7}, {20, 10}. The total product is lower, but the lowest quality is improved from 15 to 19, as well as the quality variance among different views.

Finally, to illustrate the motion capture process, the reconstructed 3D points at four different time instances are displayed in Figure 5.11. The blue markers represent the joints at the hip, knee, and ankle of the left leg, and the black markers represent the corresponding joints of the right leg. The primary objective of gait analysis is to identify the variability of gait patterns. Linear or non-linear measures, such as the standard deviation and coefficient of variation, and the Lyapunov Exponent [134], can be applied to investigate the gait variability in patients with knee anterior cruciate ligament deficiency and reconstruction, and in the elderly.

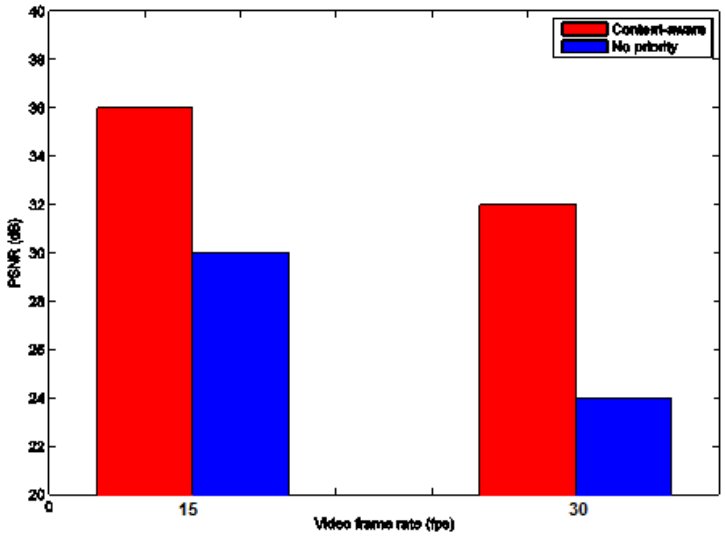
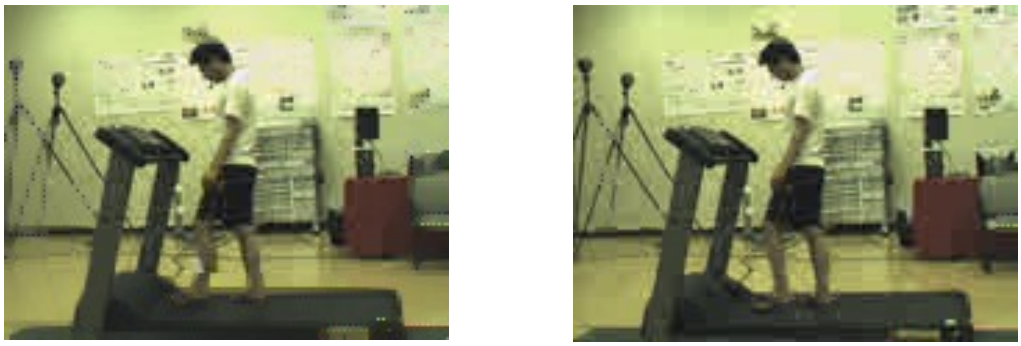
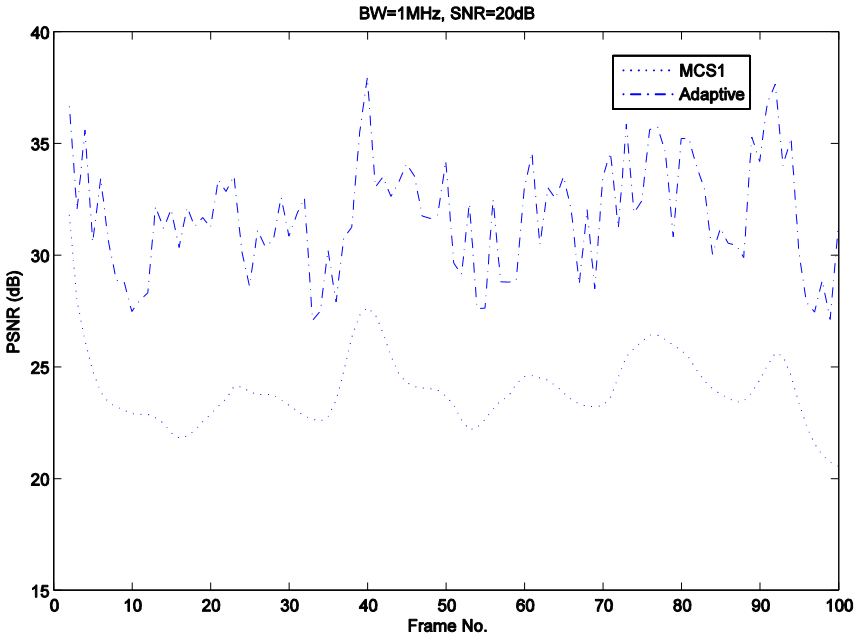


Figure 5.7 : Average PSNR under different delay constraints.

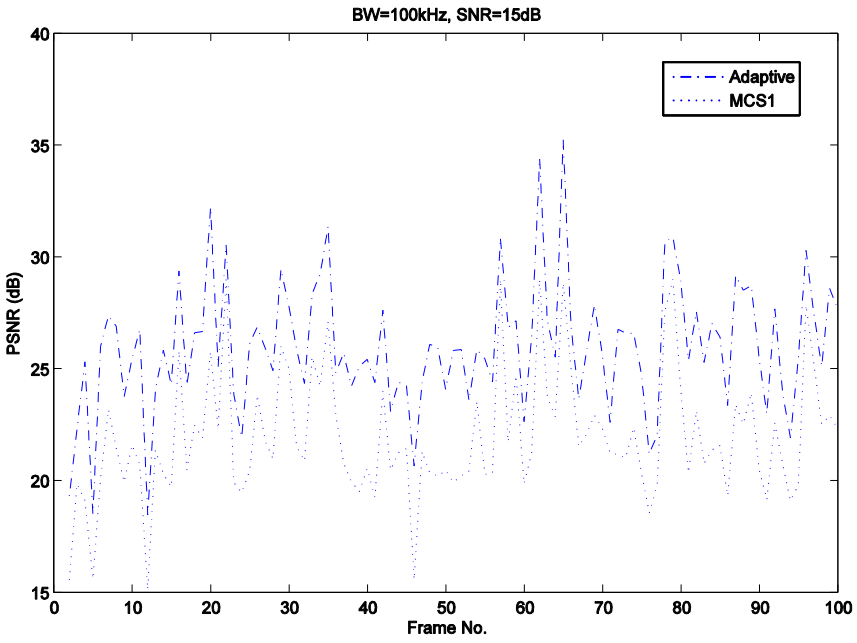


(a) Slice copy based concealment. (b) Interleaving based concealment.

Figure 5.8 : Error concealment.



(a)



(b)

Figure 5.9 : Video coding and transmission.

view 1	view 2	view 3	view 4
12,1	9,1	11,2	10,3
15,2	13,3	12,4	13,5
16,3	15,4	14,5	14,6
17,4	16,6	18,6	17,8
19,5	20,7	21,7	20,10
22,7	23,8	24,9	23,11
25,8	24,9	28,10	26,13
26,10	28,11	29,12	32,14
32,18	35,20	38,20	36,23
39,27	40,31	45,22	40,29

Figure 5.10 : Operation points.

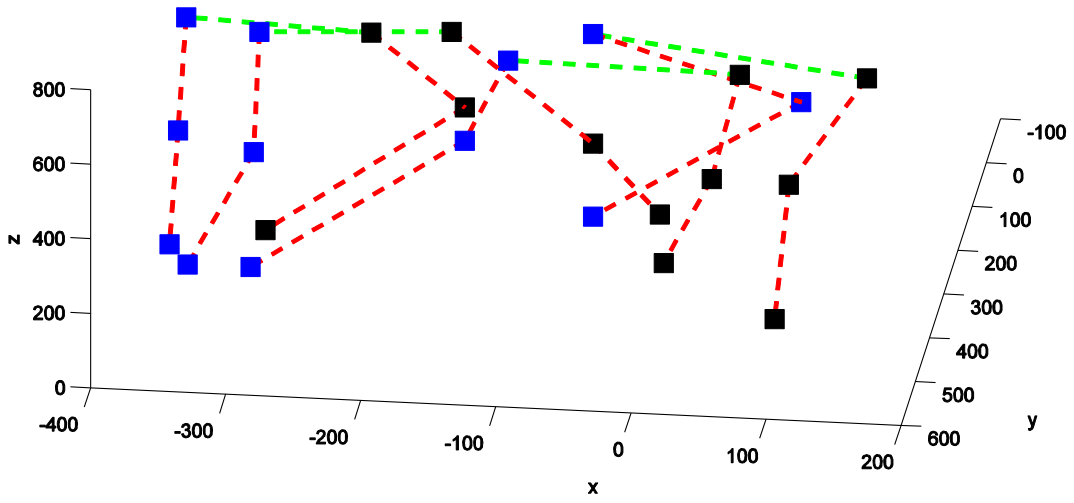


Figure 5.11 : Motion capture.

5.9 Summary

In remote healthcare monitoring, gait analysis is gaining increasing interests due to its value in health evaluation of the patient's neuromuscular system. Among existing technologies, motion capture based on multi-camera video communications over wireless sensor networks (WSN) is considered an effective and efficient means for gait analysis in remote healthcare monitoring. How to allocate the limited wireless channel resources to the multi-camera video data to ensure maximum and balanced visual quality is the key challenge in this kind of communication system. Based on the introduction of our multi-camera motion capture system designed to provide caregivers with timely access to the patient's health status through mobile communication devices, this section presented a solution for the video encoding and transmission process over WSN through cross-layer control. All components are seamlessly integrated in a unified cross-layer optimization framework dedicated for online data transmission, including coding mode and QP selection for video encoding, and MCS and path selection for video transmission. Experimental results show that the presented system design achieves better video quality than traditional video coding and transmission scheme, while the requirement for a low-cost, noninvasive and real-time healthcare monitoring system is accommodated. Part of this work is published in [135].

Chapter 6

Summary and Future Work

6.1 Summary of the Thesis Work and Our Contributions

This thesis work studied the cross-layer optimization in different wireless video surveillance systems, including the wireless video surveillance system with a single PTU camera, the binocular video object tracking and mobile 3D video transcoding system, and the multi-camera motion capture system for remote healthcare monitoring. The cross-layer control design considers the specific architecture of each surveillance system and accommodates the practical requirements by different application scenarios, such as the camera control, the computation and communication resource limits, the video quality enhancement and balance, and so on. Interdisciplinary study is conducted to incorporate different components of the system, including video object detection, data compression and transmission, and video analysis, into a delay-constrained resource optimization framework over WSN. The cross-layer controller adaptively selects the source coding parameters and channel transmission parameters according to the available system resources and the CSI information. Our contributions are summarized below:

- We implemented systematic design for different wireless video surveillance systems, including the wireless video suveillance system with a single PTU camera, the binocular video object tracking and mobile 3D video transcoding system, and the multi-camera motion capture system for remote healthcare monitoring. General components, i.e. the video capture, the video compression and transmission, and the video analysis, and the specific requirements of a practical surveillance system are considered.

- We studied the cross-layer optimization mechanism to incorporate different components of the system into a delay-constrained resource optimization framework over WSN. The cross-layer designs for the source processing and wireless transmission of the single surveillance video, the video plus depth 3D data, and the multiple simulcast videos, can guarantee enhanced video quality and timely video playback.
- Multiple video quality enhancement strategies are proposed for the surveillance systems, including the block-based interleaving error resilience, the boundary match error concealment, the down-sampling/up-sampling video coding, the dynamic rate allocation for video/depth data, the UEP for foreground/background packets, and the balanced quality for multiple video simulcast.
- Fast implementation of the video processing methods is considered for the run-time requirement of the surveillance system. For example, the 3D content generation process in the binocular video object tracking accommodates the run-time surveillance application. The background subtraction in object detection, the motion estimation in the video compression, and the boundary match in the error concealment, all utilize the special property of the PTU camera movement in the cyber-physical system to accelerate the computation.

6.2 Future Work

Wireless video surveillance is popular in various visual communication applications. While the advanced WSN infrastructure provides a strong support for surveillance video communications, new challenges are emerging in the process of compressing and transmitting large amount of video data, and in the presence of run time and energy

conservation requirements for mobile devices. Another trend in this field is the 3D signal processing technology in more advanced multiview video surveillance. The wireless communication environment posts greater difficulty for this kind of applications. How to efficiently estimate the distortion for the dedicated vision task at the receiving end using the compressed and concealed video data remains a research topic worth exploring. Our future work includes:

- Extend the resource constraint in the cross-layer control process to power/energy, consumed by both the computation and communication modules in the surveillance system.
- Further explore the cross-layer design for the multi-view video surveillance system. Study the efficient implementation for more complex 3D data processing algorithms in more advanced surveillance applications.
- Research on the rate-distortion model for other popular vision tasks, such as the 3D scene reconstruction, and the free viewpoint video-on-demand application, especially in a wireless environment.

Bibliography

[1] Electronic Code of Federal Regulations, available at: http://ecfr.gpoaccess.gov/cgi/t/text/text-idx?c=ecfr&sid=1143b55e16daf5dce6d225ad4dc6514a&tpl=/ecfrbrowse/Title47/47cfr15_main_02.tpl.

[2] M. Intag, "Wireless Video Surveillance –Challenge or Opportunity?" online available: http://www.bicsi.org/pdf/conferences/winter/2009/presentations/Wireless_Security_and_Surveillance_Challenge%20or%20Opportunity%20-%20Mike%20Intag.pdf

[3] S. Leader, "Telecommunications Handbook for Transportation Professionals -- The Basics of Telecommunications, Technical Report of Federal Highway Administration (FHWA-HOP-04-034)," 2004, available at: http://ops.fhwa.dot.gov/publications/telecomm_handbook/telecomm_handbook.pdf.

[4] J. Hourdakis, T. Morris, P. Michalopoulos, and K. Wood, "Advanced Portable Wireless Measurement and Observation Station, Report of Center for Transportation Studies in University of Minnesota (CTS 05-07)," 2005, available at: http://conservancy.umn.edu/bitstream/959/1/CTS-05-07.pdf

[5] N. Luo, "A Wireless Traffic Surveillance System Using Video Analytics," master thesis, University of North Texas, 2011. online available: http://digital.library.unt.edu/ark:/67531/metadc68005/m2/1/high_res_d/thesis.pdf

[6] C. Hartung, R. Han, C. Seielstad and S. Holbrook, "FireWxNet: A multi-tiered portable wireless system for monitoring weather conditions in wildland fire environments," Proceedings of the 4th international conference on Mobile systems, applications and services, pp. 28--41, 2006.

[7] A. Kawamura, Y. Yoshimitsu, K. Kajitani, T. Naito, K. Fujimura, and S. Kamijo, "Smart camera network system for use in railway stations," IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp.85,90, 2011

[8] N. Li, B. Yan, G. Chen, P. Govindaswamy and J. Wang, "Design and implementation of a sensor-based wireless camera system for continuous monitoring in assistive environments, Journal of Personal and Ubiquitous Computing," vol.14, Issue 6, pp. 499 - 510,Sep 2010.

[9] B.P. L. Lo, J. Sun, and S. A. Velastin, "Fusing visual and audio information in a distributed intelligent surveillance system for public transport systems," Acta Automatica Sinica, 29, (3), pp. 393–407, 2003.

[10] W. Feng, B. Code, M. Shea and W. Feng, "Panoptes: A Scalable Architecture for Video Sensor Networking Applications," ACM Multimedia, pp. 151--167, 2003.

[11] S. Hengstler, D. Prashanth, S. Fong, and H. Aghajan, "MeshEye: A hybrid-resolution smart camera mote for applications in distributed intelligent surveillance," in Proc. Int. Conf. Inf. Process. Sensor Netw. (IPSN), Cambridge, MA, 2007, pp. 360–369.

[12] X. Wang, S. Wang, and D. Bi, "Distributed Visual-Target-Surveillance System in Wireless Sensor Networks," IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, vol.39, no.5, pp.1134-1146, Oct. 2009.

[13] S. Misra, M. Reisslein, and G. Xue, "A Survey of Multimedia Streaming in Wireless Sensor Networks," IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 10, NO. 4, 2008.

[14] M. Van der Schaar and S. Shankar, "Cross-layer wireless multimedia transmission: Challenges, principles, and new paradigms," IEEE Wireless Commun. Mag., 12(4):50–58, August 2005.

[15] D. Wu, S. Ci, H. Luo, Y. Ye, and H. Wang, "Video Surveillance over Wireless Sensor and Actuator Networks Using Active Cameras," TAC, vol.56, no.10, pp.2467-2472, Oct. 2011.

[16] Y. Andreopoulos, N. Mastronarde, and M. van der Schaar, "Cross-Layer Optimized Video Streaming over Wireless Multi-hop Mesh Networks," JSAC, vol.24, no.11, pp.2104-2115, Nov. 2006.

[17] D. Wu, S. Ci, and H. Wang, "Cross-layer optimization for video summary transmission over wireless networks," JSAC, vol.25, no.4, pp.841-850, May 2007.

[18] Z. Chen and D. Wu, "Rate-Distortion Optimized Cross-layer Rate Control in Wireless Video Communication," CSVT, vol.22, no.3, pp.352-365, March 2012.

[19] E. Setton, T. Yoo, X. Zhu, A. Goldsmith, and B. Girod, "Cross-layer design of ad hoc networks for real-time video streaming," IEEE Wireless Commun. Mag., vol. 12, no. 4, pp. 59–65, Aug. 2005.

[20] H. Wang, F. Zhai, Y. Eisenberg, A.K. Katsaggelos, "Cost-distortion optimized unequal error protection for object-based video communications," CSVT, vol.15, no.12, pp. 1505- 1516, Dec. 2005.

[21] S. Cui, and A.J. Goldsmith, "Cross-layer optimization of sensor networks based on cooperative MIMO techniques with rate adaptation," IEEE 6th Workshop on Signal Processing Advances in Wireless Communications, pp.960,964, 2005.
[22] R. Madan, S. Cui, S. Lall, and A. Goldsmith ,"Cross-layer design for lifetime Maximization in Interference-Limited Wireless Sensor Networks," IEEE Transactions on Wireless Communications, vol.5, no.11, pp.3142,3152, November 2006.
[23] A. Scaglione and M. Van der Schaar, "Cross-layer resource allocation for delay constrained wireless video transmission," ICASSP, 2005.
[24] I. F. Akyildiz and I. H. Kasimoglu, "Wireless sensor and actor networks: Research challenges," Ad Hoc Netw., vol. 2, no. 4, pp. 351–367, May 2004.
[25] P. Petrov, O. Boumbarov, and K. Muratovski, "Face detection and tracking with an active camera," in Proc. 4th Int. IEEE Conf. Intell. Syst., Varna, Bulgaria, Sep. 2008, pp. 14–39.
[26] Y. Ye, S. Ci, Y. Liu, and H. Tang, "Dynamic Video Object Detection with Single PTU Camera," VCIP, Nov. 2011.
[27] A. Ortega, K. Ramchandran, and M. Vetterli, "Optimal Trellis-based Buffered Compression and Fast Approximations," IEEE Trans. on Image Processing, vol.3, no.1, pp.26-40, Jan. 1994.
[28] Y. Ye, S. Ci, and D. Wu, "Cross-layer Optimized Coding Mode Selection for Wireless Video Communications," ICMEW 2012.
[29] Y. Ye, S. Ci, Y. Liu, D. Wu, H. Wang, and A. K. Katsaggelos, "A Wireless Video Surveillance System with an Active Camera," VCIP 2012.
[30] T. Wiegand, M. Lightstone, D. Mukherjee, T.G. Campbell, and S.K. Mitra, "Rate-distortion Optimized Mode Selection for Very Low Bit Rate Video Coding and the Emerging H.263 Standard," CSVT, vol.6, no.2, pp.182-190, Apr. 1996.
[31] E.M. S. Uslubas and A.K. Katsaggelos, "A Resolution Adaptive Video Compression System," Intelligent Multimedia Communication: Techniques and Applications, vol. 280/2010: Springer Verlag, pp.167-194, 2010.
[32] M. Shen, P. Xue, and C. Wang, "Down-Sampling Based Video Coding Using Super-Resolution Technique," CSVT, vol.21,no.6, pp.755-765, June 2011.
[33] R. Zhang, S.L. Regunathan, K. Rose, "Video Coding with Optimal Inter/Intra-Mode Switching for Packet Loss Resilience," JSAC, vol.18, pp. 966-976, Jun. 2000.
[34] Z. He, J. Cai and C.W. Chen, "Joint Source Channel Rate-Distortion Analysis for Adaptive Mode Selection and Rate Control in Wireless Video Coding," CSVT, vol. 12, pp. 511 -523, Jun 2002.
[35] Y. Wang, J. Y. Tham, W. S. Lee, and K. H. Goh, "Pattern selection for error-resilient slice interleaving based on receiver error concealment technique," ICME, pp.1-4, 2011.
[36] R. Chakravorty, S. Banerjee, and S. Ganguly, "MobiStream: Error-Resilient Video Streaming in Wireless WANs Using Virtual Channels," INFOCOM, vol., no., pp.1-14, April 2006.
[37] Y. Wang, J. Y. Tham, W. S. Lee, and K. H. Goh, "Pattern selection for error-resilient slice interleaving based on receiver error concealment technique," ICME, vol., no., pp.1-4, 11-15 July 2011.
[38] B.W. Micallef, C.J. Debono, and R.A. Farrugia, "Performance of enhanced error concealment techniques in multi-view video coding systems," IWSSIP, vol., no., pp.1-4, 16-18 June 2011.
[39] H. A. Aly and E. Dubois, "Image Up-Sampling Using Total-Variation Regularization with a New Observation Model," IEEE Trans. on Image Processing, vol. 14, pp. 1647 -1659, Oct. 2005.
[40] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," PAMI, vol. 24, no. 5, pp. 603–619, May 2002.
[41] C. Stauffer andW. Grimson, "Learning patterns of activity using real-time tracking," PAMI, vol. 22, no. 8, pp. 747–757, Aug. 2000.
[42] S. Babacan and T. Pappas, "Spatiotemporal algorithm for background subtraction," ICASSP 2007, vol. 1, pp. I–1065 –I–1068.
[43] P. Suo and Y. Wang, "An improved adaptive background modeling algorithm based on gaussian mixture model," in ICSP 2008, pp. 1436 –1439.
[44] J. Gallego, M. Pardas, and G. Haro, "Bayesian foreground segmentation and tracking using pixel-wise background model and region based foreground model," ICIP 2009, pp. 3205 – 3208.
[45] J.-W. Hsieh and J.-X. Lee, "Video object segmentation using kernel-based models and spatiotemporal similarity," ICIP 2006, pp. 1821 –1824.

[46] F. Shimin, G. Qing, X. Sheng, and T. Fang, “Human tracking based on mean shift and kalman filter,” AICI 2009, vol. 3, nov.2009, pp. 518 –522.
[47] C. Harris and M. Stephens, “A combined corner and edge detector,” Proceedings of the 4th Alvey Vision Conference, 1988, pp. 147–151.
[48] J.-W. Hsieh, “Fast stitching algorithm for moving object detection and mosaic construction,” ICME 2003 - Volume 2, pp. 85–88.
[49] P. D. Kovesi, “MATLAB and Octave functions for computer vision and image processing,” Centre for Exploration Targeting, School of Earth and Environment, The University of Western Australia, available from: <http://www.csse.uwa.edu.au/_pk/research/matlabfns/>.
[50] Y. Jin, L. Tao, H. Di, N. Rao, and G. Xu, “Background modeling from a free-moving camera by multi-layer homography algorithm,” ICIP 2008, pp. 1572 –1575.
[51] C. Li, C. Xu, C. Gui, and M. Fox, “Level set evolution without re-initialization: a new variational formulation,” CVPR 2005, vol. 1, pp. 430 – 436 vol. 1.
[52] Z. Kim, “Real time object tracking based on dynamic feature grouping with background subtraction,” CVPR 2008, pp. 1 –8.
[53] Y. Sheikh, O. Javed, and T. Kanade, “Background subtraction for freely moving cameras,” ICCV 2009, pp. 1219 –1225.
[54] Z. Han, H. Wang, D. O. Wu, J. Huang, and M. Van Der Schaar, "Cross-Layer Optimized Wireless Multimedia Communications," Advances in Multimedia, vol.2007, pp.1-2, 2007.
[55] F. Pan, X. Lin, S. Rahardja, K. P. Lim, Z. G. Li, D. Wu, and S. Wu, "Fast Mode Decision Algorithm for Intra-prediction in H.264/AVC Video Coding," CSVT, vol.15, no.7, pp. 813- 822, July 2005.
[56] I. Choi, J. Lee, and B. Jeon, "Fast Coding Mode Selection With Rate-Distortion Optimization for MPEG-4 Part-10 AVC/H.264," CSVT, vol.16, no.12, pp.1557-1561, Dec. 2006.
[57] D. Krishnaswamy and M. van der Schaar, "Adaptive Modulated Scalable Video Transmission over Wireless Networks with a Game Theoretic Approach," 6th IEEE Workshop on Multimedia Signal Processing, pp.107-110, 2004.
[58] Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems, IEEE Standard 802.16, 2002.
[59] Z. Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," IEEE Trans. on Image Processing, vol.13, no.4, pp.600-612, Apr. 2004.
[60] H.264/AVC Reference Software, available online: http://iphome.hhi.de/suehring/tml/download/
[61] Y. Ye, S. Ci, D. Wu, H. Wang, A. K. Katsaggelos, "Cross-layer Design and Optimization for Video Surveillance Systems," E-Letter of the Multimedia Communications Technical Committee (MMTC), IEEE Communications Society, Special Issue on Cross-Layer Design and Optimization for Video Surveillance Systems, vol. 7, no. 4, April 2012.
[62] D. Wu, S. Ci, H. Luo, Y. Ye, H. Wang, "Video Surveillance Over Wireless Sensor and Actuator Networks Using Active Cameras," IEEE Trans. Automat. Contr. 56(10): 2467-2472, 2011.
[63] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," TPAMI, vol.25, no.5, pp. 564- 577, May 2003.
[64] C. Yang, R. Duraiswami, and L. Davis, "Efficient mean-shift tracking via a new similarity measure," CVPR2005, pp. 176- 183.
[65] T. Dang, C. Hoffmann, and C. Stiller, "Fusing optical flow and stereo disparity for object tracking," in Proc. 5th Intl. IEEE Conf. Intelligent Transportation Systems 2002. pp. 112- 117.
[66] G. Mohammadi, F. Dufaux, T. H. Minh, and T. Ebrahimi, "Multi-view video segmentation and tracking for video surveillance," in SPIE Mobile Multimedia Image Processing, Security and Applications, April 2009.
[67] C. Park and K.-H. Bae, “Quasi-Feature based Panoramic Video Creation for Multiview Object Tracking System,” International Journal of Advanced Robotic Systems, 2009.
[68] J. G. Lou, H. Cai, and J. Li, “A real-time interactive multi-view video system,” in Proc. 13th ACM Intl. Conf. Multimedia, Singapore, Nov.2005, pp. 161-170.
[69] A. Vetro, A.M.Tourapis, K. M ¨ uller, T. Chen, “3D-TV Content Storage and Transmission”, IEEE Transactions on Broadcasting, Vol. 57, No. 2, pp. 384-394, June 2011.
[70] C. Fehn, “A 3D-TV system based on video plus depth information”, Proc. of the 37th Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 1529-1533, November 2003.

[71] P. Merkle, Y. Wang, K. M ¨ uller, A. Smolic, and T. Wiegand, “Video plus Depth Compression for Mobile 3D Services”, Proc. of IEEE 3DTV Conference 2009.
[72] W. Zou, “An Overview for Developing End-to-End Standards for 3-D TV in the Home”, Information Display, Vol. 25, No.7, July 2009.
[73] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. Dodgson, “Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid,” ECCV2010, pp. 510-523.
[74] A. Geiger, M. Roser and R. Urtasun, “Efficient Large-Scale Stereo Matching”, ACCV2010, pp. 25-38.
[75] C.L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," TPAMI, vol.22, no.7, pp.675-684, Jul 2000.
[76] O. Woodford, P. Torr, I. Reid and A. Fitzgibbon, “Global stereo reconstruction under second-order smoothness priors,” CVPR2008, pp.1-8.
[77] J. Sun, H.-Y. Shum, and N.-N. Zheng. “Stereo matching using belief propagation,” TPAMI, 25(7):787–800, 2003.
[78] B.M. Smith, Li Zhang, and Hailin Jin, "Stereo matching with nonparametric smoothness priors in feature space," CVPR2009, pp.485-492.
[79] J.R. Shewchuk, In: Applied Computational Geometry: Towards Geometric Engineering. vol. 1148. Springer, Berlin (1996) pp. 203-222.
[80] J. Besag, "On the statistical analysis of dirty pictures," J. R. Stat. Soc. B, vol. 48, pp. 259-502, 1986.
[81] L. Evans, Partial Differential Equations, Providence: American Mathematical Society, 1998.
[82] S. Kosov, T. Thormahlen, and H. P. Seidel, “Accurate Real-Time Disparity Estimation with Variational Methods,” ISVC '09 Proceedings of the 5th International Symposium on Advances in Visual Computing: Part I, pp. 796-807.
[83] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” IJCV, 47(1-3):7–42, 2002.
[84] D. M. Young, “Iterative Solution of Large Linear systems,” New York: Academic, 1971.
[85] S. Liu and C. W. Chen, “3D Video Transcoding for Virtual Views”, ACM Multimedia, Florence, Italy, pp. 795-798, Oct. 2010.
[86] A. Vetro, J. Xin, and H. Sun, “Error resilience video transcoding for wireless communications,” IEEE Wireless Commun., vol. 12, no. 4, pp. 14-21, Aug. 2005.
[87] Y. Liu, Q. Huang, S. Ma, D. Zhao, W. Gao, “Joint Video/Depth Rate Allocation for 3D Video Coding based on View Synthesis Distortion Model”, Signal Processing: Image Communication, vol.24, no.8, pp. 666- 681, 2009.
[88] Y. Guo, H. Li, and Y.-K. Wang, “SVC/AVC loss simulator donation”, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6 JVT-Q069, October 2005.
[89] Y. Ye, S. Ci, Y. Liu, H. Wang, A.K. Katsaggelos, “Binocular Video Object Tracking with Fast Disparity Estimation,”, AVSS 2013.
[90] Y. Liu, S. Ci, H. Tang, Y. Ye , "A Transcoding Framework with Error-Resilient Video/Depth Rate Allocation for Mobile 3D Video Streaming," The IEEE International Conference on Communications 2012.
[91] Y. Liu, S. Ci, H. Tang, Y. Ye, and J. Liu, "QoE-oriented 3D Video Transcoding for mobile streaming," ACM Transactions on Multimedia Computing, Communications and Applications (TOMCCAP), Volume 8, Issue 3s, article 42, 2012.
[92] Y. Liu, S. Ci, H. Tang and Y. Ye, "Application-adapted Mobile 3D Video Coding and Streaming, A Survey," 3D Research Journal, Volume 3, Issue 1, pp.1-6, 2012, Springer.
[93] J.S. Rietman, K Postema, J.H. Geertzen, “Gait analysis in prosthetics (opinions, ideas and conclusions),” Prosthet Orthot Int., vol.26, pp.50–57, 2002.
[94] L. Ballan and G. M. Cortelazzo, “Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes,” 3DPVT, 2008.
[95] N. Hasler, B. Rosenhahn, T. Thormahlen , M. Wand, J. Gall, H.-P. Seidel, “Markerless motion capture with unsynchronized moving cameras,” CVPR, 2009.
[96] Z. Zhang, “Flexible camera calibration by viewing a plane from unknown orientations,” ICCV 1999.
[97] P. Perona and J. Malik, “Scale-space and edge detection using anisotropic diffusion,” PAMI, vol. 12, pp. 629–639, 1990.
[98] L. Evans, “Partial Differential Equations,” Providence: American Mathematical Society, 1998.

[99] Y. Boykov and V. Kolmogorov, “An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision,” PAMI, vol. 26, no. 9, pp. 1124 –1137, 2004.

[100] D. Wu, H. Luo, S. Ci, H. Wang, and A. Katsaggelos, “Quality-Driven Optimization for Content-Aware Real-Time Video Streaming in Wireless Mesh Networks,” GLOBECOM, Dec. 2008.

[101] J. Abate, G. L. Choudhury, and W. Whitt, “Exponential approximations for tail probabilities in queues I: Waiting times,” Oper. Res., vol. 43, no. 5, pp. 885–901, 1995.

[102] D. Wu, S. Ci, H. Wang, and A.K. Katsaggelos, “Application-centric routing for video streaming over multi-hop wireless networks,” CSVT, vol. 20, no. 12, pp. 1721-1734, Dec. 2010.

[103] J. Liu, B. Li, A. T. S. Ip, Y. Zhang, “Dynamic Simulcasting: Design and Optimization,” Combinatorial Optimization in Communication Networks Combinatorial Optimization, Volume 18, pp 565-594, 2006.

[104] D. Pisinger, “A minimal algorithm for the multiple-choice knapsack problem,” European Journal of Operational Research, 83 (1995), 394-410, 1995.

[105] M. Pollefeys, L. Van Gool, A. Zisserman, A. Fitzgibbon (Eds.), “3D Structure from Images,” LNCS, Vol. 2018, Springer-Verlag, 2001.

[106] M.D. Chang, E. Sejdic, V. Wright, T. Chau, “Measures of dynamic stability: detecting differences between walking overground and on a compliant surface,” HumMov Sci, 29:977–86, 2010.

[107] Y. Ye, S. Ci, A. K. Katsaggelos and Y. Liu, "A Multi-camera Motion Capture System for Remote Healthcare Monitoring," The IEEE International Conference on Multimedia and Expo, July 2013.