Multiview video coding for 3D telepresence

Yang, Wenxian

2006

Yang, W. X. (2006). Multiview video coding for 3D telepresence. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/2466>

<https://doi.org/10.32657/10356/2466>

Nanyang Technological University

Downloaded on 26 Apr 2023 22:53:16 SGT

MULTIVIEW VIDEO CODING FOR 3D TELEPRESENCE

A thesis submitted to
the Nanyang Technological University
in fulfilment of the requirement for the degree of
Doctor of Philosophy of Computer Engineering

by

WENXIAN YANG

Supervisor: Dr. JIANFEI CAI

Division of Computer Communications
School of Computer Engineering
Nanyang Technological University
Singapore

October 2005

Dedication

*To my father Guangrong Yang,
my mother Yuyun Xiao,
my brother Xinyao Yang,
and my husband Feng Ding —
for their love.*

Acknowledgement

I would like to express my sincere gratitude to my supervisors, Professor King Ngi Ngan and Assistant Professor Jianfei Cai, for their professional guidance, invaluable assistance, consistent encouragement and patience throughout my Ph.D. study, as well as their constructive comments on this thesis.

I am very grateful to Professor Kwanghoon Sohn of Yonsei University in Korea, for his guidance and insightful discussions during my second year of Ph.D. study, and also to Jeongeun Lim for the comments and advice on various aspects of this work.

My sincere thanks to my mentors Dr. Yan Lu and Dr. Feng Wu at Microsoft Research Asia, for their helpful suggestions and encouragements during my stay in MSRA as a visiting student.

It is my great pleasure to thank and acknowledge the staff and technicians in the Center of Multimedia and Network Technology (CeMNeT) in the School of Computer Engineering (SCE) in Nanyang Technological University (NTU). Without their continuous support in various aspects, I could not have finished my Ph.D. study. Thanks also goes to all the staff in SCE for their help and support. I appreciate the graduate students in our group and in CeMNeT for their encouragements and camaraderie.

Finally, I am eternally indebted to my family members for their love and support, which words cannot describe.

Abstract

Telepresence refers to a human/machine system in which the human operator feels physically present at a remote site. Telepresence systems have high potential in various applications of consumer entertainment.

To achieve telepresence, stereoscopic and multiview video that can give the viewers the 3D feeling are required. The major obstacle for telepresence systems is the extremely large amount of data associated with it. Efficient compression of multiview video sequences has become an integral and important part of any practical telepresence system. To code the multiview video sequences efficiently, both the view redundancy among the sequences and the temporal redundancy inside each sequence need to be exploited and reduced, which can be achieved using disparity and motion compensation, respectively. The coding efficiency can be further improved by jointly estimating and optimizing motion and disparity fields.

In this thesis, we develop several novel algorithms and two complete encoders for multiview video coding (MVC). The major contributions can be summarized in three aspects. In particular, we first propose an edge-preserving regularization scheme to calculate either 1D disparity fields or 2D motion fields. After confirming its performance by comparing with the existing algorithms, the separate regularization scheme is extended to a joint estimation scheme that calculates two disparity fields and two motion fields for two successive image pairs simultaneously.

Secondly, we develop an MPEG-4 compatible multiview video encoder which integrates with the joint disparity and motion estimation scheme. Besides, various aspects of the encoder are investigated, including a comparative study of several view-level frame prediction structures and the rate control algorithm for multiview video coding.

Thirdly, we propose a framework of scalable MVC using 4D wavelet. The wavelet-

based multiview video coding system is more flexible and extendable, and it can provide temporal, spatial, SNR (Signal-to-Noise-Ratio) and view scalabilities. In the wavelet MVC scheme, general lifting structure is introduced, based on which we propose a disparity compensated view filtering (DCVF) for wavelet decomposition along the view direction. We also propose a flexible decomposition structure based on the analysis of the temporal correlation and the view correlation. In addition, we develop the entire codec, including the design of the macroblock coding modes, subband coding, rate control, etc. The proposed scheme overcomes a variety of limitations exhibited by existing methods and provides a natural and elegant solution to MVC.

In our work, extensive experiments are carried out using both standard and non-standard multiview video sequences and the results are compared with existing systems.

List of Figures

2.1	General camera setup and the epipolar geometry.	8
2.2	Parallel camera setup and the definition of disparity.	9
2.3	Different camera configurations, (a) 1D parallel, (b) 2D parallel, (c) convergent and (d) 1D arc.	10
2.4	Coherent disparity and motion analysis.	21
2.5	Frame structure in the MPEG standards	24
2.6	Wavelet decomposition with motion trajectory for an image sequence.	27
3.1	MultiCam system for multiview video capturing.	36
3.2	Original left view of test stereoscopic/multiview image pairs or video sequences (the first frame), (a) PIANO, (b) MAN, (c) TRAIN & TUNNEL and (d) Read- ing.	38
3.3	Two successive image pairs and corresponding blocks/pixels.	46
3.4	Block- and pixel-based disparity maps by the existing and the proposed algorithms at $\lambda = 10$, PIANO.	51
3.5	Block- and pixel-based disparity maps by the existing and the proposed algorithms at $\lambda = 10$, MAN.	52

3.6	Comparison of block-based motion fields by the existing and the proposed algorithms at $\lambda = 10$	54
3.7	Comparison of proposed the separate and the joint regularization algorithms, (a) TRAIN & TUNNEL and (b) MAN.	55
4.1	Two configurations of the MPEG-2 Multiview Profile.	60
4.2	The block diagram of the MPEG-4 compatible stereoscopic video encoder.	62
4.3	GOP structure of the proposed stereoscopic video encoder.	62
4.4	View-level prediction structures for 5-view video coding.	65
4.5	Frame-level prediction structures for 5-view video coding.	66
4.6	Coding results of stereoscopic video sequence, (a) TRAIN & TUNNEL, (b) MAN and (c) Reading.	70
4.7	Decoded auxiliary view image for MAN (a) and (b) by the proposed scheme, (c) and (d) by MPEG-4 MVP.	71
4.8	Decoded auxiliary view image for Reading (a) the proposed scheme and (b) MPEG-4 MVP.	71
4.9	Coding results of stereoscopic video sequence, (a) TRAIN & TUNNEL and (b) MAN.	72
4.10	Comparison of view-level prediction structures for 5-view video coding, (a) TRAIN & TUNNEL and (b) Reading.	73
5.1	(a) <i>Prediction</i> and (b) <i>Updating</i> for DCVF.	80
5.2	High-pass subbands from the one-level decomposition of the view sequence Golf2, using (a) DCVF with GDC only, (b) DCVF and (c) MCTF.	82

5.3	Different decomposition structures along the temporal and the view directions, (a) 4T+2V and (b) 3T+V+T+V.	85
5.4	The block diagram of the proposed high dimensional scalable multiview video coding scheme.	88
5.5	Six macroblock modes for motion (disparity) coding.	89
5.6	Scan orders of a temporal-view subband, (a) T-scan, (b) V-scan, (c) head-to-tail scan for MCTF-decomposed subband, and (d) head-to-tail scan for DCVF-decomposed subband	91
5.7	The first frame of the left-most view of the original multiview video sequences (a) Flamenco1, (b) Golf2 and (c) Race1.	93
5.8	Comparison of coding results using the proposed scheme, H.264/AVC simulcast, and SVC simulcast for (a) Flamenco1, (b) Golf2 and (c) Race1.	94
5.9	Performance evaluation of DCVF (a) Flamenco1, (b) Golf2 and (c) Race1.	98
5.10	Subjective image quality evaluation, decoded image by (a) DCVF and (b) MCTF for Golf2 view sequence at 128kbps.	99
5.11	Multi-dimensional decomposition of multiview video data.	99

List of Tables

2.1	Video coding standards developed by various organizations.	25
3.1	Description of test multiview video sequences.	37
3.2	Comparison of the proposed separate regularization algorithm with exist- ing algorithms, on block-based disparity for PIANO.	53
3.3	Computational complexity analysis of the proposed algorithms.	55
4.1	VOP types and MB prediction types.	63
5.1	$HCost$ for interleaved decomposition structure.	86
5.2	Weight matrix for global rate-distortion optimization truncation.	92
5.3	Parameters for test multiview video sequences.	93
5.4	AVC Parameters for monoview video coding.	93
5.5	Comparison of coding performance of the three different decomposition structures.	97

List of Abbreviations

Abbreviations	Details
AVC	Advanced Video Coding
BBM	Block-based Matching
DCD	Disparity Compensated Difference
DCT	Discrete Cosine Transform
DCVF	Disparity Compensated View Filtering
DP	Dynamic Programming
DPCM	Differential Pulse Code Modulation
DWT	Discrete Wavelet Transform
EDU	Energy Distributed Update
EOD	Estimation of Dependency
ESCOT	Embedded Subband Coding with Optimal Truncation
FFRGMET	Feature-based Fast and Robust Global Motion Estimation Technique
FSBM	Fixed-Size Block Matching
FTV	Free-Viewpoint Television
FVV	Free-Viewpoint Video
GDC	Global Disparity Compensation
GDV	Global Disparity Vector
GGOP	Group of GOPs
GMC	Global Motion Compensation
GMV	Global Motion Vector
GOB	Group Of Blocks
GOP	Group Of Pictures
HBM	Hierarchical Block Matching
HHM	Hierarchical Hybrid Matching
HVS	Human Visual System
ISO	International Standards Organization
ITU	International Telecommunication Union
IVR	Intermediate Views Reconstruction
JVT	Joint Video Team
LDC	Local Disparity Compensation
MAC	Multiple Auxiliary Component

Abbreviations	Details
MAD	Mean Absolute Difference
MAP	Maximum A Posterior
MATT	Motion Aligned Temporal Transform
MB	Macroblock
MC	Motion Compensation
MCP	Motion Compensated Prediction
MCTF	Motion Compensated Temporal Filtering
MFA	Mean Field Annealing
MPEG	Moving Picture Experts Group
MRF	Markov Random Field
MSE	Mean Square Error
MVC	Multiview Video Coding
MVP	Multiview Profile
PDC	Pixel Differences Classification
PSNR	Peak Signal-to-Noise Ratio
R-D	Rate-Distortion
RLC	Run-Length Coding
RMSE	Root of Mean Square Error
SNR	Signal-to-Noise Ratio
SVC	Scalable Video Coding
VLC	Variable Length Coding
VOP	Video Object Plane

Contents

ACKNOWLEDGEMENT	ii
ABSTRACT	iii
LIST OF FIGURES	v
LIST OF TABLES	viii
LIST OF ABBREVIATIONS	ix
1 Introduction	1
1.1 Research Background	1
1.2 Summary of Contributions	4
1.3 Organization of the Thesis	5
2 Literature Review	7
2.1 Basics of Stereoscopic and Multiview Video	7
2.1.1 Stereoscopic Imaging	7
2.1.2 Characteristics of Multiview Video	10

2.1.3	Display Systems	11
2.2	Disparity Estimation	12
2.2.1	Differences between Disparity and Motion	12
2.2.2	Disparity Estimation Algorithms	14
2.3	Joint Estimation	19
2.4	Digital Video Coding	22
2.4.1	Fundamentals	23
2.4.2	Developments and Standards	24
2.4.3	Scalable Video Coding	26
2.4.4	Performance Measures	28
2.5	Stereoscopic and Multiview Video Coding	29
2.5.1	Existing Techniques	29
2.5.2	Stereoscopic and Multiview Rate Control	32
3	Joint Estimation and Regularization of Disparity and Motion	34
3.1	MultiCam System and Video Capturing	35
3.1.1	MultiCam System	35
3.1.2	Test Video Sequences	35
3.2	Balancing and Global Disparity Estimation	37
3.2.1	Image Balancing	37
3.2.2	Global Disparity Estimation	39
3.3	Edge-Preserving Disparity Regularization	40

3.3.1	The MRF Model and Problem Formulation	40
3.3.2	The Energy Function	41
3.3.3	Determine the Lagrange Multiplier	43
3.3.4	A Fast Implementation	44
3.4	Joint Disparity/Motion Estimation and Regularization	45
3.4.1	The Global Cost Function	45
3.4.2	The Implementations	47
3.5	Experimental Results and Analysis	49
3.5.1	Evaluation of Disparity Regularization	50
3.5.2	Evaluation of Joint Disparity and Motion Estimation	55
4	MPEG-4 Compatible Multiview Video Coding	58
4.1	Redundancies in Multiview Video Data	58
4.2	MPEG-2 Multiview Profile	59
4.3	Encoder Structure	61
4.4	Optimal Reference Selection	64
4.4.1	View-level Reference Selection	64
4.4.2	Frame-level Reference Selection	66
4.5	Rate Control	66
4.6	Experimental Results and Analysis	68
4.6.1	Evaluation of Stereoscopic Video Coding	68
4.6.2	Evaluation of Multiview Video Coding	72

5	Scalable Multiview Video Coding Using 4D Wavelet	74
5.1	Lifting based Disparity Compensated View Filtering	76
5.1.1	Generalized Lifting Transform	76
5.1.2	Global Warping-based DCVF	78
5.1.3	Adaptive Global and Local Disparity Model in DCVF	81
5.2	Adaptive Decomposition Structure	83
5.2.1	Regular 4D Wavelet Decomposition	83
5.2.2	Adaptive Decomposition Structure	85
5.3	Subband Multiview Video Coding	86
5.3.1	System Structure	87
5.3.2	MB Modes	87
5.3.3	Subband Coding	90
5.4	Experimental Results and Analysis	92
5.4.1	Evaluation of Overall Coding Performance	92
5.4.2	Evaluation of Decomposition Structure	96
5.4.3	Evaluation of DCVF	97
5.4.4	Discussion of Scalabilities	100
6	Conclusions	102
6.1	Summary	102
6.2	Future Directions	104
	BIBLIOGRAPHY	106

PUBLICATIONS	120
---------------------	------------

VITA	123
-------------	------------

Chapter 1

Introduction

1.1 Research Background

In 1992, Sheridan [1] defined “telepresence” as a human/machine system in which the human operator receives “sufficient information about the teleoperator and the task environment, displayed in a sufficiently natural way, that the operator feels physically present at the remote site.” There is a large potential in telepresence systems for broadcast and non-broadcast applications and in the consumer entertainment market. The application areas include 3D visualization, remote operations, 3D telemedicine, 3D telerobotics, FVV (free-viewpoint video)/FTV (free-viewpoint television), 3D cinema and television, immersive teleconference and surveillance, and virtual reality, etc. [2, 3]. For example, telepresence extends the videoconferencing concept so that participants can use non-verbal aspects of communication, such as eye contact, body movement, gestures, facial expressions, in the same way as they would in a face-to-face meeting. Many other applications in situations where humans are exposed to hazardous situations are also suitable candidates for telepresence. The major aspects of telepresence systems include disparity estimation, stereoscopic image pair coding, multiview video coding, intermediate views reconstruction (IVR), and the design and implementation of display systems, etc. [4]. In recent years, many researchers working in this field have presented various algorithms for

these aspects and systems have been built up.

Stereoscopic or general multiview video data is normally used to achieve the 3D feeling for telepresence systems. In visual communications, they provide a simple means of perceiving the relative depth information in a real world scene from two or more suitably obtained images of that scene. Stereoscopic and multiview video display apparatus are also being technically and commercially available. The major obstacle for using multiview video is the huge amount of data, which makes video storage and transmission extremely difficult. The amount of raw data of multiview video is that of monoview video multiplied by the number of views. For example, the raw data of an 8-view video sequence at 30 frames per second (fps) with the image size of 320×240 pixels in YUV 4:2:0 format is more than 220 mega bits per second (Mbps). Efficient compression of multiview video data has become an integral and important part of any practical telepresence system. For monoview video compression, the *Moving Picture Experts Group* (MPEG) of the *International Standards Organization* (ISO) has worked out a number of international standards such as MPEG-1 [5], MPEG-2 [6] and MPEG-4 [7]. The joint work of the MPEG and the *International Telecommunication Union* (ITU), MPEG-4 part 10, which is also known as H.264/AVC (Advanced Video Coding), represents the state-of-the-art video compression technique. Recently, based on the evidence brought forward on multiview video coding technology, the MPEG organization is working towards a new standard on multiview video coding (MVC) [8].

The key to multiview video compression is to exploit the redundancy contained in the video data. Since the multiview video data is captured at different viewing angles at the same time, there exists significant amount of redundancy within it, including the temporal redundancy along each temporal sequence, the view redundancy within view-adjacent images and the spatial redundancy within each image. However, the existing MVC systems can hardly fully utilize the various kinds of redundancies and efficiently

encode the multiview video data because the inter-frame redundancy cannot be efficiently removed.

Disparity is the name of motion along the view axis. Disparity estimation has been one of the most fundamental problems in stereoscopic image processing and a crucial building block in many visual communication applications. Most existing MVC schemes remove the temporal and the view redundancies independently using a multi-reference coding structure. However, in most natural video sequences, motion vectors are usually smaller than the disparity vectors and references from the temporal direction are preferably selected. Thus the view-directional redundancy cannot be fully utilized. In addition, the stereo-motion consistency constraint (or the “loop” constraint) provides a more robust way for calculating the two motion fields and the two disparity fields associated with two adjacent image pairs simultaneously, but is not fully exploited in most existing MVC schemes.

The other drawback of existing systems is the lack of flexibilities and scalabilities, e.g., compatibility with existing monoview video coding standards, extendability to encode an arbitrary number of views, look-around capability at the viewer side and various kinds of scalabilities. In particular, one example is the object-based schemes using 3D wireframe models, which are not compatible with the existing coding standards. In addition, since most real world video sequences are too complex to model, such object-based algorithms are only presented for videoconferencing systems with generic wireframe models for head-and-shoulder images. Another example is that a conventional stereoscopic image pair provides only the 3D view of a scene from one particular viewpoint. In order to create a look-around effect or to accommodate multiple viewers simultaneously, multiple viewing zones and the corresponding multiview images are needed. Intermediate views, depending on the position of the viewer, have to be synthesized. Moreover, scalability has become a more and more important feature for video coding and communications. Conventional

multiview video coding schemes can hardly achieve temporal, spatial, quality and view scalabilities at the same time.

1.2 Summary of Contributions

Although multiview video provides the sense of presence at a remote site, the overall demand for stereoscopic viewing, over time and across viewers in a broadcast type application, may never be high enough to warrant the encoding of all views of multiview sequences at a high quality and resolution [4]. As discussed, current telepresence systems suffer from high bandwidth, poor quality and limited functionalities. The objective of this thesis is to explore efficient algorithms for multiview video coding and build up multiview video encoders for telepresence applications. To encode the sequences into an acceptable and practical data volume, the compression method should take advantages of both the view redundancies and the temporal redundancies. The characteristics of stereoscopic and multiview video data have to be exploited. In addition, the overall encoding scheme should have a moderate computational complexity.

This thesis has made several contributions to the research of multiview video data processing and compression, including an edge-preserving disparity regularization algorithm, joint disparity and motion estimation, an MPEG-4 compatible stereoscopic/multiview video encoder and a scalable multiview video codec using 4D wavelet.

The most important component of a multiview video encoder is the disparity and motion estimation, thus we first propose an edge-preserving regularization scheme to calculate either 1D disparity fields or 2D motion fields. Based on the Markov Random Field (MRF) theory, an energy function that considers the edge values of the first order neighborhood system is proposed. The proposed regularization algorithm works under the rate-distortion theory using a Lagrange multiplier, i.e., the displacement fields can be

calculated up to a certain bit rate. Based on the edge-preserving regularization algorithm, the loop constraint and the maximum a posterior (MAP) probability, we propose a joint estimation scheme that calculates the disparity and motion fields simultaneously. A global energy function is proposed and the solution is approximated by iterative local optimizations.

Secondly, we integrate the joint estimation scheme in an MPEG-4 compatible stereoscopic and multiview video encoder. In video coding, the dominating factor for improving the compression efficiency is the motion compensation for the inter-predicted images. The joint estimation of disparity and motion can provide more robust and smooth vector fields at lower computational complexity and improve the coding results significantly compared with two conventional schemes. We also extend the MPEG-4 rate control algorithm for multiview video coding. For more than three views, different reference structures are proposed and compared since the reference structure also has a significant impact on the coding results for certain video contents.

Thirdly, a more flexible MVC scheme that can provide various kinds of scalabilities is proposed using 4D wavelet, which perfectly matches the inherent structure of high dimensional multiview video data. The wavelet based MVC scheme can adaptively exploit any existing correlations among the temporal and the view directions using combined disparity compensated view filtering (DCVF) and motion compensated temporal filtering (MCTF). It overcomes a variety of limitations exhibited by existing methods and provides a natural and elegant solution to MVC.

1.3 Organization of the Thesis

The rest of the thesis is organized as follows. In Chapter 2, we first give a brief review of the basics of stereoscopic and multiview video, including the stereoscopic imaging, the

characteristics of multiview video and the display systems. After that, several aspects of telepresence systems are discussed in details, including various techniques for disparity estimation and regularization, joint disparity and motion estimation algorithms, etc. Digital video coding techniques are also reviewed. Based on a clear view and understanding of the existing techniques, we propose an edge-preserving disparity regularization algorithm in Chapter 3. A joint disparity and motion estimation scheme is also proposed to estimate the disparity fields and the motion fields in a coherent way. The experimental results are then analyzed and compared with conventional methods. In Chapter 4, we propose an MPEG-4 compatible stereoscopic and multiview video encoder, which embeds the joint disparity and motion estimation module. Rate control for multiview sequences and optimal prediction structure are also studied. A wavelet-based MVC scheme is then proposed in Chapter 5. Disparity compensated lifting is studied and the adaptive decomposition structure is proposed. Finally, we discuss the limitations and make some conclusions in Chapter 6. Some future directions of this research are also provided.

Chapter 2

Literature Review

In this chapter, some fundamental concepts about stereoscopic and multiview video, including the camera geometry, the characteristics of multiview video and the display systems are introduced in Section 2.1. In Section 2.2, the existing techniques for disparity estimation are summarized. Joint disparity and motion estimation algorithms are then discussed in Section 2.3. We review the developments in digital video coding including the standards and other coding techniques in Section 2.4. Finally, the stereoscopic and multiview video coding techniques are summarized in Section 2.5.

2.1 Basics of Stereoscopic and Multiview Video

2.1.1 Stereoscopic Imaging

Stereopsis is defined as the ability to perceive three-dimensional depth. The relative depth information in the scene being viewed is deduced by the human brain, based on the horizontal shift between corresponding points in the 2D projections of a 3D scene on the left and the right retinae. The process of acquiring two such 2D projections is known as *stereoscopic imaging*. It imitates the human optical system in the sense that the two 2D projections are obtained by imaging the 3D scene onto two appropriately positioned

Figure 2.1: General camera setup and the epipolar geometry.

imaging sensors [9].

The relative positions and orientations of the two imaging elements and the two imaging sensors in a stereoscopic imaging setup constitute the stereoscopic imaging geometry. As shown in Figure 2.1, C and C' are the optical centers of the two cameras. Given a point m in the left image, its corresponding point in the right image is constrained to lie on a line called the *epipolar line* of m , denoted by l_m . The line l'_m is the intersection of the *epipolar plane*, defined by M , C and C' , with the right image plane R [10]. In case the two optical axes are parallel, the epipolar lines become the horizontal scan lines in the images, which means only horizontal disparity exists.

As shown in Figure 2.2, the *disparity* is defined as the relative displacement between the left and the right image points (x_l, y_l) and (x_r, y_r) belonging to the same object point (x, y, z) . For parallel camera setup, given the camera parameters, the disparity value can be calculated as:

$$d = \frac{b}{z} \cdot f \quad (2.1.1)$$

Here, d is the disparity, b is the baseline distance between the two cameras, z is the object depth and f is the focus length of the cameras.

To simplify the problem, parallel camera setup is assumed in most research works [11, 12, 13]. In applications where the cameras have to be set convergent, the images are rectified such that they look like being taken by two parallel cameras. In parallel camera setup, the baseline decides the extent of occlusion and determines the search range in disparity estimation. Long baseline provides a wide viewing angle but suffers from photometric distortion and severe occlusion. Actually, the baseline should correspond to the small distance between the human eyes [13] but always much larger due to implementation limitations. Therefore, intermediate view images with a smaller baseline will have to be synthesized for stereoscopic video sequences.

Figure 2.3: Different camera configurations, (a) 1D parallel, (b) 2D parallel, (c) convergent and (d) 1D arc.

2.1.2 Characteristics of Multiview Video

However, two views may not be sufficient to achieve the real telepresence illusion and create a look-around capability. The reason is twofold. Firstly, the occlusion problem cannot be handled successfully for two views since some points in one view may not have correspondence in the other. Disparity estimation can be more reliable with increasing number of views. Secondly, multiple viewpoints are needed at the receiver in some applications. Thus, usually three or more views are needed as input of the telepresence system.

Different camera configurations are considered in different applications using multiple views and some examples are depicted in Figure 2.3, including 1D parallel, 2D parallel, 1D arc and convergent [8]. Three cameras are commonly arranged as an equilateral. Five cameras are used in [11] where one camera is placed at the center and the other four are placed in the directions of upper, lower, right and left to the central one symmetrically,

with the same distance in between. The advantage of this configuration is that at least one of the two symmetrically located cameras will not be influenced by occlusion and can provide true data for a specific position. In [12], nine cameras are placed as a 3×3 matrix, and the depth map is estimated using the eight stereo pairs composed of the center image with another outside camera. In the multiocular system of [14], the optical axes of the N cameras are set parallel and the baseline distance between every two neighboring cameras is equal. Thus, the exact number of views where each point is (or is not) occluded can be determined to avoid errors in disparity estimation.

2.1.3 Display Systems

Stereoscopic video display system is also an important component of the telepresence systems. Although it is not a part of our research work, we give a brief introduction here.

When viewing the stereoscopic image pair, if the observer focuses the right eye on the right image and the left eye on the left image, the depth information can be perceived and a 3D image will be seen. Based on this property, various types of stereoscopic displays have been presented for entertainment and communications, including the head-mounted displays, the autostereoscopic displays and the object-tracking displays.

Head-mounted displays requiring the user to wear special glasses have been in use for several years. However, these systems suffer from uncomfortable eye-wear, control wires, cross talk and other image degrading effects such as image flicker and reduced image brightness. *Autostereoscopic* displays that require no viewing aids are regarded as a more natural long-term solution to 3D display products, but such displays have inherently limited viewing freedom. Loss of image quality in these displays typically arises from inadequate display bandwidth (which results in loss of image and depth resolution) and cross talk between image channels due to scattering and aberrations of the optical system.

Viewing freedom in 3D displays can be enhanced using *observer tracking* displays. In case of the demand of a look around capability, the image must be modified smoothly to reflect the change in viewing geometry as the observer moves. The maximized resolution and straightforward interface of observer tracking systems means that these are the most promising candidates for commercial 3D display with wide viewing freedom [15]. Conventional autostereoscopic displays cannot produce images for more than one or a few viewers at a time. If the viewer moves side-to-side into a virtual window, or if the eye moves side-to-side across a boundary of virtual window, the scene will not change corresponding to user position. Yoshihiro [16] proposed to solve this problem by super multiview concept, i.e., several views are produced by different projectors at one time and each view can only see into a virtual window. With such display, multiple viewers can see 3D images at one time as well as different scenes from various positions. A 3D display system capable of producing 45-view video images at an angular interval of 0.5 degrees at a time is developed in [16] from which multiple viewers can see 3D video images from any position in the viewing zone at one time. Furthermore, when the viewer moved around he or she could see different scenes corresponding to the eye position. Flipping of parallax could not be eliminated completely but could be reduced.

As long as display systems are technically supported and commercially available, stereoscopic and multiview video display is the last issue that we concern in this research.

2.2 Disparity Estimation

2.2.1 Differences between Disparity and Motion

To remove the spatial redundancy of the stereo pair, the disparity map has to be calculated. Although motion and disparity are quite similar since they are both vector fields describing the displacement between pixel pairs, there still exist salient differences in

their properties and thus estimation algorithms, as listed below:

1. In many cases (with the exception of fast motion) the similarity between subsequent frames within one view is much higher than the similarity between the different views, i.e., the temporal correlation is usually stronger than the view correlation. Thus, the motion compensated inter-frame prediction is most likely preferred over the disparity compensated inter-view prediction.
2. Theoretically, the disparity field of the stereoscopic video sequence needs to be calculated only once if the cameras are stationary. The later disparity fields can be calculated by the motion fields and the previous disparity field using the loop constraint.
3. In a video sequence, the intensities of corresponding points in the temporal direction can be assumed equal, which is known as the Lambertian assumption. However, this is not the case for the corresponding points in the view direction, since the cameras and the lighting conditions are different for different views.
4. The motion vectors are two-dimensional while the disparity vectors, under the epipolar constraint, can be one-dimensional. Thus, coding disparity vectors will need fewer bits than coding motion vectors.
5. In a stereo pair, every point in the left image is displaced in the right image, while only moving objects are displaced from frame to frame in the temporal direction. If we disregard the occlusions, a global disparity model can well represent the disparity field between two views for one object that has homogenous motion over it, e.g., the background.

2.2.2 Disparity Estimation Algorithms

Block-based Matching

The most commonly used technique for disparity estimation is the block-based matching (BBM), e.g., the fixed-size block matching algorithm (FSBM). In FSBM, the image is divided into square blocks and matching is performed based on the correlation function of two blocks in the two images. Each block will obtain one displacement vector. FSBM does not require any coding overhead, but only need to transmit the block disparities and disparity compensated residual. They are simple and effective to implement.

There are at least two problems associated with block-based methods that researchers have been trying to solve, the first of which is to define the cost function. Mean absolute difference (MAD) and mean square error (MSE) of pixel intensity values are most frequently used in BBM. Brunig [17] uses combined MAD and MSE measures. In [18], pixel differences classification (PDC) is selected since it reflects the minimum risk criterion and performs better than MAD in subjective quality. The second problem is to define the size of the matching block. On the one hand, small featureless blocks lead to spurious matches and the coding overhead will be large. On the other hand, larger block size will result in larger compensation error and requires higher bitrates to transmit the residual image. The block size should be large enough to include enough intensity variation for characterizing an area, but small enough to avoid projective distortion. The hierarchical block matching (HBM) method [11, 18, 19, 20, 21] is commonly adopted as a tradeoff. For example, an overlapped HBM disparity estimator can reduce blocking artifacts by using multiple vectors for a block [11, 18, 22]. In [23], an appropriate window is selected by evaluating the local variation of the intensity and the disparity using a statistical model of the disparity distribution. Woo [24] proposed to use a locally adaptive window with variable shapes. The algorithm searches for a window that produces the estimate

of disparity with the least uncertainty for each pixel of an image.

Still, there are several drawbacks for BBM methods due to their intrinsic properties [25]. Firstly, blindly dividing the image into square blocks of usually 8×8 or 16×16 pixels without taking into account its semantic content results in serious blocking artifacts, as the square blocks of pixels may have little resemblance to the true disparity of the objects that form the image. Blocks that fall on objects at different depths may suffer from blockiness at low bit rates. Although overlapped block matching can partially solve this problem, it leads to large coding overhead and compensation error [26]. Secondly, inaccurate disparity estimation and compensation can result because only a single disparity is transmitted for each block. The accuracy of the disparity vector may not be sufficient to describe the actual disparity. Intensity noise and occlusions lead to disparity vectors that minimize the disparity compensated difference (DCD) energy but may not correspond to the true disparity. Finally, the properties of the Human Visual System (HVS) are not taken into considerations. Since the HVS is sensitive to object boundaries, which are usually related to abrupt intensity changes, these blocking errors can be visually annoying. BBM methods may also fail to provide accurate matching results because the correspondences are locally ambiguous due to noise, occlusion, lack of texture and repetitive texture [27]. They cannot take the advantage of large areas of constant disparity, which usually occur in the background.

Feature-based Matching

Feature-based techniques [28, 29, 30, 31, 32] can avoid such distortion and matching errors. A variety of features, including zero-crossings, edges, curves, corner points and distinctive patterns can be detected, extracted and matched between the image pair. The other pixels are interpolated from neighboring features.

Feature-based methods can be classified into two categories by the type of features

considered: edge-based and region- (or area-) based methods. In edge-based methods, a matching path usually consists of the left-occlusion-path, matching-path and right-occlusion-path. The matching is aimed at minimizing the matching-path cost, which is determined by the feature vectors of two corresponding spans. For example, in [33], a feature vector, including average Y, U, V, the energy of AC component for Y and so on, is set up for each span and used for matching. Dynamic Programming (DP) [34, 35] has been applied in edge feature based matching. Besides the DP methods, Ji [31] proposed to use wavelet to perform multiresolution edge extraction and matching. In region-based matching [27, 36], the image is segmented into homogeneous regions, and the contours (or shape) of segmented regions and textures are used as features. Since regions have higher discrimination capability than edges and there are fewer features to be matched, region-based approaches are more accurate, reliable and efficient than edge-based matching. The reduction of the number of features to be matched, along with the region's high discrimination power, greatly decreases the number of mismatches and false targets.

Other features may also be used for the correspondence problem. For example, there is a one-to-one relationship between the sets of the boundary points of the connected components of the stereo images under certain conditions. So the boundary of a parameter-dependent connected component of an image can be used as the matching feature [30]. A similar method is adopted in [28] where feature points are detected and connected in between to form straight lines. After that, straight-line correspondence is established and triplets of matched points are used to construct a polygon model in the left image. Then the exhaustive search method is applied to the right image to find a matching polygon. Cun [32] combines the classic feature matching technique and the modified genetic algorithm, and proposed a chromosome construction algorithm and a fitness function for stereo matching. Edges extracted by Canny operator are considered as the matching feature.

The disadvantage of feature-based matching is that feature matching itself yields only coarse disparity maps and the disparity vectors for non-feature points have to be interpolated. Linear and bi-linear interpolation are easy to implement while non-linear interpolation [37, 38] can usually obtain better approximation of disparity values from its neighborhood. Kardouchi [29] combines the feature matching with Delaunay triangulation. After the feature points are found and matched, a dense vector field is computed by planar interpolation over Delaunay triangles. The advantage is that the blocking artifacts can be eliminated using feature-based matching. In addition, feature-based matching can embed the occlusion models directly into the disparity estimation process.

Mixed Disparity Estimation

From the above analysis of the block- and feature-based disparity estimation techniques, it is natural to combine the superiorities of the two and overcome their deficiencies. Such hybrid matching methods [39, 40] can inherit the accuracy of the feature-based approaches, simplify the procedure of feature matching and produce robust and dense disparity fields.

In [39], a hierarchical genetic algorithm is presented to estimate reliable disparity fields with low computational cost. The image is divided into sets of feature points and non-feature points. A cost function combining area and feature matching is proposed, which applies the morphological gradient for feature points and the disparity Laplacian function for non-feature points. In [40], a hierarchical hybrid matching algorithm (HHM) is proposed which include four steps: (1) segmentation and feature extraction, (2) mapping line segments into blocks, (3) feature matching and global block matching, and (4) local block matching. The mixed disparity estimation algorithms provide a tradeoff regarding to the accuracy and the complexity of the disparity estimation.

Object-based Matching

In recent years, the concept of object is becoming more and more important and object-based methods are applied in coding applications. In object-based or generally content-based methods, the properties of the HVS is considered such that the important objects are given more considerations than the background. Several semantically meaningful objects can be extracted from the image on which disparity estimation is performed. In video coding, by transmitting the shape, motion and color of the objects in an image, it is possible to avoid the annoying coding errors, such as mosquito effects and blocking artifacts, produced by block-based hybrid coding. Furthermore, important image areas such as the facial details in face-to-face communications can be reconstructed with higher image quality than the background. Ohm [13] presented an object-based system for stereoscopic viewpoint synthesis, which identifies foreground and background regions first and then applies the HBM algorithm with global and local disparity estimation to the foreground. Many current content-based matching algorithms adopt the concept of Video Object Plane (VOP) in MPEG-4, where the disparity estimation is performed on a VOP basis. Different matching algorithms can be applied to the foreground and the background VOPs to achieve better results on the more important foreground object.

Model-based schemes that define the 3D space structural models of the scene can also be categorized as object-based methods. More frequently, a 3D wireframe model is mapped onto the 2D space to form a triangular or quadrilateral mesh. Firstly, two models are built for the objects in the left and the right images respectively, using static or adaptive triangular/quadrilateral meshes. Then, the disparity and motion vectors of the vertices of the mesh are estimated and the affine parameters are computed. However, since natural scenes can hardly be modeled, almost all the efforts in model-based video coding so far have been concentrated in head-and-shoulders images.

Stochastic and Deterministic Methods

From the mathematical point of view, disparity estimation is an ill-posed problem and various constraints are needed to obtain a unique and stable solution. The stochastic approaches based on the MRF model of the image are well suited for such problems. The MAP based estimations have been widely discussed for various applications such as field estimation, segmentation, object tracking, restoration, etc. Lee et al. [41, 42] proposed the dense correspondence field estimation with the stochastic diffusion based on MAP. The stochastic methods are normally performed on a pixel-basis by iterative optimization and the main disadvantage is the extremely high computational complexity. Such time consuming approaches can obtain precise results for applications that require high fidelity of displacement fields, such as IVR, mixed reality, 3D scene structure from stereo images, etc.

However, for stereoscopic and multiview video coding for 3D telepresence, we should avoid the high complexity for disparity estimation. Deterministic approaches are proposed as approximations of the stochastic process so that the computational complexity can be lowered, e.g., the Mean Field Annealing (MFA) methods [43, 44] and the regularization method [45]. In the regularization algorithms, the displacement field is iteratively updated to minimize a cost function, which usually contains a similarity term, a smoothness term and other various constraints [39]. The extent of smoothness is controlled by the Lagrange multiplier, which can be optimally set under the rate-distortion theory [19, 46].

2.3 Joint Estimation

In multiview video coding, some existing disparity and motion estimation methods do not utilize both redundancies or use them separately. For example, the main view is inde-

pendently coded using motion compensation and the auxiliary view is coded by disparity compensation [27, 47]. Only disparity fields are estimated for the auxiliary view and the temporal redundancy is not removed. In [48], the authors suggest the use of disparity information only while do not consider the motion information even within the main view. A hybrid HVS coder with 3D-DCT (Discrete Cosine Transform) is used instead of motion compensated transform coding for the main view, and the auxiliary view is encoded using disparity prediction.

However, researches show that jointly utilize the view and temporal redundancies is one of the key factors to achieve high compression ratio in stereoscopic and multiview video coding [49, 50, 51]. In approaches where the disparity fields are independently estimated at subsequent time instances, the decoded images may suffer from annoying artifacts due to temporal inconsistencies. Temporal consistency restrictions have been reported to produce more natural results and the relations between the motion and the disparity fields can be utilized either in the estimation phase or in subsequent steps. It is natural and usually advantageous to integrate motion compensation with the disparity estimation and compensation, and utilize disparity and motion information in a coherent way [49, 52, 53, 54, 55].

The theoretical foundation of joint disparity and motion estimation is the stereo-motion consistency constraint, which is also known as the “loop” constraint. As shown in Figure 2.4 [56], the simplest relation between disparity and motion fields for two image pairs at two successive time samples can be expressed as:

$$\vec{M}_l + \vec{D}_t - \vec{M}_r - \vec{D}_{t+1} \approx \vec{0} \quad (2.3.1)$$

where $\vec{M}_l$ and $\vec{M}_r$ are the motion fields of the left and the right image pairs, $\vec{D}_t$ and $\vec{D}_{t+1}$ are the disparity fields of the stereo image pair at t and $t + 1$. Simultaneous estimation of all the fields allows for complex field models that contain any kind of interaction, which enhances the accuracy [54]. However, it is a challenge as the dimensionality of the

Figure 2.4: Coherent disparity and motion analysis.

solution space is extremely large.

Various algorithms have been proposed based on the loop constraint. A typical way to include temporal information during disparity estimation is to add the temporal smoothness constraint in the cost function of the disparity matching process [49, 57, 58]. In [49], the disparity field of the first stereoscopic pair is estimated separately. With this known disparity field, the two dense motion fields for both left and right sequence, and the disparity field of the next stereoscopic pair are jointly estimated at each time instance by a multiscale iterative relaxation algorithm. A temporal consistency restriction is imposed over the disparity fields and thus over the interpolated stereo images. Tsai [58] derived a constraint from the inherent stereo-motion redundancy of the 3D structure. A coherence condition between motion and disparity in stereo sequences may be expressed as a linear combination of four vectors (two motion and two disparity vectors) in two successive frame pairs. This constraint can be set as a reliability test. In [59], the algorithm first

calculates a temporal independent depth map from the current image pair by means of correspondence analysis and triangulation. Then the 3D motion of the model objects is estimated and compensated. A second depth map, i.e., the motion compensated depth map is computed from the motion compensated model objects. The final depth map, which is used to update the 3D shape parameters of the model objects, is calculated by averaging both depth maps depending on the reliability of the respective depth estimated with a Kalman filter approach. In the compensation phase, motion compensation is used for disparity occluded areas and disparity compensation is used for motion occluded areas. If both are available, the average or other values (e.g., using filters and other image processing tools) derived from both the disparity and the motion fields is used. In case of both disparity and motion occluded regions, the disparity values for these points can be estimated by the multiscale gradient-based algorithm.

2.4 Digital Video Coding

In this section the fundamentals of digital video coding are reviewed and the existing stereoscopic and multiview encoders are summarized. *Video* refers to time-varying images and has traditionally been recorded, stored, and transmitted in the analog form. The advent of digital video opens up a number of opportunities for interactive video communications and services [60]. Digital video is preferred over analog video because of the following reasons [9]:

1. digital transmission can be made more robust than analog transmission,
2. digital copying does not degrade the quality of the copy,
3. further processing on digital video can be carried out at a pixel-by-pixel level,
4. finer tradeoff between perceived quality and transmission bandwidth becomes pos-

sible.

The major bottleneck for digital video applications has been the huge storage and transmission bandwidth requirement.

2.4.1 Fundamentals

A video sequence is composed of pictures updated at a certain bit rate, normally with a number of pictures grouped together (group of pictures or GOP). Each picture is composed of several groups of blocks (GOB), or slices. Each GOB contains a number of macroblocks (MB) of 16×16 pixels, and each MB is composed of four 8×8 luminance blocks, which represent the intensity information, and corresponding chrominance blocks which represent the color information.

The coding scheme widely used in most video coding standards is a combination of *intraframe* and *interframe* coding. *Intraframe* coding exploits the spatial redundancy within one frame and by DCT or DWT (discrete wavelet transform). DCT is applied to each block to transform the pixel values into a set of DCT coefficients. The coefficients are then quantized and zigzag scanned to provide a sequence of symbols, with each symbol representing a number of zero coefficients followed by one nonzero coefficient. These symbols are then converted into bits by entropy coding, e.g., variable-length coding (VLC). *Interframe* coding removes the temporal redundancy between adjacent frames by motion compensation (MC). Since typical video data contains a large amount of redundancy along the temporal axis, the video contents of temporally adjacent frames are very similar. In motion-compensated prediction, several most recently decoded frames are used as predictions for the current frame. A motion vector for each block or macroblock is searched and encoded, together with the residual data.

There are normally three different frame types. Intra coded frames are called intra

Figure 2.5: Frame structure in the MPEG standards

pictures (*I* frames). Pictures that are coded using a previous picture as a reference or prediction are called inter or predicted pictures (*P* frames). *B* frames are bi-directionally coded using two reference pictures. One typical frame prediction structure in the MPEG standards is described in Figure 2.5. However, a *P* frame may also contain some intra-coded macroblocks. For a certain macroblock, it may be impossible to find a good enough matching area in the reference picture to be used for prediction, especially in cases of occlusion or rapid motion. In this case, direct intra coding might be more efficient.

The coding the *P* and *B* frames includes the coding of DCT coefficients of the residual image and the coding of motion vectors. Since motion vectors of neighboring blocks tend to be similar, the horizontal and vertical components of motion vectors are differentially coded using a predefined VLC table.

2.4.2 Developments and Standards

Video coding for telecommunication applications has evolved through the development of the ITU-T H.261, H.262 (MPEG-2), and H.263 video coding standards (and later enhancements of H.263 known as H.263+ and H.263++) [61]. Throughout this evolution, continued efforts have been made to maximize the coding efficiency while dealing with the diversification of network types and their characteristic formatting and loss/error

Table 2.1: Video coding standards developed by various organizations.

Organization	Standard	Typical Bit Rate	Typical Applications
ITU-T	H.261	$p \times 64$ kbps, $p = 1 \dots 30$	ISDN Video Phone
ISO	IS 11172-2 MPEG-1 Video	1.2 Mbps	CD-ROM
ISO	IS 13818-2 MPEG-2 Video ^a	4-80 Mbps	SDTV, HDTV
ITU-T	H.263	64 kbps or below	PSTN Video Phone
ISO	IS 14496-2 MPEG-4 Video	24-1024 kbps	A variety of applications
ISO/ITU-T	H.264/AVC	> 64 kbps	A variety of applications
^a ITU-T also actively participated in the development of MPEG-2 Video. In fact, ITU-T H.262 refers to the same standard and uses the same text as IS 13818-2.			

robustness requirements. Table 2.1 summarizes these standards. The early video coding standards, including H.261, MPEG-1 and MPEG-2, use the same hybrid DCT-MC framework as described. The new generation of video coding standards contains many optional modes and supports a larger variety of functionalities.

H.263 combines the features of H.261 with several new methods, including the half-pixel motion compensation and other techniques. In terms of PSNR (peak-signal-to-noise-ratio), H.263 can provide about a 3 dB gain over H.261 at very low bit rates and actually superior coding efficiency at all bit rates. H.263 was chosen as the basis of the MPEG-4 video design. The major difference between MPEG-4 and the previous standards lies in the representation and compression of the shape information. In MPEG-4 the concept of content-based coding is used, where each input picture is decomposed into a number of arbitrarily shaped regions, i.e., VOPs. The shape of each VOP is encoded using context-based arithmetic coding. MPEG-4 thus provide content-based interactivity, universal access and high compression coding of general material for a wide bit rate range.

H.264/AVC is the newest video coding standard of the ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group. The main goals of the H.264/AVC standardization effort have been enhanced compression performance and provision of a “network-friendly” video representation addressing “conversational”

(video telephony) and “nonconversational” (storage, broadcast, or streaming) applications. H.264/AVC has achieved a significant improvement in rate-distortion efficiency relative to existing standards.

2.4.3 Scalable Video Coding

Scalability of video is the property that allows a video decoder to decode portions of the coded bitstreams to generate decoded video of quality commensurate with the amount of data decoded. In the application scenario of streaming video over the Internet, video server has to deal with the problems of various device capabilities, diverse network environments and bandwidth fluctuations. Scalable video coding (SVC) provides an elegant solution to this problem, which compresses the video into a scalable bitstream so that video data can be decoded at different resolutions from extracted sub-bitstreams.

It is necessary to offer scalability with at most a slight loss in source coding efficiency versus the state-of-the-art non-scalable coding approaches, as well as not requiring extra complexity versus the baseline. As indicated in [62], it has been a dream within the video coding research community that streaming delivery of video on networks will benefit from scalable video coding functionalities. Recent research on subband/wavelet inter-frame coding has shown that combining temporal-axis filter banks, motion compensation without the prediction loop, and embedded bit plane coding of the resulting subbands, the dream can come true. This framework is called the motion compensated temporal filtering (MCTF). Research has shown that MCTF coding schemes would still improve when run in a non-scalable mode for one specific bit rate, and the efficiency penalty for providing scalable streams is significantly less than in traditional (closed loop) MC prediction. Even though MCTF goes back to research done in the early 1990s, a breakthrough was made three years ago when several research teams demonstrated MCTF with the lifting scheme that can combine the necessary subpixel interpolation with DWT. Cur-

Figure 2.6: Wavelet decomposition with motion trajectory for an image sequence.

rently the Joint Video Team (JVT) is working towards a standard for SVC based on these concepts as an extension of the H.264/AVC video coding standard.

Besides, recently a large amount of efforts have been invested in the application of 3D wavelet in scalable video coding [63, 64, 65]. In the 3D wavelet SVC schemes, the whole video is decomposed into various spatial-temporal subbands through several MCTF and subsequent spatial wavelet transforms. Figure 2.6 shows a typical temporal wavelet decomposition along with motion trajectory, where each column represents a frame. Wavelet-based video coding has been proven to be an excellent way to achieve both good coding performance and full scalabilities. The performance of 3D wavelet coding schemes can be comparable to the state-of-the-art non-scalable video coding standard MPEG-4 AVC/H.264 part10 [66]. Meanwhile, the temporal, spatial and SNR scalabilities can be supported easily with the subband decomposition and bitplane coding.

In the SVC framework proposed in [67], the raw video data is first input to the motion estimation module, which performs motion estimation on variable block sizes from 16×16 to 4×4 with quarter pixel precision. Then, the Barbell lifting is employed for the temporal wavelet decomposition. The Barbell lifting contains three steps: *split*, *prediction* and *updating*. Video data is split into two subsets, X and Y, in the first step.

Then, in the *prediction* step, the subset X is used to predict the subset Y and the high-pass wavelet subband H is calculated as the prediction error. The overlapped motion alignment is enabled in the prediction step. The third step, *updating*, uses the high-pass wavelet subband H to update subset X to ensure the preservation of moments in the low-pass subband L . After the temporal wavelet decomposition, the resulting high-pass and low-pass frames are further decomposed by 2D spatial wavelet transform with the 9/7 filters. Finally, all subbands are encoded with the 3D embedded subband coding with optimal truncation (3D-ESCOT) algorithm [68].

2.4.4 Performance Measures

The performance of a lossless compression method can be evaluated based on the compression ratio alone. Usually we use average bits-per-pixel (bpp) defined below as the measure for compression efficiency:

$$\text{Average bpp} = B/M \quad (2.4.1)$$

Here, B =Total number of bits needed to encode a stream; M =Total number of pixels in that stream.

The performance of lossy coding methods can be evaluated in terms of the compression ratio (or bitrate), objective quality and perceptual quality. The quality refers to the accuracy of the decoded video frames and can be measured objectively by PSNR, which is calculated as

$$PSNR(dB) = 10 \log_{10} \frac{255^2}{\frac{1}{M \times N} \sum_{m=1}^M \sum_{n=1}^N (F(m, n) - \tilde{F}(m, n))^2} \quad (2.4.2)$$

Here, F and $\tilde{F}$ denote the original video frame and the decoded video frame, respectively.

Besides, the performance evaluation should also consider complexity, flexibility and generality of the system [33]. The computation complexity can be relatively measured

relatively by the total processing time, giving the configurations of the environment. The flexibility is characterized by a ready capability to adapt to new, different, or changing requirements. For example, the system parameters control module should be able to be adjusted or replaced according to the specific applications easily. Generality is defined as not confined by specialization or limitation. All these factors should be considered in building up the telepresence system.

2.5 Stereoscopic and Multiview Video Coding

2.5.1 Existing Techniques

The left and right views that constitute a stereoscopic image pair are obviously very similar. They are often described as “highly redundant” in that most of the information contained in either is repeated in the other, or as “highly correlated” in that either is for the most part easily predicted from the other by application of some external information about the relationship (the relative perspective) between them [69].

HVS theory indicates that one good eye and one bad eye together are better than the good eye alone, i.e., the information they provide in a sense adds rather than averages. In addition, a binocular 3D stereoscopic image pair with one sharp member and one blurred member can successfully stimulate the depth perception. From the research in HVS, it is clear that a “reasonable approximation” is enough. These observations provide the possibility that we may find representations which takes up only a little more storage space and transmission bandwidth than required by either left or right view alone. By this way, the main stream is encoded losslessly or slightly lossy while the auxiliary stream is encoded very lossy to lower the bitrate.

Existing practical stereoscopic and multiview video coding systems typically consist of four modules [50]: (1) coding one of the views (the main view) using a conventional

method (e.g., MPEG), (2) calculating the disparity map between corresponding points in the main view and the auxiliary view, (3) coding the disparity map, and (4) coding the disparity compensated residual. The bitstreams to be transmitted thus consist of the main stream, the disparity map and the residual in the auxiliary stream, and other information for error resilience etc.

Standard-Compatible MVC

Multiview video coding has been included in several video coding standards [70], e.g., MPEG-2 MVP (multi-view profile) and MPEG-4 MAC (multiple auxiliary component). Compatibility with image and video coding standards is a desirable feature [71], which implies that a decoder without the multiview profile should be able to decode the main view. Based on the framework provided by the video standards, many standard-compatible MVC schemes have been proposed [57, 72, 73, 74, 75, 76, 77, 78].

In [57], the left image stream is encoded independently by an MPEG-type coding scheme. For the right image stream, the reference frames (I and P) are encoded by the subspace projection technique and the rest of the frames are not encoded. An interpolation algorithm is developed to exploit the great redundancy within stereo sequences to reconstruct those un-coded frames at the decoder to achieve a high compression ratio. In this scheme, the motion and disparity fields are estimated by HBM with a multiresolution structure, and coded by an entropy coding technique. Ohm [72] analyzed the compression of stereoscopic and multiview video data using the MPEG family of standards. In the MPEG-2 MVP, it is straightforward to perform stereo and multiview channel encoding using the temporal scalability syntax. Frames from the main stream are defined as the base layer, and those from the auxiliary stream as enhancement layer(s). The enhancement-from-base-layer prediction thus turns out as a disparity compensated prediction instead of a motion compensated prediction. Motion compensated prediction

within the same channel is performed in case that disparity compensated prediction fails. For the integration with MPEG-4, the disparity map could be encoded by the use of the tools provided for grayscale alpha data encoding. The schemes presented in [75, 76] are based on the MPEG-2 standard and the work in [77] is based on the MPEG-4 standard. More recently, an H.264-based multiview video coding scheme was proposed in [78], which utilizes the multiple reference structure in H.264.

Non-Standard-Compatible MVC

In addition, various non-standard-compatible stereoscopic and multiview video coding techniques [79, 80, 81, 82] have also been proposed. In both [79] and [80], object-based stereoscopic video coding algorithms which rely on modeling of the object structure using 3D wire-frame models are proposed. These object-based techniques perform well for head-and-shoulder sequences in teleconferencing applications but are unsuitable for other video contents since segmentation and building up 3D models for the natural scenes are extremely difficult. In [81], the authors studied the structure and motion estimation in multiview teleconferencing-type sequences. In [82], a sprite generation algorithm was proposed in which the temporal and view-directional correlations are simultaneously utilized by dynamically generating a sprite image as a reference in the prediction.

Scalable MVC

Scalability is also considered in the encoding of stereoscopic [83] and multiview [84] image data to facilitate its transmission over heterogeneous networks and to cater end users with various devices. In [83], the stereoscopic image pairs are encoded using wavelet transform to provide an embedded bitstream. The open-loop structure, where the right image is predicted by the original left image, and the close-loop structure, where the reconstructed left image is used for prediction are compared and analyzed. With half-pixel accuracy

disparity estimation, the close-loop prediction, and the zerotree coding, the proposed algorithms, while providing fully embedded bitstream, also achieve better rate-distortion (R-D) performance than conventional coding schemes.

Disparity-compensated lifting for wavelet compression of multiview images is studied in [84]. The light fields can be represented as a set of views in 2D images, which are re-arranged as a multiview image set. Object-based coding using shape adaptation is also proposed. The experimental results prove that the disparity-compensated lifting with wavelet coding provides not only better encoding efficiency but also various kinds of scalabilities for the multiview image set.

2.5.2 Stereoscopic and Multiview Rate Control

To fit the encoded video stream in a specified bandwidth and at the same time obtain smooth and optimal visual quality, the rate control module allocates bits among different frames in the temporal axis in video coding. For a multiview video sequences encoder, rate allocation is especially important for bitrate and quality control since it allocates bits among different pictures in the view axis as well as in the temporal axis.

The rate control problem in video coding has been extensively studied in recent years. A quadratic rate distortion model for MPEG-4 rate control is proposed in [85]. After that, extension to multiple video objects rate control [86] and various improvements [87] are proposed. Rate control for stereoscopic and multiview video sequences coding has also been discussed in the literature. For multiview video rate control, an extension from MPEG-4 is quite straightforward by setting a larger number of picture types and defining the constant proportions among different picture types [88].

However, rate control of multiview video coding should be considered as a dependent bit allocation problem since the decoded image quality of the main view frames exerts

a large influence on the gain in the auxiliary view frames as references for disparity compensation. When the number of encoded bits is reduced to some extent for the predicted image, and these saved bits are assigned to the reference image, the reference image quality can be improved without any degradation of the predictor, as long as the improvement of the reference image quality introduces enough bit-reduction effect for the predicted image coding [89, 90]. Thus higher coding efficiency can be obtained by exploring the dependency of the disparity compensated image pairs.

In [89], the authors proposed rate control technologies for stereoscopic video coding based on MPEG-2 MVP, but it does not contain rigorous theoretical justifications. In [91], a blockwise dependent quantization for stereoscopic image coding was proposed. The authors also pointed out that in stereo image coding, an optimal independent bit allocation of the reference image may not be the optimal, as it affects the performance of the target image. However, dynamic programming (e.g., the Viterbi algorithm) which is typically used to solve the dependent quantization problem [90, 91], is very time consuming and not suitable for telepresence applications. In [92] the authors observed from empirical data that the variance of the motion-compensated residue grows linearly with the coding error of the reference frame. Later, in [93], frame dependency between the reference picture and the predicted picture is further studied for an embedded wavelet video coder for better rate-distortion tradeoff. In [94], the authors proposed to add an EOD (estimation of dependency) function to the Lagrangian cost function to characterize the frame dependency. In [95], the dependency of the reference frame and the predicted frame is estimated. The distortion for the target frame is simply the sum of the original prediction difference and the quantization errors of the reference frame. In the rate control of stereoscopic video coding, human vision factors are also considered [96], where four indicators are defined and calculated to refine the quantization parameter.

Chapter 3

Joint Estimation and Regularization of Disparity and Motion

In this chapter, we first give a short description of our multiview video capturing system and test data in Section 3.1. The preprocessing steps for disparity estimation, including the image balancing and the global disparity estimation, are elaborated in Section 3.2. After that, we focus on the disparity and motion estimation since it determines the decorrelation efficiency and constitutes one of the crucial components for stereoscopic and multiview video coding. We present an edge-preserving regularization scheme to calculate the disparity fields for a stereoscopic image pair in Section 3.3, where an energy function is proposed based on the MRF theory. A fast algorithm is also proposed where the textured regions and the homogeneous regions of the images are identified and regularized differently. The regularization scheme is then developed into a joint disparity and motion estimation scheme for stereoscopic video sequences in Section 3.4. An overall energy function based on the MAP probability and the loop constraint is minimized iteratively to calculate two motion and two disparity fields simultaneously in a mutually beneficial way. Evaluation results are discussed in Section 3.5.

3.1 MultiCam System and Video Capturing

3.1.1 MultiCam System

The MultiCam system¹ as shown in Figure 3.1 is setup to capture multiple views video data simultaneously. The system consists of three component cameras controlled by three motion control motors respectively. The motion control motors are software controlled. Three CPUs are used to control the system and a trigger is used to ensure data synchronization. Cam1 and Cam2 can move horizontally at a range of $(62\text{ mm}, 500\text{ mm})$ and $(-62\text{ mm}, -500\text{ mm})$. Cam3, which is the middle camera in Figure 3.1(b), can move in all x, y, z directions. All the three cameras can rotate from -180 degrees to 180 degrees and the focus range is from 5 mm to 50 mm . The 3D positions of the three cameras are also recorded before video capturing. The video data is recorded in either YUV or RGB format.

A simple camera geometry is used to capture the 3-view video sequence Reading, where the three cameras line on a horizontal line with equal baseline distance 62 mm between two neighboring cameras. The focus length for all the three cameras are 16 mm and the approximate object distance is 3 m . The MultiCam system is hardware-calibrated and the disparity values on vertical direction can be limited to two or three pixels.

3.1.2 Test Video Sequences

Our test video sequences used in Chapters 3 and 4 include not only standard stereoscopic test image pairs and video sequences, but also the multiview video sequences captured by the MultiCam System. The parameters of the test data are presented in Table 3.1 and

¹The MultiCam system is setup under the University Academic Research Fund, (URC) RG3/02-Depth-Based Video Segmentation, of Nanyang Technological University, Singapore.

(a) MultiCam System

(b) Three Cameras

Figure 3.1: MultiCam system for multiview video capturing.

Table 3.1: Description of test multiview video sequences.

Test Data	Resolution	Frame Rate	No.Views	No.Frames	Shape
PIANO	720×576	N/A	2	1	No
MAN	512×512	25 fps	2	25	Yes
TRAIN & TUNNEL	720×576	25 fps	2	250	No
Reading	768×288	25 fps	3	250	No

the example images for the original test image pairs/video sequences (the first frame) are shown in Figure 3.2.

In the sequence Reading, the background is static and homogeneous. There exist normal speed motion for the shoulder, the hands (with the pen) and the book, together with slow speed motion for the head. For simplicity, the interlaced video data is separated into even and odd fields, and only the even field is used in our work.

3.2 Balancing and Global Disparity Estimation

3.2.1 Image Balancing

The corresponding pixels between the stereoscopic image pair may not have the same intensity values because of the lighting conditions and inherent camera differences. The purpose of image balancing is to eliminate such potential signal difference between the image pairs. The intensity values of the right image are linearly transformed so that the mean and the variance are equal to those of the pixel intensities in the left image. The linear transform is performed as follows [97]:

$$\hat{I}_R(i, j) = \frac{\sigma_L}{\sigma_R} I_R(i, j) + (m_L - \frac{\sigma_L}{\sigma_R} m_R) \quad (3.2.1)$$

where σ_L and σ_R are the variance, m_L and m_R are the mean of pixel intensities of the left and the right images, respectively. $I_R(i, j)$ and $\hat{I}_R(i, j)$ represent the intensity values of pixel (i, j) of the right image before and after balancing, respectively. The resulting

(a) PIANO

(b) MAN

(c) TRAIN & TUNNEL

(d) Reading

Figure 3.2: Original left view of test stereoscopic/multiview image pairs or video sequences (the first frame), (a) PIANO, (b) MAN, (c) TRAIN & TUNNEL and (d) Reading.

intensity values are limited to the range of $[0, 255]$. The image balancing is performed as a preprocessing step and the auxiliary video sequence is input to the encoder after balancing w.r.t. the main video sequence.

3.2.2 Global Disparity Estimation

Since the two views are captured at the same time, the global disparity can well describe the movement brought by the difference in viewing angles, especially for the stationary background scene.

For some stereoscopic and multiview video data, e.g., Reading, when the baseline is large comparative to the object depth, the disparity value between the two views can be very large and leads to inefficient disparity compensation. Thus we compensate the auxiliary view using the global disparity models first as preprocessing.

Generally, the global disparity (or motion) model can be represented as

$$\begin{aligned} x_1 &= \frac{d_1x + d_2y + d_3}{d_7x + d_8y + 1} \\ y_1 &= \frac{d_4x + d_5y + d_6}{d_7x + d_8y + 1} \end{aligned} \quad (3.2.2)$$

In MPEG-4 sprite coding, the following four global motion models are introduced:

1. translational transformation for mostly translational motion. Only one point is needed and thus one global motion vector (GMV) is encoded. $d_1 = d_5 = 1$, $d_2 = d_4 = 0$, $d_7 = d_8 = 0$;
2. isotropic transformation to represent isotropic magnification, rotation and translation motion, where four parameters are needed and thus two GMVs are encoded. $d_1 = d_5$, $d_2 = -d_4$, $d_7 = d_8 = 0$;
3. affine transformation, where three GMVs are encoded to represent the six parameters. $d_7 = d_8 = 0$;

4. perspective transformation, four GMVs are encoded for the eight parameters.

For the Reading sequences, the translational model is selected to represent the global disparity between adjacent views. One global disparity vector (GDV) is used to represent the two model parameters and encoded losslessly for each view. After calculating the global disparity model parameters, the reference frame is warped for the prediction of the current frame. The boundary pixels that cannot find corresponding pixels after warping are padded by neighboring pixels.

3.3 Edge-Preserving Disparity Regularization

3.3.1 The MRF Model and Problem Formulation

The theoretical foundation of the proposed disparity regularization algorithm is the MRF theory. It provides a convenient and consistent way of modeling context dependent entities such as image pixels and other spatially correlated features. Based on the MRF theory, the disparity estimation can be regarded as a labeling problem, where one label (i.e., a disparity vector) is assigned to each of the sites in image S (i.e., the image grid). The image S is considered as a rectangular lattice of size N and the disparity field $D = \{D_{i,j}, (i,j) \in S\}$ denotes a family of random variables defined on S .

An MRF is characterized by its local property (the Markovianity), i.e., the sites in S are related to one another via a neighborhood system. According to the Hammersley-Clifford theory [98], the probabilistic consideration of the MRF model is converted to Gibbs energy distribution in a form like:

$$U(d) = \sum_{c \in C} V_c(d) = \sum_{(i,j) \in C_1} V_1(d_{i,j}) + \sum_{(i,j) \in C_2} V_2(d_{i,j}) \quad (3.3.1)$$

where $V_c(d)$ is the clique potential of all the possible first order clique sets which are the single-site C_1 and the double-site C_2 . In the disparity estimation problem, the objective

is to maximize the joint probability which is equivalent to the minimization of the energy function. How to choose the forms and the parameters of the potential function to properly cooperate the constraints is a major topic in MRF modeling.

3.3.2 The Energy Function

Based on the MRF theory, we define the energy function as

$$F(D) = F_1(D) + \lambda F_2(D) \quad (3.3.2)$$

Here, F is the cost function on the displacement field D . λ is the Lagrange multiplier that controls the extent of smoothness and can be determined by the rate-distortion theory. F_1 and F_2 are the similarity term and the smoothness term for the displacement field, respectively. Let f_1 and f_2 represent the similarity and the smoothness term of one block, $d(i, j)$ represents the displacement vector for one block. To calculate the disparity value $d(i, j)$ of a block, the similarity term (MAD) is:

$$f_1(\hat{d}(i, j)) = \frac{1}{b^2} \sum_{x=0}^{b-1} \sum_{y=0}^{b-1} |I_L(i+x, j+y) - I_R(i+x, j+y+\hat{d}(i, j))| \quad (3.3.3)$$

where $\hat{d}(i, j)$ is the test value of disparity $d(i, j)$ on site (i, j) , I_L and I_R are the pixel intensities of the left and right images, and b is the block size.

The smoothness constraint has been one of the most popular prior assumptions in low level vision. However, in disparity estimation, objects at different depth have different disparity vectors, which will be ignored and smoothed out by generic smoothness measures. Thus, our proposed smoothness measure considers the object edges. The edge-preserving smoothness term is defined as

$$f_2(\hat{d}(i, j)) = \frac{|\nabla^2 \hat{d}(i, j)|}{|w_{i,j} - \frac{1}{n} \sum_{(x,y) \in N_{i,j}} w_{x,y}| + \kappa} \quad (3.3.4)$$

where

$$\nabla^2 \hat{d}(i, j) = d(i-1, j) + d(i+1, j) + d(i, j-1) + d(i, j+1) - 4 \cdot \hat{d}(i, j) \quad (3.3.5)$$

is the second derivative of the vector field. $w_{i,j}$ and $w_{x,y}$ are the edge-preserving weights of the current block and the neighborhood blocks. $N_{i,j}$ is the neighborhood of block (i, j) and n is the number of neighboring blocks. The disparity field on site (i, j) only depends on the disparity field of the four neighborhood sites $N_{i,j}$. The first order neighborhood is considered in the block-based disparity estimation problem. In the first order neighborhood system, every (interior) site (i, j) has four neighbors $(i-1, j)$, $(i, j-1)$, $(i+1, j)$, $(i, j+1)$. The second derivative of the vector field is used for the smoothness term because it represents the roughness of data well. The regularization scheme can also be applied in the pixel-based format, where the second order neighborhood is selected. The Sobel edge value at pixel (i, j) and the average Sobel edge values of all the pixels in block (i, j) are used as the edge-preserving weights $w_{i,j}$. In Eqn. (3.3.4), κ is a small constant to avoid dividing by zero and maintain the stability of the cost function. Here we set $\kappa = 1$.

The motivation behind using edges for matching is to predict the disparity information of the moving edges such that the accuracy and robustness of the estimation can be increased. The edges play an important role in improving the image quality. The human visual system recognizes objects by temporal filtering of edge signals and one major advantage of using edges is that they are tied more closely to physical features. However, one problem of the conventional block matching disparity estimation is that the disparity vectors of the blocks located on the boundaries of moving objects cannot be estimated accurately. The smoothness term further smooth out the disparity changes at edges. Thus, we use the edge values as the weights so that discontinuities at edges are not over-smoothed.

The edge values weights function as follows. When the difference of the edge values between the current block and the neighboring blocks is small, the disparity values within that region should also be smooth, and thus $|\nabla^2(\hat{d}(i, j))|$ will be emphasized. Otherwise, if the edge values are very different, there will also exist a jump in the disparity values,

in this case the similarity cost f_1 will be dominating. Thus the smoothness cost f_2 is inversely proportional to the difference of edge values.

The regularization process typically iterates for four or five times. The disparity values are initialized to zero so that in the first iteration f_2 will be zero and the regularization process retrogresses to FSBM, where only the similarity cost f_1 is minimized. As mentioned, the algorithm can also be applied to 2D motion fields. The smoothness term will be the sum of the horizontal component and the vertical component and the similarity term will remain the same.

3.3.3 Determine the Lagrange Multiplier

The Lagrange multiplier λ in the proposed energy function 3.3.2 is very important since it determines the extent of smoothing as a tradeoff between distortion and bit rate, which are directly related to the quality of the reconstructed image and the coding efficiency respectively [45]. For larger λ , the disparity map will be smoother so that the reconstruction quality will be lowered and consequently the bit rate required for encoding the disparity map will be smaller, and vice versa.

The determination of the optimal Lagrange multiplier under the target bit rate has been extensively studied [19, 46, 99]. The problem can be formulated as looking for an optimal Lagrange multiplier λ which can minimize the cost function:

$$Cost(\lambda) = Dist(\lambda) + \lambda Rate(\lambda) \quad (3.3.6)$$

Note that this λ is the same as in Eqn. (3.3.2), thus $Dist$ and $Rate$ here are obtained by estimating the disparity field using Eqn. (3.3.2), and they are functions of λ . Here, $Cost$ is the total cost, $Rate$ refers to the coding bit rate of the smoothed displacement field using λ and $Dist$ is the corresponding distortion. The minimization problem can be solved using any bisection search algorithm and we use the convex hull method. In

our experiments here, the bit rate is measured by bits per block (bpb) for coding the disparity field, and the distortion is measured by RMSE (root of mean square error) of the reconstructed image comparing to the original image.

The process to calculate the optimal λ under a given bit rate $Rate_{target}$ is as follows:

1. We define two values of λ , $\lambda_0 < \lambda_1$, such that $Rate(\lambda_0) > Rate(\lambda_1)$ and $Dist(\lambda_0) < Dist(\lambda_1)$.
2. The next value of λ to be tested by iteration is:

$$\lambda_{new} = \frac{Dist(\lambda_1) - Dist(\lambda_0)}{Rate(\lambda_0) - Rate(\lambda_1)} + \epsilon \quad (3.3.7)$$

where ϵ is a vanishingly small positive number.

3. If $\lambda_{new} < \lambda_0$, λ_{new} is increased by $\frac{\lambda_1 - \lambda_0}{2}$. However, if $\lambda_{new} > \lambda_1$, λ_{new} is decreased by $\frac{\lambda_1 - \lambda_0}{2}$.
4. The procedure is terminated if the difference between $Rate(\lambda_{new})$ and $Rate_{target}$ is within an error bound, and the displacement vector field for λ_{new} is the optimal solution.
5. If $Rate(\lambda_{new}) > Rate_{target}$, $\lambda_0 = \lambda_{new}$, otherwise $\lambda_1 = \lambda_{new}$. Then go to Step 2.

3.3.4 A Fast Implementation

As observed, the disparity fields for the low textured regions (also referred to as homogeneous regions) become very smooth after regularization. To reduce the complexity, we separate the textured and homogeneous regions first and apply a simpler regularization scheme to the homogeneous regions.

For the detection of homogeneous regions, the Sobel edge operator is applied to the original image to obtain an edge value for each pixel, thus forming a Sobel edge map.

Otsu's threshold selection method [100] is then applied to the Sobel edge map to determine a threshold Th_{Otsu} . The homogeneous regions are defined as pixels whose edge values and the edge values of their four neighborhoods are lower than Th_{Otsu} and the other pixels are defined as textured regions. For block-based case, the average edge value of all pixels in this block is compared with Th_{Otsu} to determine the region type.

To calculate the disparity vector of a block in the homogeneous regions, the average of the four neighborhood disparity vectors is tested. If the similarity error f_1 obtained by this vector is within the error bound, which is determined empirically, it is regarded as the true displacement vector. Otherwise the vector of this block will be calculated the same way as those for textured regions.

Note that the first iteration of regularization for the fast algorithm, i.e., the initialization, is not exactly FSBM since the disparity vectors for homogeneous and textured regions are calculated differently.

3.4 Joint Disparity/Motion Estimation and Regularization

The joint estimation of disparity and motion fields is formulated as a maximum a posterior problem. In this section, firstly we build up a global cost function for the joint regularization and then use local optimization to approximate the global optimization process.

3.4.1 The Global Cost Function

As mentioned, the first iteration of the regularization process can be regarded as FSBM, after we obtained the current disparity, the left and the right motion fields using this method, the next disparity field is initialized using the loop constraint Eqn. (2.3.1). This

Figure 3.3: Two successive image pairs and corresponding blocks/pixels.

initialization step provides coarse displacement fields, especially for the next disparity field, in which errors are accumulated. They are refined in the joint regularization stage.

As shown in Figure 3.3, in the description of the proposed joint regularization algorithm, the original images are noted as

I_1 : the left image at time $t + 1$,

I_2 : the right image at time $t + 1$,

I_3 : the left image at time t ,

I_4 : the right image at time t .

The two disparity fields and the two motion fields associated with I_1, I_2, I_3, I_4 are named as

D_c : current disparity field, at time t , from left to right,

D_n : next disparity field, at time $t + 1$, from left to right,

M_l : the left motion field, from time $t + 1$ to t ,

M_r : the right motion field, from time $t + 1$ to t .

The purpose of the joint estimation is to maximize the conditional probability:

$$\begin{aligned} & \max_{M_l, M_r, D_c, D_n} P(M_l, M_r, D_c, D_n | I_1, I_2, I_3, I_4) \\ & \Leftrightarrow \max P(I_2, I_3, I_4 | M_l, M_r, D_c, D_n, I_1) P(M_l, M_r, D_c, D_n | I_1) P(I_1) \end{aligned} \quad (3.4.1)$$

and is equivalent to the minimization of the energy function

$$\min(F(D_n) + F(M_l) + F(M'_r)) \quad (3.4.2)$$

In the joint estimation, the estimation of M_r is not based on the original image I_2 , but on I'_2 which is reconstructed from I_1 and D_n , so we use M'_r instead of M_r in Eqn. (3.4.2).

3.4.2 The Implementations

Since the temporal and spatial complexity of finding the optimal solution for the global energy function (3.4.2) is very high, we solve the problem step by step and obtain local optimal results instead of the global optimal result to reduce the computational complexity. The joint regularization procedure is described as follows. Assume that the four points P_1, P_2, P_3, P_4 in the four images I_1, I_2, I_3, I_4 correspond to the same object point as shown in Figure 3.3. The two paths to obtain the coordinates of P_4 from P_1 via the displacement fields are

$$\begin{aligned} P_1(i, j) &\rightarrow P_2(i, j + D_n(i, j)) \\ &\rightarrow P_4(i + M_{rv}(P_2), j + D_n(i, j) + M_{rh}(P_2)) \\ &\equiv P_4(x_1, y_1) \end{aligned} \quad (3.4.3)$$

$$\begin{aligned} P_1(i, j) &\rightarrow P_3(i + M_{lv}(i, j), j + M_{lh}(i, j)) \\ &\rightarrow P_4(i + M_{lv}(i, j), j + M_{lh}(i, j) + D_c(P_3)) \\ &\equiv P_4(x_2, y_2) \end{aligned} \quad (3.4.4)$$

The subscript v and h represent the vertical and the horizontal components of the motion field, respectively.

The results of these two paths should be identical, which manifests in another form the loop constraint. From Eqn. (3.4.3) and Eqn. (3.4.4) we define a loop cost F_{loop} . The F_{loop} for one point (e.g., $P_1(i, j)$) is noted as f_{loop} and defined as:

$$f_{loop}(i, j) = g(|x_1 - x_2| + |y_1 - y_2|) \quad (3.4.5)$$

and

$$g(x) = \begin{cases} x & \text{if } x > TH_{cc} \\ 0 & \text{otherwise} \end{cases} \quad (3.4.6)$$

TH_{cc} is a predefined threshold, which was set to 4 experimentally in our work. Ideally, the f_{loop} should be zero.

The joint regularization process is implemented in an iterative way. Each iteration can be described by the following steps:

1. Assume the motion vectors are fixed and we update the disparity fields. For every (block/pixel) P_1 :
 - (a) The corresponding P_3 is fixed;
 - (b) Minimize the cost $F(D_n) + F(D_c) + F_{loop}$, where $F(D_n)$ and $F(D_c)$ are defined through Eqn. (3.3.2) to Eqn. (3.3.5), applied on disparity fields, and F_{loop} is calculated by Eqn. (3.4.5).
2. Assume the disparity values are fixed and we update the motion fields. For every (block/pixel) P_1 :
 - (a) The corresponding P_2 is fixed;
 - (b) Minimize the cost $F(M_l) + F(M_r) + F_{loop}$, where $F(M_l)$ and $F(M_r)$ are defined through Eqn. (3.3.2) to Eqn. (3.3.5), applied on motion fields.
3. Regularize D_c and M_r separately for one iteration as described in Section 3.3, with the same λ .

In the first two steps, the test set for finding the best disparity/motion vectors is composed of those values from neighboring pixels/blocks. Since iterations are done by scanning the next left image, the next disparity field and the left motion field are scanned pixel by pixel so that every value can be tested and updated. However, for the current disparity and the right motion fields, some values will be tested several times while some will not be tested and updated, thus they degrade the overall smoothness and lead to an increase in entropy. For this reason, we consider the third step above necessary to obtain well-smoothed current disparity and right motion fields. Another issue is the order we update the disparity and motion fields. We consider that the initial motion fields to be more reliable than the initial disparity field since motion vectors are smaller than disparity vectors in most applications.

The loop cost plays a very important role in the joint disparity and motion estimation algorithm. In the separate regularization scheme, since matching by only the similarity term will produce some unreliable results, especially in the case of inadequate texture information or repetitive texture, the smoothness cost is added to the total cost function. However, at edge regions where the edge values of the current pixel/block and its neighborhoods are very different, the smoothness cost will be insignificant. Thus the loop cost is used to remove such outliers. By checking the four vectors in the four displacement (disparity or motion) fields that are associated with the same object point, the loop cost can enhance the matching confidence and provide true displacement vectors.

3.5 Experimental Results and Analysis

We evaluate our proposed separate disparity regularization algorithm using the PIANO image pair and MAN stereoscopic video sequences. The joint disparity/motion estimation algorithm is tested using the stereoscopic video sequences TRAIN & TUNNEL and MAN. The disparity estimation problem is simplified into one-dimensional problem under the

epipolar constraint for parallel cameras. The search range for disparity estimation is ± 16 pixels in the horizontal direction and for the motion fields, the search range is ± 8 pixels in both directions.

We implemented the separate regularization and the joint regularization processes for block-based disparity and motion fields and the block size is 8. The results of the separate disparity regularization are compared with two existing algorithms by the bit rate for coding the regularized disparity field and the PSNR, which is obtained by reconstructing the original image using the disparity field. After that, we compared our separate regularization scheme with the joint regularization scheme, considering the computational complexity besides bit rate and PSNR.

3.5.1 Evaluation of Disparity Regularization

Firstly we compare our proposed separate disparity regularization algorithm with two existing disparity estimation algorithms. The cost functions for both the two existing algorithms are composed of a similarity term, which is the same as the similarity term (i.e., MAD) in our proposed method, and a smoothness term. The existing algorithm 1 uses the second derivative of the vector field as the smoothness term without edge-preserving property, i.e., the smoothness term is defined as $f_2 = |\nabla^2 \hat{d}(i, j)|$. The existing algorithm 2 was proposed by Kim [45] and the smoothness term also considers the edges, but in a different way.

Although we adopt the block-based method for video coding, the pixel-based method is also implemented to demonstrate the superior performance of the proposed algorithm. Both block- and pixel-based disparity maps obtained by the proposed separate regularization algorithm and existing algorithm 1 are shown in Figure 3.4 and Figure 3.5. The resulting disparity maps of the existing algorithm are noisy in homogeneous regions while over-smoothed or blurred at object edges. Our proposed method obtains very homoge-

Figure 3.4: Block- and pixel-based disparity maps by the existing and the proposed algorithms at $\lambda = 10$, PIANO.

neous vectors for low-textured regions, the depth of the objects can be well distinguished and the edges are clearly preserved.

The PSNR and the coding bit rate of the disparity maps at different λ values for the three algorithms are shown in Table 3.2. The results for the Existing AL 2 are copied directly from [45], and the results for the proposed algorithm are generated by inputting the desired bit rate for coding the vector field (bpb) to search for the optimal λ . The outputs include the PSNR, the obtained Lagrange multiplier, and the number

(a) Block-based disparity, existing algorithm

(b) Pixel-based disparity, existing algorithm

(c) Block-based disparity, proposed algorithm

(d) Pixel-based disparity, proposed algorithm

Figure 3.5: Block- and pixel-based disparity maps by the existing and the proposed algorithms at $\lambda = 10$, MAN.

Table 3.2: Comparison of the proposed separate regularization algorithm with existing algorithms, on block-based disparity for PIANO.

Proposed Separate Regul.				Existing AL 1			Existing AL 2		
bpb	PSNR	λ	Iterations	bpb	PSNR	λ	bpb	PSNR	λ
2.81	28.61	1	5	3.11	27.62	2	2.80	28.19	2
2.48	28.50	3	4	2.95	27.60	3	2.49	28.14	4
2.31	28.39	6	4	2.90	27.59	4	2.30	28.10	6
2.12	28.23	12	3	2.82	27.57	5	2.11	28.05	9
1.99	28.02	22	3	2.77	27.54	6	1.99	27.99	13
1.91	27.98	30	5	2.73	27.51	8	1.90	27.92	18
1.86	27.90	38	5	2.71	27.48	10	1.86	27.87	21
1.82	27.81	46	4	2.62	27.42	14	1.83	27.81	25

of iterations required. For the Existing AL 1, however, we selected several values of the Lagrange multiplier to obtain the results. Obviously, smoothing by only the second derivative term (Existing AL 1) performs much worse than the edge-preserving methods (Existing AL 2 and the proposed algorithm). As can be seen from Table 3.2, at similar bit rate for the disparity map, the quality of the reconstructed image obtained by our proposed method is better comparing to the two existing methods.

To select the Lagrange multiplier at certain bit rate, experiments showed that four or five iterations are needed before the optimal Lagrange multiplier is obtained that provides the desired bitrate. This demonstrates that the proposed regularization process is consistent with the rate-distortion theory.

Furthermore, we tested the proposed regularization scheme on block-based motion estimation and the resulting motion fields are shown in Figure 3.6. The horizontal motion fields are also presented as gray scale images by cropping and stretching the values in the range of $[0, 255]$ to show more clearly the improvements. Thus, the proposed algorithm also provides more smooth and reliable vector fields for motion estimation.

From these results, we can confirm that the proposed disparity/motion estimation and

(a) Block-based motion by existing algorithm for TRAIN & TUNNEL

(b) Block-based motion by proposed algorithm for TRAIN & TUNNEL

(c) Block-based motion by existing algorithm for MAN

(d) Block-based motion by proposed algorithm for MAN

(e) Horizontal motion field by existing algorithm for MAN

(f) Horizontal motion field by proposed algorithm for MAN

Figure 3.6: Comparison of block-based motion fields by the existing and the proposed algorithms at $\lambda = 10$.

Figure 3.7: Comparison of proposed the separate and the joint regularization algorithms, (a) TRAIN & TUNNEL and (b) MAN.

Table 3.3: Computational complexity analysis of the proposed algorithms.

		Separate Regularization	Joint Regularization
Theoretical	Addition	731.125	442.125
	Multiplication	1.25	1.25
	Absolute	728.4688	439.4688
Actual	Addition	640.568	420.0605
	Multiplication	0.9942	0.7854
	Absolute	631.8628	411.9824

The above results are on a block-basis.

regularization method provides vector fields that are both objectively and subjectively better than the existing methods and are very suitable for video coding applications.

3.5.2 Evaluation of Joint Disparity and Motion Estimation

The comparison of the separate and the joint regularization is shown in Figure 3.7 and the computational complexity is analyzed in Table 3.3.

The complexity is measured by the theoretical analysis and the actual numbers of additions/subtractions, multiplications, and absolute conversions. The separate regu-

larization column shows the average values of complexity of separate regularizing two disparity fields and two motion fields. The joint regularization column is the average complexity for regularizing the four fields simultaneously. The parameters for calculating the complexity include the search range, the block size and the definition of neighborhood, which determine the number of test vectors. In actual computation, since some of the test values might be the same, the cost function will be calculated only once so that the computation time might be reduced compared to this analysis. The actual complexity values are obtained on the TRAIN & TUNNEL sequences, with $\lambda = 10$ for all cases. Experiments are performed on a PC platform with a 2.8 GHz P4 CPU. As shown in Table 3.3, the complexity of the joint regularization scheme is less than two third of that of the separate regularization scheme, which is achieved by calculating the fourth field by the other three using Eqn. (2.3.1).

To compare the reconstruction PSNR and bit rate of our joint regularization scheme with the separate one, firstly we calculate the four vector fields of two successive image pairs using the separate regularization algorithm. We calculate the average reconstruction PSNR and the average bit rate of the four vector fields at the same λ . Then, we also calculate the average PSNR and bit rate of the four vector fields obtained by the joint regularization scheme. In Figure 3.7, the two curves are the PSNR-bit rate curves for these two methods obtained at different λ values. As shown in Figure 3.7, the average quality of the reconstructed images obtained by the joint regularization scheme is about 0.2 to 0.3 dB lower in PSNR than those obtained by the separate regularization scheme at same bit rate. The reason is that, while minimizing the two costs simultaneously together with the loop constraint will provide true disparity and motion vectors, it also leads to larger distortion since the importance of similarity is lowered. However, for our proposed application area such as videoconferencing systems, we can regard the quality as fairly acceptable. In addition, the residual coding in video coding systems will compensate the

image quality.

These results show that the proposed joint regularization scheme reduces the computational complexity while approximately maintains the image qualities. The performance is promising for stereoscopic and multiview video coding applications.

Chapter 4

MPEG-4 Compatible Multiview Video Coding

Based on the joint disparity and motion estimation algorithm, we propose an MPEG-4 based stereoscopic and multiview video sequences encoder in this chapter. Firstly, we give a short analysis on the redundancies exist in multiview video data in Section 4.1, and then a brief review on the MPEG-2 MVP is presented in Section 4.2. The structure of the proposed codec is described in Section 4.3. After that, various frame prediction structures are comparatively studied in Section 4.4 and an extension of the monoview video coding rate control scheme is proposed in Section 4.5. Finally the experimental results and concluding remarks of the proposed encoder are given in Section 4.6.

4.1 Redundancies in Multiview Video Data

In information theory, *redundancy* is the number of bits used to transmit a message minus the number of bits of actual information in the message, and *correlation* is a measure of the strength of linear relationship between two random variables. When the two signals are highly correlated, there also exists high redundancy when representing the two signals since one can be predicted by the other.

The motivation of data compression is to decorrelate and remove redundancies of the input signals. In multiview view video data, there exist four types of correlations.

1. *Spatial correlation*, which refers to the likeness between adjacent pixels within each image,
2. *Temporal correlation*, which means that most pixels of two temporally adjacent image frames stay the same,
3. *Spectral correlation*, which refers to the smoothness of frequency components,
4. *View correlation*, which means that most pixels of pictures captured from slightly different angles remain the same.

The multiview video data can be described as “highly correlated” or “highly redundant”, and any efficient compression algorithm should take all kinds of correlations into considerations.

4.2 MPEG-2 Multiview Profile

As an addition profile developed for MPEG-2, the MPEG-2 multiview profile provides a standard scheme for stereoscopic video coding by using the temporal scalability tools. The temporal scalability in MPEG-2 refers to layered coding of two or more layers. Each layer has same or different temporal resolutions and same spatial resolution. The temporal scalability builds on the block motion compensated DCT coding structure of non-scalable video coding in MPEG-2 by adding interlayer prediction. Using the temporal scalability tools, the MPEG-2 MVP is a two-layer video coding scheme, i.e., the base layer and the enhancement layer. The base layer video is coded as an MPEG-2 main profile bitstream while the enhancement layer video exploits the correlation between the two views to improve the compression efficiency.

Figure 4.1: Two configurations of the MPEG-2 Multiview Profile.

There are two major configurations in the MPEG-2 MVP as shown in Figure 4.1. Configuration 1 uses the disparity-compensated prediction, i.e., the enhancement layer adopts a picture structure composed of only B pictures, using both forward and backward reference with respect to temporally adjacent decoded base layer pictures. The macroblock modes for B pictures in the enhancement layer are similar to that for non-scalable coding, including forward, backward and bidirectional prediction modes.

Configuration 2 adopts a disparity-/motion-compensated prediction scheme, where the enhancement layer uses a picture structure mainly composed of B pictures except that the first picture is a P picture. One of the two predictions for enhancement layer B pictures is motion compensated prediction from decoded immediate previous enhancement layer picture, and the other prediction is obtained w.r.t. the immediate previous decoded base layer picture in display order. Thus, the forward mode implies reference w.r.t the previous decoded enhancement layer picture (by motion) whereas the backward mode implies reference w.r.t. the base layer (by disparity).

Experimental results show that configuration 2 achieves better coding performance than configuration 1, and increases the image quality about 1.6 dB higher in PSNR compared to the simulcast MPEG-2 coding scheme [101, 102]. In this research, since our proposed encoder is based on MPEG-4, for fair comparison, we implemented the MPEG-2 MVP method on the MPEG-4 platform, and refer to this combined working platform

in this thesis as MPEG-4 MVP. Configuration 2 is adopted for the picture reference structure.

4.3 Encoder Structure

To enable compatibility with existing monoview video coding standards, the proposed stereoscopic and multiview video encoder is based on the MPEG-4 platform. There are 19 visual profiles defined in MPEG-4 [103], and our scheme is based on the *Core Visual Profile* [104]. The reason is that the *Core Visual Profile* supports object-based encoding by the concepts of VOPs and arbitrary shape coding, which is suitable for head-and-shoulder video sequences typically used in videoconferences.

The block diagram of the proposed encoder for coding two views is shown in Figure 4.2. As shown in the figure, the main view is encoded using an MPEG-4 encoder while the only difference is that the motion vectors are calculated coherently in the joint disparity and motion estimation module. Thus the main view bitstream can be decoded using the MPEG-4 decoder. The auxiliary views are encoded by joint motion and disparity compensation. Since the encoder is based on VOPs, the shape coding in MPEG-4 is adopted to encode the contours of the VOPs. For the auxiliary views, the shape is also encoded after disparity or motion compensation. The disparity and motion vectors are losslessly encoded by DPCM (Differential Pulse Code Modulation) entropy coding using a predefined VLC table. The intra frames and the residual images are encoded by DCT transform, quantization and run-length coding (RLC) of the DCT coefficients, etc.

Since the theoretically optimal GOP structure can be defined only under certain application requirements and source data characteristics, in our proposed encoder, user can define the GOP structure by setting the M and N parameters. N is the intra distance which is the length of GOP, and M is the prediction distance. An example of the GOP

Figure 4.2: The block diagram of the MPEG-4 compatible stereoscopic video encoder.

Figure 4.3: GOP structure of the proposed stereoscopic video encoder.

structure is shown in Figure 4.3, where $N = 9$ and $M = 3$.

As shown in Figure 4.3, we introduce new picture types I_D , P_D and B_D for the auxiliary view, where the subscript D means that the picture is also predicted from the main view images by disparity compensation. Thus, I_D frames are predicted by disparity field from the corresponding I frames in the main view and P_D , B_D frames are predicted jointly by disparity and motion fields from both views.

Different MB prediction modes of texture data for different VOP types are defined in Table 4.1. After joint disparity and motion estimation, all possible MB modes for

Table 4.1: VOP types and MB prediction types.

VOP Type		MB Prediction Type			
Main View	I	Intra	-	-	-
	P	Intra	Inter	Inter4V	-
	B	Direct	Interpolate	Forward	Backward
Auxi. View	I_D	Intra	Inter	Inter4V	-
	P_D, B_D	Interpolate	Motion	Disparity	-

the current macroblock are checked and the prediction mode that obtains the minimum distortion is selected. For P and P_D frames, only one motion (disparity) vector has to be encoded for one macroblock for Inter mode, and four motion (disparity) vectors are encoded for Inter4V mode. For the Interpolate mode of B_D frames, both the disparity vector and the motion vector are encoded. For B frames in the main view, there is no need to encode any motion vectors for Direct mode since it can be calculated from the two reference frames. Although the Direct mode is very efficient for decorrelation of the video sequence with constant motion, it does not apply for the B_D frames since there is no continuity in motion and disparity.

The proposed encoder has an open-loop structure because the joint estimation is performed on the original images. To obtain a close-loop implementation so that the results are reproducible at the decoder, refinement of the disparity and motion vectors is required. However, this may affect the smoothness of the vector fields achieved by the joint estimation and regularization. As a tradeoff, the refinement is performed within a small search range with half-pixel accuracy during the encoding process using the reconstructed reference images.

The encoder works similarly for multiple views for each pair of corresponding main view and auxiliary view. In this thesis, 3-view and 5-view video coding are considered.

4.4 Optimal Reference Selection

In most previous works, the reference frames used to predict a desired frame were both fixed and heuristically chosen. These reference frames do not necessarily yield the best prediction and compression performance suffers accordingly. Prediction performance is related to the correlation between the two frames. For one video sequence, the temporally adjacent frames will be most similar, but it is not as straightforward for multiview video data.

To keep compatibility with MPEG-4, we only consider the adaptive reference frame selection for the auxiliary views in multiview video coding. The problem can be considered from two aspects, i.e., the view-level reference selection and the frame-level reference selection.

4.4.1 View-level Reference Selection

View-level reference structure refers to the problem of given a (n, m) pair, where n is the total number of views and m is the number of main views, how to select the main views. In this research, we only consider the situation that all the views are captured by parallel cameras with equal baseline distance between the adjacent pairs. Thus, the basic principle for selecting the reference picture is based on the fact that the further the distance between the two views is, the more occlusion occurs and the worse the disparity estimation will be.

For 2-view video coding, the configuration is quite straightforward, i.e., the left view is coded as the main view and the right view is coded as the auxiliary view. The configuration for 3-view video can be regarded as special cases of 5-view video, thus the view level reference structure is investigated for 5-view video here. We consider four configurations as shown in Figure 4.4, in which the bold lines represent the main views and the others

Figure 4.4: View-level prediction structures for 5-view video coding.

are the auxiliary views predicted from the main views.

Config 1 adopts directly the idea from the GOP structure in monoview video coding. In particular, in Config 1, the rightmost view 4 is predicted from the leftmost view 0, and the other intermediate views are predicted bi-directionally from both view 0 and view 4. Since B frames in motion video sequences has the highest compression ratio, it is reasonable to expect that the three middle views in Config 1 can be efficiently predicted. Thus, the Config 1 is expected to perform well especially when the disparity between the views is very small.

In both Config 2 and Config 3, the middle view is coded as main view and the difference is whether the outmost views are predicted from the main view or the adjacent auxiliary views. However, both schemes have their advantages. For Config 2, since the main view is normally encoded at a higher quality than the auxiliary views, predicting from the main view will give better performance. For Config 3, the disparity vectors will be smaller by using the adjacent views.

In Config 4, there are two main views and each auxiliary view is predicted by its adjacent main view(s). This will be suitable for multiview video sequences captured with very large baseline between adjacent views.

Figure 4.5: Frame-level prediction structures for 5-view video coding.

4.4.2 Frame-level Reference Selection

Frame-level reference structure considers the selection between disparity and motion vectors, forward motion and backward motion vector, etc. This is similar to the multiple reference structure in H.264/AVC coding. The selection is done at macroblock level based on the minimum distortion criterion. As shown in Figure 4.3, the auxiliary view frames are predicted either by disparity alone (I_D frames) or jointly by disparity and motion fields (P_D and B_D frames). In Config 1 and Config 4, a new kind of auxiliary view which is predicted by two main views is introduced. For this auxiliary view, the I_D frames are predicted by bi-directional disparity and the P_D/B_D frames are predicted by two disparity fields and one motion field, as shown in Figure 4.5.

4.5 Rate Control

The rate control algorithm for video coding cannot be directly applied to multiview video coding since more picture types are included in the multiview video. In this section, we generalize the rate control algorithm in MPEG-4 Verification Model 18.0 [104] for multiview video coding.

Let K be the total number of different picture types in multiview video coding. The

R-Q model defined in [85] is applied to the type- k picture as:

$$R_k = a_k Q_k^{-1} + b_k Q_k^{-2}, \quad (k=1,2,3,\dots,K) \quad (4.5.1)$$

Here, R_k and Q_k are the estimated bits and average quantization parameter for type- k picture, respectively. a_k and b_k are the model parameters. Similar to the solution for monoview video coding, the encoder also collects the bit rate and average quantization step for each type of picture at the end of encoding each frame for MVC. The model parameters (a_k, b_k) for $(k = 1, 2, 3, \dots, K)$ can be estimated by the linear regression method and the following formula

$$\begin{aligned} a_k &= \frac{\sum_{i=1}^{N_k} Q_i R_i - b_k \sum_{i=1}^{N_k} Q_i^{-1}}{N_k} \\ b_k &= \frac{N_k \sum_{i=1}^{N_k} R_i - (\sum_{i=1}^{N_k} Q_i^{-1})(\sum_{i=1}^{N_k} Q_i R_i)}{N_k \sum_{i=1}^{N_k} Q_i^{-2} - (\sum_{i=1}^{N_k} Q_i^{-1})^2} \end{aligned} \quad (4.5.2)$$

Here, N_k is the number of type- k pictures observed in coding the previous frames. Q_i and R_i represent the actual bits and average quantization parameters in the past.

Similar to the solution proposed for MPEG-4 monoview video coding rate control, We can assume that there is a constant ratio between the average quantization parameters between different picture types.

$$\frac{C_1}{Q_1} = \frac{C_2}{Q_2} = \dots = \frac{C_K}{Q_K} = c \quad (4.5.3)$$

Let R denote the usable bits in the current GOP, which should be distributed among the K picture types each of N_k frames.

$$\sum_{i=1}^K N_i R_i = R \quad (4.5.4)$$

By Eqn. (4.5.1), (4.5.3) and (4.5.4), we can solve Q_k for picture type- k by:

$$\alpha_k Q_k^{-2} + \beta_k Q_k^{-1} + \gamma_k = 0 \quad (4.5.5)$$

where,

$$\begin{aligned}\alpha_k &= \sum_{i=1}^K b_i N_i \left(\frac{C_k}{C_i}\right)^2 \\ \beta_k &= \sum_{i=1}^K a_i N_i \left(\frac{C_k}{C_i}\right) \\ \gamma_k &= -R\end{aligned}\tag{4.5.6}$$

To solve Eqn. (4.5.5), we define δ_k as

$$\delta_k = \beta_k^2 - 4\alpha_k\gamma_k\tag{4.5.7}$$

If $\delta_k < 0$, the quadratic function is downgraded to simple equation and

$$\begin{aligned}Q_k^{-1} &= \frac{-\gamma_k}{\beta_k} \\ R_k &= \alpha_k Q_k^{-1}\end{aligned}\tag{4.5.8}$$

Otherwise,

$$\begin{aligned}Q_k^{-1} &= \frac{\sqrt{\delta_k} - \beta_k}{2\alpha_k} \\ R_k &= a_k Q_k^{-1} + b_k Q_k^{-2}\end{aligned}\tag{4.5.9}$$

4.6 Experimental Results and Analysis

4.6.1 Evaluation of Stereoscopic Video Coding

The sequences TRAIN & TUNNEL, MAN and Reading are used to evaluate the performance of the proposed MPEG-4 compatible multiview video encoder. Without loss of generality, we set the intra distance $N = 9$ frames and the prediction distance $M = 3$ frames for the GOP structure. The block size is set to 8 in all cases. MAN is a stereoscopic videoconferencing sequence thus both luminance and shape information are encoded. The background of MAN is very homogeneous and thus the images can be easily segmented.

We applied a region growing segmentation method with manual adjustments in several frames where the segmentation algorithm does not work well. For TRAIN & TUNNEL and Reading, only luminance information is encoded. The 5-view video sequences are generated from the 2-view (MAN and TRAIN & TUNNEL) or 3-view (Reading) video sequence using the IVR algorithm proposed in [105].

The results are compared with the MPEG-4 MVP scheme and the Smooth DE scheme, as in Figure 4.6. The Smooth DE scheme, taking the same reference structure as the proposed scheme, adopts the second derivative smoothing disparity estimation scheme as described and evaluated in Chapter 3. In Figure 4.6, the bitrate is the sum of bits per second (bps) for the two views and the PSNR is the overall value calculated from the average distortion for all the frames in all the views. As shown in the figure, the proposed encoder increases the PSNR about 1 to 1.5 dB for comparing to the MPEG-4 MVP scheme at low bit rate and nearly 1 dB at higher bit rates. The gain comparing to the Smooth DE method is from 0.2 to 0.5 dB at lower bit rates and from 0.5 to 1 dB at higher bit rates. The performance gain comes from both the main view and the auxiliary views. The decoded main view quality of our proposed encoder is always better than that of the conventional method because more bits are used for encoding the main view. In addition, the quality of the auxiliary view of our proposed encoder is always better than that of the conventional encoder although less bits are used. The reason is twofold. Firstly, the prediction frames from the main view in our proposed encoder has higher quality which can provide more efficient estimation and compensation. Secondly, in our method, disparity and motion fields for the auxiliary views are obtained from joint regularization but not by full search, thus the bits for coding the vectors for P_D and B_D frames are reduced.

The Smooth DE scheme performs better than the MPEG-4 MVP at lower bit rates because the disparity and motion fields are smoothed and cost less bits. However, at

Figure 4.6: Coding results of stereoscopic video sequence, (a) TRAIN & TUNNEL, (b) MAN and (c) Reading.

higher bit rates, since the smoothed displacement vector fields lose details and accuracy especially at the edges, the reconstructed PSNR values of the decoded pictures become lower than those of MPEG-4 MVP.

To compare the coding results subjectively, we also show one frame of the decoded images obtained by MPEG-4 MVP and our proposed encoder for MAN and Reading in Figure 4.7 and Figure 4.8, respectively.

Figure 4.7: Decoded auxiliary view image for MAN (a) and (b) by the proposed scheme, (c) and (d) by MPEG-4 MVP.

Figure 4.8: Decoded auxiliary view image for Reading (a) the proposed scheme and (b) MPEG-4 MVP.

Figure 4.9: Coding results of stereoscopic video sequence, (a) TRAIN & TUNNEL and (b) MAN.

4.6.2 Evaluation of Multiview Video Coding

TRAIN & TUNNEL and Reading are used to test the 3-view video coding and the coding results are also compared with MPEG-4 MVP coding scheme as shown in Figure 4.9. Our proposed scheme increases the overall PSNR about 1 dB for TRAIN & TUNNEL and nearly 2 dB for Reading. Compared with stereoscopic video coding, the performance gain for coding three views is slightly increased, this is reasonable since the gain comes only from the other auxiliary view. Actually, assuming the compression ratio of the main view and the auxiliary view, the performance gain for coding five views can be predictable if the number of the main and the auxiliary views are known. Thus, we only compare the coding results of the four configurations for 5-view video coding.

The comparison of the four configurations in view-level prediction for 5-view video coding is shown in Figure 4.10 for TRAIN & TUNNEL and Reading. For Reading, since the disparity range is very large, Config 4, where view 1 and view 3 are encoded as main views, obtains the best result. Config 1 has the worst result since the direct disparity between the two outmost views (view 0 and view 4) is very large and thus disparity prediction for view 4 from view 0 almost fails. For TRAIN & TUNNEL, since

Figure 4.10: Comparison of view-level prediction structures for 5-view video coding, (a) TRAIN & TUNNEL and (b) Reading.

the disparity range is small, Config 1 has the best result, which is more obvious at higher bitrates. Config 4 has the worst result since two main views cost a lot of bits. For Config 3, the results are unstable and largely depend on the rate control algorithm. The reason is that, in Config 3, view 1 and view 3 are predicted from the main view, view 2, but they are also the reference views for view 0 and view 4 respectively. The reconstructed image qualities of view 1 and view 3 have great effects on the image qualities of view 0 and view 4.

From these results, we confirm that the best view-level prediction structure very much depends on the properties of the multiview video data, i.e., the disparity range. For a small baseline distance in the applications such as robotic stereovision, Config 1 will be a good choice. For a large baseline distance, as in videoconferencing, Config 4 can achieve better performance. In order to fairly evaluate the different prediction structure, the properties of the video data have to be taken into considerations.

Chapter 5

Scalable Multiview Video Coding Using 4D Wavelet

In this chapter, we propose a new framework of MVC using high dimensional wavelet, which perfectly matches the inherent structure of the high dimensional video data.

Although the existing MVC schemes do exploit the correlations between adjacent views to some extent through inter-view prediction, there are some constraints for practical applications. Firstly, since temporal correlation is always stronger than view correlation in most natural video sequences, the multi-reference coding structure cannot efficiently remove the view redundancy. So far, no significant improvement in coding efficiency of MVC has been reported in the literature. Secondly, these schemes are difficult to extend to code an arbitrary number of views. The GGOP (Group of GOP) structure adopted in most of the MPEG-compatible schemes has to be fixed and pre-defined. Thirdly, in conventional multi-reference schemes, the saving in bit rate comes from the auxiliary views which are predicted by the main view. When the number of views increases, the performance gain comparing to simulcast coding is linear since only the correlation between the auxiliary view and the main view is exploited. Considering all the views together as integral will be more efficient in the overall coding performance than selecting several of them as the main views. Due to these limitations, a more effi-

cient MVC scheme is desirable not only for high coding performance, but should also be easily extendable, scalable and flexible for real applications.

In this chapter, we propose an MVC framework using high dimensional wavelet transform. The video data in multiview applications is high dimensional. For example, using a 2D camera array, the data captured has five dimensions, including 2D view, 1D temporal and 2D spatial. Wavelet, with its intrinsic high multi-resolution structure, is more feasible to extend to high dimensions than DCT. It provides a more natural and flexible method for coding multiview video data. However, the existing MCTF cannot be applied directly to the view direction since disparity is inherently different as temporal motion. The block-based motion model cannot well represent the differences introduced by different camera positions. Thus, we propose a disparity compensated view filter based on the generic lifting transform. In addition, DCVF and MCTF are adaptively utilized to decompose the multiview video data into 4D wavelet coefficients according to the coding cost.

In Section 5.1, a unified lifting structure for image sequences using arbitrary transform models is defined. In particular, same as the conventional MCTF, the block-based motion model is used in the lifting along the temporal direction. Moreover, an adaptive global and local disparity model is proposed for the lifting along the view direction, i.e., DCVF. In Section 5.2, through analyzing the temporal and view correlations, we propose an adaptive decomposition structure. The entire system of the proposed MVC scheme is then introduced in details in Section 5.3. Extensive experimental results and concluding remarks are given in Section 5.4.

5.1 Lifting based Disparity Compensated View Filtering

The lifting technique is an efficient implementation of wavelet transform with low memory and computational complexity, which has been incorporated with motion compensation and formed the MCTF by some researchers in the literature [106, 107]. In this section, we first introduce the generic lifting transform, followed by the proposed DCVF for view decomposition.

5.1.1 Generalized Lifting Transform

In general, the lifting-based wavelet transform is composed of two steps: prediction and updating. The former updates the odd frame of a video sequence to high-pass subbands and the latter updates the even frames to low-pass subbands. Here we use the 5/3 bi-orthogonal filter as an example to discuss the generic lifting transform. The generic lifting transform can also use Haar and the 9/7 filter. Let $I_{2k}[n]$ and $I_{2k+1}[n]$ represent the even and the odd frames in a video sequence. Without taking pixel alignment into account, the lifting steps can be formulated by

$$\begin{aligned} H_k[n] &= I_{2k+1}[n] - \frac{1}{2}(I_{2k}[n] + I_{2k+2}[n]) \\ L_k[n] &= I_{2k}[n] + \frac{1}{4}(H_{k-1}[n] + H_k[n]) \end{aligned} \quad (5.1.1)$$

where $H_k[n]$ and $L_k[n]$ are the high-pass subband and the low-pass subband, respectively. The inverse transform can be easily derived from Eqn. (5.1.1).

The generic lifting transform that is applicable to a video sequence requires pixel mapping among adjacent frames, which can be obtained from a motion model. Let $M_{k_1 \rightarrow k_2}$ denotes the mapping from frame I_{k_1} to I_{k_2} , which is usually related to a transform model, e.g. the block-based translational motion model, to compensate for the difference

between two frames. The lifting steps can be further formulated by

$$\begin{aligned} H_k[n] &= I_{2k+1}[n] - \frac{1}{2}(M_{2k \rightarrow 2k+1}(I_{2k})[n] + M_{2k+2 \rightarrow 2k+2}(I_{2k+2})[n]) \\ L_k[n] &= I_{2k}[n] + \frac{1}{4}(M_{2k-1 \rightarrow 2k}(H_{k-1})[n] + M_{2k+1 \rightarrow 2k}(H_k)[n]) \end{aligned} \quad (5.1.2)$$

The high-pass subbands corresponding to the predictive residuals will approach zero when the motion is accurately modeled. The low-pass subbands will be very close to the original image except that some high frequency signals have been removed. Some side information that represents the pixel mapping will also be produced in addition to the high-pass and low-pass subbands. The key problem of the generic lifting transform is how to calculate the pixel mapping $M_{2k \rightarrow 2k+1}$ and $M_{2k+1 \rightarrow 2k}$ that usually describe the same motion but with reverse directions.

The generic lifting transform has been extensively studied and applied to the temporal de-composition of video sequence in the field of SVC. Similar to the traditional hybrid video coding, the block-based translational model is normally adopted in the temporal decomposition, where each block has a rigid translational motion vector [67]. In this case, the pixel mapping can be taken as the block-based motion-compensated prediction (MCP). However, there are still some issues to be considered when the generic lifting transform is used in temporal decomposition. Firstly, the sub-pixel MCP leads to the ambiguity in representing both the pixel mapping and its inverse with only one set of rigid motion vectors. Secondly, different from the traditional hybrid coding schemes, both the prediction and updating steps in lifting should be considered in order to achieve the high coding efficiency. These open issues likewise exist and may be more difficult to be solved when the generic lifting transform is performed on the video sequence with special motion models, e.g., the multiview video.

5.1.2 Global Warping-based DCVF

Although the generic lifting transform with the block-based translational model can be directly used for the decomposition of the view video sequence, the performance cannot be guaranteed and may be even far from optimized without considering the inherent features of the multiview video. As discussed earlier, the global disparity is one of the major properties of the multiview video. Therefore, we propose a warping-based lifting transform as DCVF by utilizing the global disparity, based on the assumption that the disparity between views can be perfectly represented by a global disparity model.

In the proposed scheme, we use the six-parameter affine transformation to represent the global disparity. Although theoretically the global disparity parameters can be derived from the camera parameters, we employ the global motion estimation (GME) technique proposed in [108] to calculate the model parameters for the purpose of generality. To avoid the effects of distorted image areas from the wide-angle lens of the cameras, a 16-pixel wide boundary of each frame is excluded when calculating the model parameters. Complying with the generalized lifting transform, the lifting steps in the global warping-based DCVF consists of the prediction and updating operations.

The prediction operation on the odd frame I_{2k+1} of a view sequence is achieved from the even frames I_{2k} and I_{2k+2} . According to the generic lifting transform, the key problem lies in the definition of $M_{2k \rightarrow 2k+1}$ and $M_{2k+2 \rightarrow 2k+1}$. Here we take $M_{2k \rightarrow 2k+1}$ as an instance in description. Suppose $p = (x, y)$ is the location of one pixel in frame I_{2k+1} . The vector $X = (a_1, a_2, a_3, b_1, b_2, b_3)^T$ consists of the affine transformation parameters from I_{2k+1} to I_{2k} . $T(p, X)$ denotes the transformation of p with respect to X , resulting in the location $p' = (x', y')$ in I_{2k} , i.e., $p' = T(p, X)$. Particularly, we have:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_3 \\ b_3 \end{pmatrix} \quad (5.1.3)$$

Further, we can derive the predicting value for the pixel at p in I_{2k+1} from I_{2k} to be:

$$M_{2k \rightarrow 2k+1}(I_{2k})[p] = I_{2k}[T(p, X)] \quad (5.1.4)$$

$p'(x', y')$ usually corresponds to a floating coordinate in I_{2k} , thus we use the bilinear interpolation algorithm to calculate the intensity value of $I_{2k}[p']$. Particularly, we use the four integer pixels around p' , i.e. p_0, p_1, p_2 and p_3 , as that used in MPEG-4 [104]. After the prediction step in the lifting transform, the high-pass subband H_k is produced.

The updating of the even frame I_{2k} is achieved from the high-pass subbands H_{k-1} and H_k , for the purpose of removing part of high frequency components in I_{2k} as well as improving the coding efficiency. Similarly, the key problem is the definition of $M_{2k-1 \rightarrow 2k}$ and $M_{2k+1 \rightarrow 2k}$. Here we take $M_{2k+1 \rightarrow 2k}$ as an example. Let $p = (x, y)$ represent the location of one pixel in H_k . The vector Y consists of the affine transform parameters from I_{2k} to H_k . Since H_k and I_{2k+1} have the same coordinate system, Y is exactly the inverse of X that has been used in the prediction operation. Similarly, we can derive the updating value for the pixel at p in I_{2k} from H_k to:

$$M_{2k+1 \rightarrow 2k}(H_k)[p] = H_k[T(p, Y)] \quad (5.1.5)$$

$T(p, Y)$ usually corresponds to a floating coordinate in H_k , and thus we also calculate $H_k[T(p, Y)]$ with bilinear interpolation. After the updating step, the low-pass frame L_k is produced.

The above prediction and updating operations compose the global warping-based DCVF. Figure 5.1 intuitively illustrates how the two lifting steps are performed based on the global warping. For the prediction, I_{2k} is first warped towards I_{2k+1} w.r.t the affine transform parameters to generate the warped frame I'_{2k} . Low-pass extrapolation (LPE) padding is employed to fill the blank regions in I'_{2k} that do not correspond to any pixels in I_{2k} . After the warping, I'_{2k} and I_{2k+1} are in the common coordinate system and the

Figure 5.1: (a) *Prediction* and (b) *Updating* for DCVF.

prediction of the pixel in I_{2k+1} can be directly derived from the co-located pixel in I'_{2k} . The updating can also be performed based on global warping.

Compared with the traditional MCTF, the global warping-based DCVF presents many advantages. For example, if the sub-pixel MCP is employed in the MCTF, the rigid translational model might result in the mismatch between the prediction and updating steps. However, in the proposed global warping-based DCVF, the affine transformation and its inverse can perfectly produce the co-located pixels in both prediction and updating steps, which can inherently solve the mismatch problem in MCTF.

5.1.3 Adaptive Global and Local Disparity Model in DCVF

As mentioned, the global warping-based DCVF is based on the assumption that the disparity between two adjacent views can be perfectly represented by the global disparity model. In real applications, any global disparity model cannot guarantee the perfect matching for all regions, even though in most cases they dominate the accuracy of pixel mapping between the two views. Considering the local disparity information can further refine the matching accuracy between two adjacent views. Therefore, we propose an adaptive global and local disparity model in the lifting-based DCVF.

In the adaptive disparity model, each MB corresponds to either local or global disparity mode. If the local disparity mode is selected for an MB, the local disparity vectors equivalent to the local motion vectors of the block-based translational model are used to represent the pixel mapping, and the MB is referred to as a local disparity compensation (LDC) MB hereafter; otherwise, the affine transform parameters w.r.t. the whole frame are used, and the MB is referred to as a global disparity compensation (GDC) MB hereafter. The MB type can be selected based on the tradeoff between the matching accuracy and the coding cost.

The lifting steps on the LDC MBs can be the same as that in MCTF. Particularly, we employ the barbell lifting transform in [67] for the LDC MBs. The lifting steps on the GDC MBs are derived from the previous global warping-based DCVF. Since the proposed global DCVF also complies with the generic structure of the barbell lifting, we can implement the lifting steps on the LDC and GDC MBs in a common way within the same framework but with different disparity parameters.

For the updating operation of the LDC MB, we employ the energy distributed update (EDU) technique developed in [109], in which every pixel in the high-pass subband H_k (corresponding to I_{2k+1}) updates four pixels in the even frame I_{2k} (corresponding to

Figure 5.2: High-pass subbands from the one-level decomposition of the view sequence Golf2, using (a) DCVF with GDC only, (b) DCVF and (c) MCTF.

the low-pass subband L_k) under the energy distribution constraint. To implement the updating operations on the LDC and GDC MBs within the same framework, we also derive the EDU-based updating operation for the GDC MBs as follows. Let $p = (x, y)$ denote the location of one pixel in a GDC MB of the high-pass frame H_k . Thus, p corresponds to a floating location p' in I_{2k} w.r.t. the affine transformation parameters X from H_k (i.e., I_{2k+1}) to I_{2k} . According to the routine of EDU-based updating, $H_k[p]$ should be scaled and added to every one of the four pixels around p' in I_{2k} , i.e., p'_i , $i = 0, 1, 2, 3$. The problem is how to calculate the scaling weight that corresponds to the contribution of $H_k[p]$ to the pixel at p'_i in I_{2k} . Based on the updating operation defined in the global warping-based DCVF, the pixel at p'_i in I_{2k} is updated with four pixels around its corresponding floating location $p_i = (x_i, y_i)$ in H_k . Thus, we calculate the scaling weight w_i depending on the distance between p_i and p as follows:

$$w_i = \begin{cases} (1 - |x - x_i|) \cdot (1 - |y - y_i|), & \text{if } |x - x_i| \leq 1 \text{ \& } |y - y_i| \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (5.1.6)$$

$w_i = 0$ indicates that $H_k[p]$ need not update the pixel at location p'_i in I_{2k} , because the four integer pixels around p_i in H_k does not include the pixel at p . In this way, the updating steps on the LDC and GDC MBs are implemented within the same framework.

Figure 5.2 shows the high-pass subbands produced from one-level decomposition of a

view sequence, using the DCVF with GDC only, the DCVF with both GDC and LDC and the MCTF, respectively. As shown in Figure 5.2(a), while the energies at the most regions are very small, the energy at the foreground region is still relatively large. It is because the affine transformation model cannot perfectly match both the dominated background and the small foreground regions due to the different scene depths. In this case, the LDC mode is required. As shown in Figure 5.2(b), the whole frame including both the foreground and background regions presents very low energies, which is very desirable in the sense of coding efficiency. Moreover, while the high-pass subband produced using the DCVF has the energies no more than that produced using the MCTF, as shown in Figure 5.2(c), the DCVF requires much fewer disparity parameters than the MCTF.

5.2 Adaptive Decomposition Structure

Assuming that the multiview video consists of a 2D matrix with each element indicating a video frame, a straightforward extension is to apply wavelet decomposition along the view direction by aligning pixels across different views in addition to the temporal decomposition. Towards this goal, the proposed DCVF can be jointly utilized with the traditional MCTF to decompose the 2D image matrix along view axis and temporal axis, respectively. In this section, we will firstly present a fixed decomposition structure and then propose a more efficient adaptive decomposition structure based on the correlation analysis.

5.2.1 Regular 4D Wavelet Decomposition

In the multi-level MCTF, the low-pass subband signals are used for the higher level decomposition. Based on the investigation on 3D wavelet-based video coding, it is unnecessary to decompose the high-pass subband signals once more. Therefore, it is also reasonable to leave along the high-pass subband signals and only apply the higher level

decomposition to the low-pass frames in the multi-level DCVF. Since the temporal correlation of the multiview video is usually stronger than the view correlation in lower layers, we have proposed a regular decomposition structure, in which the DCVF is performed after the multi-level MCTF has fully extracted the temporal correlations. Figure 5.3(a) shows an example of such a decomposition structure that consists of four levels of MCTF decomposition and the following two levels of DCVF decomposition. For convenience, we refer to this decomposition structure as $4T+2V$, where T and V represent the temporal decomposition and the view decomposition, respectively. The number following T or V in the figure represents the band index from the decomposition. Obviously, only the last subband, i.e., the subband at (T-LLLL, V-LL) in Figure 5.3(a), is composed of the low-pass subbands, and all the others are composed of the high-pass subbands.

The above scheme cannot always achieve the optimum decomposition structure with respect to the coding efficiency, because it is based on the assumption that the temporal correlation is always stronger than the view one at the beginning several levels of decomposition. However, for some special multiview videos, the view correlation is probably stronger than the temporal one, and accordingly, the view decomposition should be performed first. Actually, in most cases, the temporal decomposition and the view decomposition should be performed in the manner of interleaving according to their coding costs. For instance, the temporal distance between two low-pass frames is increased by a factor of two after one-level temporal decomposition, leading to the fact that the average coding cost of the following temporal decomposition is larger than the current one. Meanwhile, the average coding cost of the view decomposition remains almost the same, because the temporal and view decompositions are approximately independent. Thus, it is possible that the coding cost of temporal decomposition becomes larger than that of the view decomposition. In this case, the DCVF should be selected in the following decomposition. After the view decomposition is performed, it is possible that the

Figure 5.3: Different decomposition structures along the temporal and the view directions, (a) $4T+2V$ and (b) $3T+V+T+V$.

MCTF becomes more efficient than the DCVF once again. In this way, the interleaved decomposition structure, as shown in Figure 5.3(b), should be more efficient.

5.2.2 Adaptive Decomposition Structure

To decide the optimal decomposition structure, we define a cost function:

$$H\text{Cost} = SAD(H) + \lambda R(MV) \quad (5.2.1)$$

to evaluate the efficiency of de-correlation after one-level decomposition. $SAD(H)$ is the distortion measured by the sum of absolute differences of samples in the high-pass subband H obtained from the decomposition. $R(MV)$ is the bit rate for coding the motion or disparity information. λ is the Lagrangian parameter, which is the same as that used in rate-distortion optimized motion/disparity estimation. Here, the final cost $H\text{Cost}$ is the average value calculated on each MB.

Based on the analysis of the coding cost, the decomposition structure can be adaptive

Table 5.1: $HCost$ for interleaved decomposition structure.

Sequence	Flamenco1		Golf2		Race1	
Direction	Temporal	View	Temporal	View	Temporal	View
1 st	305.24	732.24	246.91	902.88	531.31	1201.53
2 nd	416.60	722.66	233.70	900.05	759.03	1201.88
3 rd	564.94	715.29	308.81	900.81	1089.23	1201.88
4 th	735.85	714.93	372.53	898.01	1486.69	1202.48
5 th	745.82	970.40	689.46	903.90	1508.57	1595.53
6 th	926.55	975.16	1119.03	914.72	1981	1577.02
Selected Structure	3T+V+2T		5T+V		3T+V+T+V	

to the video content to achieve higher coding. Table 5.1 shows some statistics from the MPEG testing multiview video sequences. The cost values of the temporal and view decompositions at each level are respectively calculated using Eqn. (5.2.1). These results are obtained from a GGOP, which contains the GOPs in all view at the same time interval. The decomposition with the smaller cost value is selected for the current level. Then, the next level decomposition can be performed on the current low-pass subband frames. The statistics in Table 5.1 show that the cost values of the view and temporal direction can be regarded as independent. Thus, it is reasonable that the adaptive decomposition structure can achieve better coding performance.

5.3 Subband Multiview Video Coding

In this section, we first figure out the proposed MVC architecture which is based on the proposed DCVF and the adaptive decomposition structure, and then present the MB mode design and the subband coding scheme.

5.3.1 System Structure

The proposed MVC scheme is built up on the SVC scheme proposed in [67], which has been adopted as the common reference software for the MPEG exploration activity on inter-frame wavelet coding. The multiview video can be taken as a matrix, with each element indicating a frame. The proposed DCVF and the original MCTF can be jointly utilized to decompose the raw matrix into a coefficient matrix with each element indicating a low-pass or high-pass subband, which can be further spatially decomposed. Thus, the multiview video is finally transformed into the 4D wavelet coefficients.

The block diagram of the proposed scheme is shown in Figure 5.4. The lifting-based decomposition structure is performed based on the GOP. In this scheme, the input data is analyzed in both temporal and view directions, and the decomposition structure is decided by the cost function defined in Eqn. (5.2.1). After either temporal or view decomposition, the high-pass subbands are sent to the buffer for spatial wavelet decomposition and entropy coding. If the decomposition is not the last level, the low-pass subbands are further decomposed along either temporal or view direction; otherwise, they are also sent to the buffer for spatial wavelet decomposition and entropy coding. The motion vectors obtained from the MCTF and the disparity information obtained from the DCVF are also encoded. The 3D-ESCOT algorithm is adopted for entropy coding and a rate-distortion weight is assigned to each temporal-view-spatial subband for the purpose of overall bit allocation.

5.3.2 MB Modes

The MB modes in the conventional MCTF include FwD, BwD, DirecT, Skip and BiD [67]. In the FwD or BwD mode, the forward or backward motion compensated prediction is used in the lifting steps, and the forward or backward motion vectors are encoded. In the

Figure 5.4: The block diagram of the proposed high dimensional scalable multiview video coding scheme.

BiD mode, bidirectional motion-compensated prediction is used, and both the forward and backward motion vectors are encoded. Since motion vectors among the neighboring MBs have strong correlations, the Skip mode is employed where no motion vectors need to be encoded. In the Skip mode, bidirectional motion-compensated prediction is used, and the forward and backward motion vectors are derived from the coded surrounding MBs. In the DirecT mode, the forward and backward motion vectors have the same magnitude but opposite sign, and hence only one motion vector needs to be encoded.

Due to the existence of the GDC MBs in the proposed DCVF, it is not efficient to directly employ the MB modes designed for the MCTF. Therefore, we introduce some new MB modes for the GDC MBs and exclude some others for the LDC MBs. Particularly, we introduce three new MBs modes for the GDC MBs, i.e., FwGDC, BwGDC and BiGDC, where the forward, backward or bidirectional global disparity-compensated prediction is used, respectively. Since the disparity information can be derived from the global disparity parameters encoded at the frame header, no local disparity vectors need to be encoded in any of the three modes. Moreover, we exclude the DirecT mode while preserving the other four modes for the LDC MBs. The reason lies in that the disparity in the consecutive view frames is not as continuous as the motion of the temporal frames.

Figure 5.5: Six macroblock modes for motion (disparity) coding.

As a summary, there are totally seven MB modes designed for the GDC and LDC MBs in the proposed DCVF, as shown in Figure 5.5.

The rate-distortion optimization (RDO) technique in [67] is also employed to select the best mode of every MB. As for the coding of the MB mode, the seven MB modes are considered together in the variable length code design. To follow the syntax in the original codec, we employ the Golumb-Rice codes by ordering the seven modes based on the probabilities in statistics. We merge the two modes FwGDC and BwGDC into one mode SiGDC since they have very similar probabilities in statistics. Thus, if SiGDC is signaled, an additional flag has to be signaled too to indicate the alternative of the two modes. Moreover, when the right reference is unavailable in DCVF, only FwD and FwGDC modes are used. For these frames, one bit is enough to signal the MB mode of every MB. In the conventional MCTF, the influence of the unavailability of the right reference is negligible since only a few frames do not have the right references. However, for the multiview video, the number of frames that do not have the right reference may occupy a large portion. For example, for an eight-view video, one third of prediction

operations do not have right reference in the DCVF.

5.3.3 Subband Coding

In the proposed MVC scheme, the multiview video data is decomposed into 4D wavelet coefficients, which should be efficiently organize for entropy coding. As shown in Figure 5.3, every temporal-view subband after the temporal and view decompositions is still two dimensional. For simplicity, we scan the coefficient frames in every subband into a sequence, i.e., we re-organize the 4D wavelet coefficients into the 3D wavelet coefficients, so that 3D-ESCOT can be directly utilized. Since several frames are jointly coded and the context relies on the adjacent frames in 3D-ESCOT, the scan order may have a close impact on the coding efficiency. Therefore, we study the various scan orders and select the scan order in terms of the actual decomposition structure.

Particularly, we consider three scan orders, i.e. the T-scan, the V-scan and the head-to-tail scan, as shown in Figure 5.6. For the subband generated by MCTF, the correlation between temporal adjacent frames is stronger than that between view adjacent frames. Similarly, for the subband generated by DCVF, the correlation along view direction is stronger than that along temporal direction. However, in either T-scan or V-scan, there almost exists no correlation at the transition point from one sequence to another. The entropy coding would be inefficient if the context is across such transition points. Comparing with T-scan and V-scan, we observe a marginal performance gain for head-to-tail scan. In the head-to-tail scan, the subband generated by MCTF is scanned in temporal order for even rows and inverse temporal order for odd rows, and the subbands generated by DCVF is scanned in view direction for even columns and inverse view direction for odd columns. In this way, the entropy coding in 3D-ESCOT can be adopted in the proposed scheme.

A rate-distortion (R-D) optimized truncation technique is used to assemble different

Figure 5.6: Scan orders of a temporal-view subband, (a) T-scan, (b) V-scan, (c) head-to-tail scan for MCTF-decomposed subband, and (d) head-to-tail scan for DCVF-decomposed subband

subband coded bits into different layers so as to improve the overall coding performance. The basic idea is to truncate the compressed bits of the different subbands at the same slope of the R-D curves by a weight matrix. The definition of the weight matrix depends on the decomposition structure. We take the decomposition structure $4T+2V$ as shown in Figure 5.3(a) as an example. Suppose the $5/3$ filter, i.e. $(\frac{1}{2}, 1, \frac{1}{2})$ for the low-pass synthesis and $(-\frac{1}{8}, -\frac{1}{4}, \frac{3}{4}, -\frac{1}{4}, -\frac{1}{8})$ for the high-pass synthesis, is used for both MCTF and DCVF. Based on these filter coefficients, the transform gain of each subband in terms of one- or multi-level DWT can be calculated. For example, $WL = 1.5$ and $WH = 0.7188$ for the low- and high-pass subbands in the one-level DWT, and $WLLLL = 10.6875$ and $WLLH = 3.0430$ for the low-pass and high-pass subbands in four-level DWT. To extend this scheme to the 4D wavelet decomposition structure, the weight of the subband from both MCTF and DCVF is defined as the product of the two weights from the two DWTs. For instance, the weight of subband (T-LLH, V-H) in Figure 5.3(a) is defined as $WLLH \times WH = 7.6822$. The weight matrix for the whole $4T+2V$ decomposition structure is shown in Table 5.2.

Table 5.2: Weight matrix for global rate-distortion optimization truncation.

	VBand=0	1	2
TBand=0	0.7188	-	-
1	0.9219	-	-
2	1.5859	-	-
3	3.0430	-	-
4	7.6822	9.8525	29.3906

5.4 Experimental Results and Analysis

5.4.1 Evaluation of Overall Coding Performance

To evaluate the proposed MVC scheme, we present the experimental results on three MPEG 3DAV test multiview video sequences [8], as described in Table 5.3. Each test multiview video sequences include 8 views with 256 frames per view. Some sample frames of the three multiview videos are shown in Figure 5.7.

In the evaluation of overall coding performance, the adaptive decomposition structure is used in the proposed MVC scheme. The temporal GOP size is set to be 128, and hence the maximum number of levels for the temporal decomposition can be up to 7. The view GOP size is set to be 8, and hence the maximum number of levels for the view decomposition can be up to 3. Particularly, we set the total number of levels for the temporal and view decomposition to be 6. The proposed MVC scheme is compared with the simulcast video coding, in which every view is independently coded using H.264 AVC/MPEG-4 part10 as a reference, and the test model JM7.6 is selected as the reference software. Table 5.4 shows the coding parameters used to generate the H.264 anchors [8]. For fair comparison, the H.264 anchors are also encoded with an I-frame period of 128 frames. In addition, since the proposed MVC scheme is developed from the wavelet-based video coding in [67], the simulcast video coding using the SVC codec is also taken as a reference in comparison. We set the same test conditions for the proposed MVC scheme

Table 5.3: Parameters for test multiview video sequences.

Sequence	Class	Image Property	Camera Parameter	Bitrate(kbps)
Flamenco1	A(easy)	$320 \times 240, 30$ fps	N:8,D:20cm,A:1D parallel	64,128,256
Golf2	A(easy)	$320 \times 240, 30$ fps	N:8,D:20cm,A:1D parallel	64,128,256
Race1	B(difficult)	$320 \times 240, 30$ fps	N:8,D:20cm,A:1D parallel	128,256,512

Table 5.4: AVC Parameters for monoview video coding.

Feature/Tool/Setting	AVC Parameters
Rate control	Yes
R-D optimization	Yes
Specific settings	Loop filter, CABAC
Search range	± 32 for CIF/VGA
#Reference picture	5
I-frame period	1 I-frame per GOP
GOP structure	IBBP...

and the original SVC codec.

Figure 5.8 shows the rate-distortion curves of the coding results on the three test sequences. We can observe that the proposed algorithm always outperforms SVC simulcast coding. The coding gain is around 1 dB in average and can be up to 2 dB at the low bit rate coding. The coding gain slightly decreases at high bit rates and the reason is due to

Figure 5.7: The first frame of the left-most view of the original multiview video sequences (a) Flamenco1, (b) Golf2 and (c) Race1.

Figure 5.8: Comparison of coding results using the proposed scheme, H.264/AVC simulcast, and SVC simulcast for (a) Flamenco1, (b) Golf2 and (c) Race1.

the performance of DCVF. Actually, the coding gain of the proposed MVC scheme over the SVC simulcast coding scheme mainly comes from the DCVF for the view decomposition and the adaptive decomposition structure. As will be discussed later, the global disparity model in the DCVF works relatively better at lower bit rates than at higher bit rates. As a consequence, the contribution from the DCVF decreases at higher bit rates, thus the overall coding gain of the proposed MVC scheme over the SVC simulcast coding slightly decreases sometimes.

The proposed MVC scheme can also achieve the coding gain of 1.3 dB in average and up to 2 dB comparing to the H.264 simulcast coding. However, for the coding of Flamenco1, the H.264 simulcast coding slightly outperforms the proposed MVC scheme at high bit rates. Actually, the H.264 simulcast coding also outperforms the SVC simulcast coding about 1 dB for this multiview video. We can observe that there exists some illumination changing across frames in Flamenco1. In this case, the directional spatial prediction adopted in H.264 may be more efficient than the motion-compensated prediction as well as the MCTF, which leads to the fact that the SVC simulcast anchors become much worse than the H.264 anchors. As a consequence, the overall coding gain of the proposed MVC scheme over the H.264 anchors decreases too. Actually, the intra prediction has been an open problem to be addressed in the wavelet-based video coding.

The memory cost and computing complexity are discussed as follows. Since the simulcast coding scheme can encode and decode the multiple views one by one, its memory cost is always less than the proposed MVC scheme. Here we only discuss the case that the multiple views are simultaneously processed. Suppose that the number of decomposition levels of the proposed MVC scheme is the same as that of the SVC simulcast coding, the memory cost should be very similar. Comparing with the H.264 video coding, the memory cost of the proposed scheme is higher due to the hierarchical wavelet decomposition across frames. Actually, it also remains one major problem of the existing wavelet-based

scalable video coding schemes. As for the computing complexity, the extra complexity of the proposed MVC scheme compared to the SVC simulcast coding mainly comes from the encoder part, i.e. the decision of the adaptive decomposition structure, which can be replaced by an estimation instead of the calculation of the coding cost. The decoding complexity of the proposed MVC scheme is very similar to the SVC simulcast coding. Compared with the H.264 coding with 5 reference frames, our testing has shown that the two schemes have the similar running time.

5.4.2 Evaluation of Decomposition Structure

Since the proposed MVC scheme supports the flexible decomposition structure, the coding performances with respect to the different decomposition structures are also evaluated, including the T-first, the V-first and the interleaved structures. The T-first and the V-first can be taken as the special cases of the interleaved structure. Particularly, we employ the 4T+2V and T-only (i.e. 6T) structures to evaluate the T-first case, and the 2V+4T structure to evaluate the V-first case. In the evaluation, the interleaved decomposition structure is obtained based on the cost function Eqn. (5.2.1). Table 5.1 shows the cost value w.r.t. each level of decomposition, which is calculated from one GOP of every test multiview video. The experiments in this subsection are also performed on the same GGOPs of these multiview videos.

Table 5.5 illustrates the distortion and the final cost value w.r.t. every decomposition structure at the same bit rate. We can observe that the cost value approximately represents the de-correlation efficiency as well as the final R-D performance. As shown in Table 5.5, the interleaved decomposition structure has the best R-D performance. Actually, when the decomposition structures are very close, the corresponding cost values as well as the overall R-D performances are also very close, e.g. the 4T+2V structure and the 3T+V+T+V structure for Race1. In our experiments, the interleaved structure is

Table 5.5: Comparison of coding performance of the three different decomposition structures.

Sequence	Decomposition Structure	Bitrate	PSNR	HCost(Total)
Flamenco1 128kbps	T-first (4T+2V)	127.92	38.03	483.90
	V-first (2V+4T)	128.00	34.89	965.39
	Interleaved (3T+V+2T)	127.96	38.14	465.62
Golf2 128kbps	T-first (4T+2V)	127.94	39.56	332.79
	V-first (2V+4T)	127.89	35.20	806.62
	Interleaved (5T+V)	127.99	39.76	324.30
Race1 256kbps	T-first (4T+2V)	255.85	34.17	572.21
	V-first (2V+4T)	255.98	32.46	1165.93
	Interleaved (3T+V+T+V)	255.96	34.28	562.98

decided by the cost function using the Lagrange multiplier. The chosen decomposition structure at a target bit rate may not be the most efficient one at the other bit rates. However, the decomposition structures for one multiview video decided at the different bit rates may only be slightly different, which have very limited effects on the overall coding efficiency. Thus we use the same structure for all bit rates in the evaluation of the overall coding performance.

5.4.3 Evaluation of DCVF

The efficiency of the proposed DCVF for the view decomposition is evaluated, in which the MCTF is taken as a reference. Moreover, the DCVF under the different settings are also tested, including the DCVF without LDC (i.e., GDC only) and the DCVF without updating. The experiments are performed on the view sequences of the multiview videos. Figure 5.9 shows the R-D performances w.r.t. the four coding schemes. We can observe that the proposed DCVF always outperforms the MCTF for the coding of view sequences. At the low bit rate, the coding gain can be up to 2 dB. It proves that the global disparity model is very efficient in for the view sequence coding, especially at the low bit rates. Moreover, the coding result of the DCVF becomes worse when either the LDC or the

Figure 5.9: Performance evaluation of DCVF (a) Flamenco1, (b) Golf2 and (c) Race1.

updating is turned off. Particularly, the gap between the DCVF with and without LDC increases while the bit rate increases, which indicates that the importance of the LDC in DCVF enlarges while the bit rate increases. The subjective image qualities can be compared from Figure

To compare the results subjectively, we also include the decoded images for the view sequence of Golf2 at 128 in Figure 5.10 comparing DCVF and MCTF.

Figure 5.10: Subjective image quality evaluation, decoded image by (a) DCVF and (b) MCTF for Golf2 view sequence at 128kbps.

Figure 5.11: Multi-dimensional decomposition of multiview video data.

5.4.4 Discussion of Scalabilities

In this part we discuss how the multi-dimensional representation of the multiview video data can be used to support various kinds of scalabilities.

Figure 5.11 illustrates a simplified (i.e., 1-level MCTF followed by 1-level DCVF) 4D decomposition of the multiview video data. As shown in the figure, *temporal scalability* can be obtained by discarding the temporal high pass subbands. By decoding the temporal low pass subbands, a sub-sequence at half the input frame rate can be obtained for each view. Similarly, *view scalability* can be obtained by discarding the view high pass subbands. For example, view 0 and view 2 will be decoded and view 1 will be discarded. *Spatial scalability* can be obtained by discarding the spatial high pass subband images while *SNR scalability* is supported by bitplane coding. Besides, *combined scalability* can also be supported. For example, if the input multiview video data contains 8 views and each view is SD at 60Hz, it is possible to extract either 4 views with QCIF at 15Hz or 2 views with CIF at 30 Hz. However, the level of scalability is limited by the number of decomposition along the corresponding direction. For example, extracting 2 views with CIF at 30 Hz from the input with 8 views, SD and 60 HZ is not supported for the 5T+V decomposition structure. Moreover, the level of scalability is limited by the order of decompositions. A 4T+2V decomposition structure can only provide 2 views or 4 views at 15 Hz but not at 30 Hz or 60 Hz for the same input. This is due to the required match between the encoder and the decoder.

In conclusion, we propose a flexible multiview video coding scheme using high dimensional wavelet in this Chapter. The experimental results have demonstrated that the proposed scheme not only provides a variety of scalabilities but also improves the coding efficiency and the visual quality of the decoded images. The results reported in this work are only for 8-view video sequences, which means that the view-sequences only

contains eight frames each. Since lifting is not efficient on such a short sequence because of the boundary effect, it constraints the overall performance gain although the embedded global disparity improves the coding efficiency of t-LLLL band significantly especially at low bit rates.

For the various decomposition structures supported by the proposed system, i.e., T-first, V-first and interleaved, the interleaved structure obtained by comparing the decorrelation efficiency achieves the best coding performance at the cost of increased complexity, which can be solved by using a fast algorithm to approximately evaluate the correlations among the images.

Chapter 6

Conclusions

6.1 Summary

In this thesis, we have studied the fundamentals of stereoscopic imaging and multiview video processing, reviewed the existing techniques for digital video coding, and investigated the state-of-the-art methods in stereoscopic and multiview video compression in the literature. We have proposed an edge-preserving disparity regularization algorithm, and subsequently a joint estimation algorithm which is very suitable for coherent disparity and motion estimation for interframe prediction in multiview video coding. We have built up two complete MVC systems. One is MPEG-4 compatible and the other is more flexible and provides scalabilities by using high dimensional wavelet coding.

For the disparity regularization algorithm, we analyzed the properties of the disparity vectors for different video objects at different depth, the impact of edge values on the smoothness of the disparity fields and the rate-distortion property of disparity compensation. An edge-preserving energy function is then proposed based on the MRF theory. The experimental results have proven the accuracy and the robustness of our proposed regularization scheme.

The relationship between motion and disparity fields can be utilized for more effective

estimation of both displacement fields. The stereo-motion consistency constraint, or the “loop” constraint, has been investigated. For coherent estimation of both fields, a global energy function is proposed based on the MAP probability and a sub-optimal solution is obtained by iterative local optimizations. The loop constraint is used to calculate the initial disparity field and remove the outliers during the regularization process to make the algorithm more robust. Compared with the separate regularization of disparity and motion fields, the joint estimation reduces the computational complexity significantly with slight performance degradation in term of PSNR.

We have further proposed an MPEG-4 compatible stereoscopic and multiview video coding system to provide good image quality for both main and auxiliary views at low bit rates. The main view is coded as an MPEG-4 bitstream and the auxiliary view is predicted jointly by disparity and motion. The joint estimation algorithm is embedded to calculate smoother displacement fields for interframe prediction and compensation, which significantly reduces the bit rate for coding the disparity and the motion vectors. We extend the rate control scheme in monoview video coding for the coding of multiple views. For the 5-view video coding, we compared various view-level reference structures and discussed the optimal reference structures regarding to the multiview video properties such as the baseline distance.

We have also proposed a more flexible multiview video codec based on high dimensional wavelet. By adaptively combining the MCTF and the DCVF, both temporal redundancy and view redundancy are fully utilized and efficient compression is achieved. Since the wavelet decomposition perfectly matches the structure of high dimensional data of multiview video, the MVC based on wavelet is not constrained by the multiview configurations and can be easily extended for more complex applications. In addition, taking the advantages of the DWT and bitplane coding, the proposed scheme can provide a variety of scalabilities including temporal scalability, spatial scalability, SNR scalability

as well as view scalability.

6.2 Future Directions

Our studies on joint disparity and motion estimation show that the performance of the block-based joint estimation algorithm is much worse than the pixel-based counterpart. This is because the block-based algorithm blindly divides the image into fixed size square blocks without taking into account the semantic content of the image. Blocks at object boundaries may contain pixels that undergo entirely different true disparity vectors. One solution is to use overlapped block matching and compensation, and another is to use object-based methods, which can achieve more accurate disparity fields and more efficient compression since the disparity within one object varies slowly and smoothly.

Our studies on MPEG-compatible multiview video coding show that there exist inevitable limitations on the MVC schemes in order to remain the compatibility with existing monoview video coding standards. In particular, the MPEG-compatible multiview video coding is normally constrained by the frame prediction structure and the camera configurations, and the performance gain is limited. The extension of our MPEG-4 based MVC scheme to a more general form deserves careful considerations. The comparison on various reference structures for multiview video coding indicates that the bit allocation between different views plays a critical role in MVC. For more accurate bit rate control and better coding efficiency, the dependent rate control scheme has to be investigated for multiview video coding instead of extending the existing rate control for monoview video. In addition, although we compare different reference structures, it is desired to give an analytical method to determine the optimal prediction structure for a given scene in a rate-distortion sense.

Still there is much work to be done to optimize the performance of our proposed

wavelet based MVC scheme. For example, the proposed MVC decomposes the multiview video data in the temporal and view direction separately. It is advantageous to explore the temporal and view directional redundancy coherently on the 4D wavelet domain to achieve better compression efficiency. Another possible improvement of the wavelet MVC scheme is again related to rate control. To achieve optimal overall performance, in scalable video coding using MCTF, the image quality of the low pass subband images are always better than that of the high pass subband images. This phenomena also exists for DCVF, which results in noteworthy differences of decoded image qualities among different views. In order to have a good tradeoff between the overall performance and the variance of the PSNR, the R-D weight matrix has to be carefully designed.

Besides the extensions for those particular systems or algorithms we developed, there are also many problems remaining for a general multiview video coding system. For example, it is clear that the correlations among views are increased with the increase of the number of cameras. An interesting question is: if the number of cameras goes to infinity, what will be the overall coding rate? Theoretical studies are needed to find out this asymptotic value. In short, multiview video coding has its unique features. To explore the features of multiview video, it requires the integration from both the tradition video coding area and the area of computer vision.

Bibliography

- [1] T. Sheridan, *Telerobotics, Automation and Human Supervisory Control*. The MIT Press, London, 1992.
- [2] A. Puri, R. V. Kollarits, and B. G. Haskell, “Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4,” *Signal Processing: Image Communication*, vol. 10, pp. 201–234, 1997.
- [3] ISO/IEC JTC1/SC29/WG11 N6501, “Requirements on multi-view video coding,” Redmond, USA, July 2004.
- [4] Ru-Shang Wang, “Multiview/stereoscopic video analysis, compression, and virtual viewpoint synthesis,” Ph.D. dissertation, Polytechnic, Newyork, June 1999.
- [5] ISO/IEC IS 11172 (MPEG-1), “Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbits/s,” *Technical Report, Motion Picture Experts Group*, 1993.
- [6] ISO/IEC IS 13818 (MPEG-2), “Generic coding of moving pictures and associated audio information,” *Technical Report, Motion Picture Experts Group*, 1994.
- [7] ISO/IEC 14496-2:1999, “Information technology — coding of audio/visual objects,” 1999.

- [8] ISO/IEC JTC1/SC29/WG11 MPEG2004 N6494, “Preliminary call for evidence on multiview video coding,” Redmond, USA, July 2004.
- [9] Sriram Sethuraman, “Stereoscopic image sequence compression using multiresolution and quadtree decomposition based disparity- and motion-adaptive segmentation,” Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, Pennsylvania, July 1996.
- [10] Gang Xu and Zhengyou Zhang, *Epipolar Geometry in Stereo, Motion and Object Recognition A Unified Approach*. Kluwer Academic Publishers.
- [11] Jong-Il Park and Seiki Inoue, “Acquisition of sharp depth map from multiple cameras,” *Signal Processing: Image Communication*, vol. 14, pp. 7–19, 1998.
- [12] Yuichi MAKAMURA, Tomohiko MATSUURA, Kiyohide SATOH, and Yuichi OHTA, “Occlusion detectable stereo - occlusion patterns in camera matrix,” in *IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, June 1996, pp. 371–378.
- [13] Jens-Rainer Ohm and Ebroul Izquierdo M., “An object-based system for stereoscopic viewpoint synthesis,” *IEEE Transactions on Circuits and Systems for Video Technology*, no. 5, pp. 801–811, October 1997.
- [14] Nikos Grammalidis and Michael G. Strintzis, “Disparity and occlusion estimation in multiocular systems and their coding for the communication of multiview image sequences,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 3, pp. 328–344, June 1998.
- [15] Graham J. Woodgate, David Ezra, Jonathan Harrold, Nicolas S. Holliman, Graham R. Jones, and Richard R. Moseley, “Autostereoscopic 3D display systems with

- observer tracking,” *Signal Processing: Image Communication*, vol. 14, pp. 131–145, 1998.
- [16] Yoshihiro Kajiki, Hiroshi Yoshikawa, and Toshio Honda, “Autostereoscopic 3-D video display using multiple light beams with scanning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 254–260, March 2000.
- [17] Michael Brunig and Wolfgang Niehsen, “Fast full search block matching based on combined SAD and MSE measures,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 3653, January 1999, pp. 439–449.
- [18] H. Fan and King N. Ngan, “Disparity map coding based on adaptive triangular surface modeling,” *Signal Processing: Image Communications*, vol. 14, no. 1-2, pp. 119–130, November 1998.
- [19] Chia-Wen Lin, Er-Yin Fei, and Yung-Chang Chen, “Hierarchical disparity estimation using spatial correlation,” *IEEE Transactions on Consumer Electronics*, vol. 44, no. 3, pp. 630–637, August 1998.
- [20] Yongdong Zhang and Guiling Li, “An efficient hierarchical disparity estimation algorithm for stereoscopic video coding,” in *IEEE Asia-Pacific Conference on Circuits and Systems, (APCCAS)*, 2000, pp. 744–747.
- [21] Marco Accame, Francesco G. B. De Natale, and Daniele D. Giusto, “Hierarchical block matching for disparity estimation in stereo sequences,” in *IEEE International Conference on Image Processing (ICIP)*, vol. 2, 1995, pp. 374–377.
- [22] Wootack Woo and Antonio Ortega, “Modified overlapped block disparity compensation for stereo image coding,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 3653, January 1999, pp. 570–581.
-

- [23] Takeo Kanade and Masatoshi Okutomi, “A stereo matching algorithm with an adaptive window: theory and experiment,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 9, pp. 920–932, September 1994.
- [24] Woontack Woo and Antonio Ortega, “Overlapped block disparity compensation with adaptive windows for stereo image coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 194–200, March 2000.
- [25] Weiguo Wu, Atsushi Yokoyama, Teruyuki Ushiro, Takayuki Yoshigahara, and Yoko Miwa, “A stereo matching method with deformable window and its application to 3D measurement of human face,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 4067, 2000, pp. 378–388.
- [26] Luis Torres and Murat Kunt, *Video Coding, The Second Generation Approach*. Kluwer Academic Publishers.
- [27] Woontack Woo and Antonio Ortega, “Stereo image compression based on disparity field segmentation,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, 1997, pp. 391–402.
- [28] Sekhavat Sharghi and Farhad Kamangar, “Stereo correspondence using geometric relational matching,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 3653, 1999, pp. 582–592.
- [29] Mustapha Kardouchi, Janusz Konrad, and Carlos Vazquez, “Estimation of large-amplitude motion and disparity fields: application to intermediate view reconstruction,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 4310, 2001, pp. 340–351.

- [30] Yang Wang and Prabir Bhattacharya, “Hierarchical stereo correspondence using features of gray connected components,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 3024, 1997, pp. 403–412.
- [31] Yanwen Ji, Anthony Tung-Shuen Ho, and Tao Yu, “Using a novel multi-resolution hybrid matching method to improve stereo matching accuracy of satellite images,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 4067, 2000, pp. 1024–1032.
- [32] Cun Xie and Jia-Sheng Hu, “Stereo matching method using modified genetic algorithm,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 4115, 2000, pp. 646–653.
- [33] Yuecheng Zhang and Yun He, “A new real-time algorithm for multiview disparity based 3D-object segmentation,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 4310, 2001, pp. 702–712.
- [34] J. Karathanasis, D. Kalivas, and J. Vlontzos, “Disparity estimation using block matching and dynamic programming,” in *IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, 1996, pp. 728–731.
- [35] Andre Redert, Chun-Jen Tsai, Emile Hendriks, and Aggelos K. Katsaggelos, “Disparity estimation with modeling of occlusion and object orientation,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 3309, 1999, pp. 798–808.
- [36] Til Aach and Andre Kaup, “Disparity-based segmentation of stereoscopic foreground / background image sequences,” *IEEE Transactions on Communications*, vol. 42, no. 2/3/4, pp. 673–679, 1994.

- [37] Wen-Nung Lie and Bo-Er Wei, “Intermediate view synthesis from binocular images for stereoscopic applications,” in *IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 5, 2001, pp. 287–290.
- [38] C. L. Pagliari, M. M. Perez, and T. J. Dennis, “Reconstruction of intermediate views from stereoscopic images using a rational filter,” in *IEEE International Conference on Image Processing (ICIP)*, vol. 2, 1998, pp. 627–631.
- [39] L. J. Luo, D. R. Clewer, D. R. Bull, and C. N. Canagarajah, “A hierarchical genetic disparity estimation algorithm for multiview image systems,” in *IEEE International Conference on Image Processing (ICIP)*, vol. 2, 2000, pp. 768–771.
- [40] Guoping Li and Yun He, “A hierarchical combined feature- and area-based stereo matching algorithm,” in *IEEE International Symposium on Circuits and Systems (ISCAS)*, vol. 2, May 2002, pp. 227–280.
- [41] Sang Hwa Lee, Yasuaki Kanatsugu, and Jong-Il Park, “A hierarchical method of MAP-based stochastic diffusion and disparity estimation,” in *IEEE International Conference on Image Processing (ICIP)*, 2002, pp. 541–544.
- [42] —, “Hierarchical stochastic diffusion for disparity estimation,” in *IEEE Workshop on Stereo and Multi-baseline Vision*, 2001, pp. 111–120.
- [43] Jun Zhang and Gerald G. Hanauer, “The application of mean field theory to image motion estimation,” *IEEE Transactions on Image Processing*, vol. 4, no. 1, pp. 19–33, January 1995.
- [44] Ikhlas M. Abdclqader, Sarah A. Rajala, Wesley E. Snyder, and Griff L. Bilbro, “Energy minimization approach to motion estimation,” *IEEE Signal Processing Magazine*, vol. 28, pp. 291–309, 1992.

- [45] Mi-Hyun Kim and Kwang-Hoon Sohn, “Edge-preserving directional regularization technique for disparity estimation of stereoscopic images,” *IEEE Transactions on Consumer Electronics*, vol. 45, no. 3, pp. 804–811, August 1999.
- [46] Yair Shoham and Allen Gersho, “Efficient bit allocation for an arbitrary set of quantizers,” *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. 36, no. 9, pp. 1445–1453, 1988.
- [47] Jens-Rainer Ohm, Karsten Gruneberg, Emile Hendriks, M. Ebroul Izquierdo, Dimitris Kalivas, Michael Karl, Dionysis Papadimitos, and Andre Redert, “A realtime hardware system for stereoscopic videoconferencing with viewpoint adaptation,” *Signal Processing: Image Communication*, vol. 14, pp. 147–171, 1998.
- [48] Soo-Chang Pei and Chin-Lun Lai, “Very low bit-rate coding algorithm for stereo video with spatiotemporal HVS model and binary correlation disparity estimator,” *IEEE Journal on Selected Areas in Communications*, vol. 16, no. 1, pp. 98–107, January 1998.
- [49] I. Patras, “Construction of multiple views using jointly estimated motion and disparity fields,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 3024, 1997, pp. 380–390.
- [50] Mel Siegel, S. Sethuraman, J. S. McVeigh, and Angel Jordan, “Compression and interpolation of 3D-stereoscopic and multi-view video,” in *Stereoscopic Displays and Virtual Reality Systems IV*, vol. 3012, February 1997, pp. 227–238.
- [51] Dimitrios Tzovaras and Michael G. Strintzis, “Motion and disparity estimation using rate distortion theory for very low bit rate and multiview image sequence coding,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 3024, 1997, pp. 352–359.

- [52] A. Tamatoui and C. Labit, “Coherent disparity and motion compensation in 3DTV image sequence coding schemes,” in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, vol. 4, April 1991, pp. 2845–2848.
- [53] Andre R, Emile Hendriks, and Jan Biemond, “Correspondence estimation in image pairs,” *IEEE Signal Processing Magazine*, vol. 16, no. 3, pp. 29–46, May 1999.
- [54] Y. -L. Chan and W. -C. Siu, “Edge oriented block motion estimation for video coding,” in *IEE Proceedings on Vision, Image and Signal Processing*, vol. 144, no. 3, June 1997, pp. 136–144.
- [55] I. Patras, N. Alvertos, and G. Tziritas, “Joint disparity and motion field estimation in stereoscopic image sequences,” in *13th International Conference on Pattern Recognition*, vol. 1, 1996, pp. 359–363.
- [56] Takeshi Naemura, Masahide Kaneko, and Hiroshi Harashima, “Compression and representation of 3-D images,” *IEICE Transactions on Information and Systems*, vol. E82-D, no. 3, pp. 558–567, March 1999.
- [57] Ebroul Izquierdo M., “Stereo matching for enhanced telepresence in three-dimensional videocommunications,” *IEEE Transactions on Circuits and Systems for Video Technology*, no. 4, pp. 629–643, August 1997.
- [58] Chun-Jen Tsai and Aggelos K. Katsaggelos, “Optical flow estimation for multi-channel video sequences,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 3024, 1997, pp. 360–368.
- [59] L. Falkenhagen, “3D object-based depth estimation from stereoscopic image sequences,” in *Internation Workshop on stereoscopic and three-dimensional imaging*, September 1995.

- [60] A. Murat Tekalp, *Digital Video Processing*. Prentice Hall PTR.
- [61] Wiegand, T., Sullivan, G.J., Bjntegaard, G., and Luthra, A., “Overview of the H.264/AVC video coding standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, no. 7, pp. 560–576, July 2003.
- [62] Editorial, “Special issue on subband/wavelet interframe video coding,” *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 557–559, August 2004.
- [63] Secker and D. Taubman, “Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting,” in *IEEE International Conference on Image Processing (ICIP)*, vol. 2, 2001, pp. 1029–1032.
- [64] L. Luo, J. Li, S. Li, Z. Zhuang, and Y. -Q. Zhang, “Motion compensated lifting wavelet and its application in video coding,” in *International Conference on Multimedia and Expo (ICME)*, 2001, pp. 365–368.
- [65] —, “A two-pass optimal motion threading technique for 3D wavelet video coding,” in *IEEE International Symposium on Circuits and Systems*, vol. 4, 2002, pp. 819–822.
- [66] R. Xiong, F. Wu, S. Li, Z. Xiong, and Y. -Q. Zhang, “Exploiting temporal correlation with block-size adaptive motion alignment for 3D wavelet coding,” in *Visual Communication and Image Processing, Proceedings of SPIE*, vol. 5308, January 2004, pp. 144–155.
- [67] R. Xiong, F. Wu, J. Xu, S. Li, and Y. -Q. Zhang, “Barbell lifting wavelet transform for highly scalable video coding,” in *Picture Coding Symposium*, December 2004.
- [68] J. Xu, Z. Xiong, S. Li, and Y. -Q. Zhang, “3-D embedded subband coding with optimal truncation (3-D ESCOT),” *J. Applied and Computational Harmonic Anal-*

- ysis: Special Issue on Wavelet Applications in Engineering*, vol. 10, pp. 290–315, May 2001.
- [69] Mel Siegel, Priyan Gunatilake, S. Sethuraman, and Angel Jordan, “Compression of stereo image pairs and streams,” in *Stereoscopic Displays and Virtual Reality Systems*, vol. 2177, February 1994, pp. 258 – 268.
- [70] J. -R. Ohm, “Stereo/multiview encoding using the MPEG family of standards,” in *Stereoscopic Displays and Virtual Reality Systems VI*, vol. 3639, January 1999, pp. 242–253.
- [71] Haluk Aydinoglu and Monson H. Hayes, “Multi-view image coding using local orthogonal bases,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 3024, 1997, pp. 340–351.
- [72] J.-R. Ohm, “Stereo/multiview video encoding using the MPEG family of standards,” January 1999.
- [73] Qin Jiang and Monson H. Hayes, “Stereo sequence coding,” in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, 1999, pp. 894–901.
- [74] Yun-Jeong Song, “Improved disparity estimation algorithm with MPEG-2’s scalability for stereo image sequences,” *IEEE Transactions on Consumer Electronics*, vol. 42, no. 3, pp. 306–311, August 1996.
- [75] A. Puri, R. V. Kollarits, and B. G. Haskell, “Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4,” *Signal Processing: Image Communication*, vol. 10, pp. 201–234, 1997.
- [76] J. Lim, K. Ngan, W. Yang, and K. Sohn, “Multiview sequence CODEC with view scalability,” *Signal Processing: Image Communication*, vol. 19, no. 3, pp. 239–256, March 2004.
-

- [77] W. Yang and K. Ngan, “MPEG-4 based stereoscopic video sequences encoder,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, May 2004, pp. 741–744.
- [78] Guoping Li and Yun He, “A novel multiview video coding scheme based on H.264,” in *Fourth International Conference on Information, Communications and Signal Processing, Fourth IEEE Pacific-Rim Conference on Multimedia (ICICS-PCM)*, December 2003, pp. 493–497.
- [79] Sotiris Malassiotis and Michael G. Strintzis, “Object-based coding of stereo image sequences using three-dimensional models,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 6, pp. 892–905, December 1997.
- [80] Ebroul Izquierdo and Jens-Rainer Ohm, “Image-based rendering and 3D modeling: a complete framework,” *Signal Processing: Image Communication*, vol. 15, pp. 817–858, 2000.
- [81] R. S Wang and Y. Wang, “Multiview video sequence analysis, compression, and virtual viewpoint synthesis,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, pp. 397–410, April 2000.
- [82] N. Grammalidis, D. Beletsiotis, and M. G. Strintzis, “Sprite generation and coding in multiview image sequences,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, pp. 302–311, March 2000.
- [83] Nikolaos V. Boulgouris and Michael G. Strintzis, “A family of wavelet-based stereo image coders,” *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 12, no. 10, pp. 898–903, October 2002.
- [84] Chuo-Ling Chang, Xiaoqing Zhu, Prashant Ramanathan, and Bernd Girod, “Inter-view wavelet compression of light fields with disparity-compensated lifting,” in *Vi-*

- sual Communication and Image Processing, Proceedings of SPIE*, vol. 5150, 2003, pp. 694–706.
- [85] Tihao Chiang and Y. -Q. Zhang, “A new rate control scheme using quadratic rate distortion model,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 246–250, February 1997.
- [86] Anthony Vetro, Huifang Sun, and Yao Wang, “MPEG-4 rate control for multiple video objects,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 186–199, February 1999.
- [87] Feng Pan, Z. Li, K. Lim, and G. Feng, “A study of MPEG-4 rate control scheme and its improvements,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 5, pp. 440–446, May 2003.
- [88] Jeongeun Lim and Kwang-Hoon Sohn, “Accurate bit-rate and quality control for 3d multiview sequences,” *Visual Communications and Image Processing (VCIP)*, vol. 5308, pp. 1049–1059, January 2004.
- [89] Sei Naito and Shuichi Matsumoto, “Advanced rate control technologies for 3D-HDTV digital coding based on MPEG-2 multi-view profile,” in *IEEE International Conference on Image Processing (ICIP)*, vol. 1, 1999, pp. 281–285.
- [90] Kannan Ramchandran, Antonio Ortega, and Martin Vetterli, “Bit allocation for dependent quantization with applications to multiresolution and MPEG video coders,” *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 533–545, September 1994.
- [91] Woontack Woo and Antonio Ortega, “Optimal blockwise dependent quantization for stereo image coding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 6, pp. 861–867, September 1999.
-

- [92] Liang-Jin Lin, Antonio Ortega, and C. -C. Jay Kuo, "A gradient-based rate control algorithm with applications to MPEG video," in *IEEE International Conference on Image Processing (ICIP)*, vol. 3, 1995, pp. 392–395.
- [93] J. Po-Yuen Cheng and C. -C. Jay Kuo, "Rate control for an embedded wavelet video coder," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 4, pp. 696–702, August 1997.
- [94] Zhejie Yang, Feng Wu, and Shipeng Li, "Rate distortion optimized mode decision in the scalable video coding," in *IEEE International Conference on Image Processing (ICIP)*, vol. 3, 2003, pp. 781–784.
- [95] Zhihai He and Chang Wen Chen, "Efficient quality analysis for video streaming over wireless ad hoc networks," in *Visual Communications and Image Processing (VCIP), Proceedings of SPIE*, vol. 5308, 2004, pp. 996–1006.
- [96] Belle L. Tseng and Dimitris Anastassiou, "Perceptual adaptive quantization of stereoscopic video coding using MPEG-2's temporal scalability structure," in *International workshop on Stereoscopic and Three Dimensional Imaging*, September 1995.
- [97] Won-Ho Kim, Jae-Young Ahn, and Sung-Woong Ra, "An efficient disparity estimation algorithm for stereoscopic image compression," *IEEE Transactions on Consumer Electronics*, vol. 43, no. 2, pp. 165–172, May 1997.
- [98] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society*, vol. B, no. 36, pp. 192–293, 1974.
- [99] Dimitrios Tzovaras and Michael G. Strintzis, "Motion and disparity field estimation using rate-distortion optimization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 2, pp. 171–180, April 1998.

- [100] Nobuyuki Otsu, “A threshold selection method from gray-level histogram,” *IEEE Transactions on Systems, Man, Cybernetics*, vol. SMC-9, no. 1, pp. 62–66, January 1979.
- [101] Xuemin Chen and Ajay Luthra, “MPEG-2 multi-view profile and its application in 3DTV,” *SPIE/IS-T Multimedia Hardware Architectures*, pp. 212–223, February 1997.
- [102] Hiroyuki Imaizumi and Eisuke Nakasu, “Experimental resultal results of stereoscopic video coding using temporal scalability,” *ISO/IEC JTC1/SC29/WG11 MPEG95/m0449*, Dallas Nov. 1995.
- [103] ISO/IEC JTC1/SC29/WG11 N4030, “MPEG-4 overview - (v.18 - singapore version),” March 2001.
- [104] ISO/IEC JTC1/SC29/WG11 N3908, “MPEG-4 video verification model version 18.0,” January 2000.
- [105] H. S. Kim, S. -S. Kim, J. -Y. Son, and K. H. Sohn, “Feature-based disparity estimation for intermediate view reconstructure of multiview images,” in *International Conference on Imaging Science, Systems and Technology*, vol. 2, June 2001, pp. 1–8.
- [106] B. Pesquet-Popescu and V. Bottreau, “Three-dimensional lifting schemes for motion compensated video compression,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2001, pp. 1793–1796.
- [107] Andrew Secker and David Taubman, “Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression,” *IEEE Transactions on Image Processing*, vol. 12, no. 12, pp. 1530–1542, December 2003.

- [108] F. Dufaux and J. Konrad, “Efficient, robust, and fast global motion estimation for video coding,” *IEEE Transaction on Image Processing*, pp. 497–501, March 2000.
- [109] B. Feng, J. Xu, F. Wu, and S. Yang, “Energy distributed update steps (edu) in lifting based motion compensated video coding,” in *International Conference on Image Processing*, October 2004, pp. 2267–2270.

Journal Papers

1. W. Yang, Y. Lu, F. Wu, J. Cai, K. Ngan and S. Li, “Scalable multiview video coding using 4D wavelet”, submitted to IEEE Transaction on Circuits and Systems for Video Technology.
2. W. Yang, K. Ngan and J. Cai, “An MPEG-4-compatible stereoscopic/multiview video coding scheme”, IEEE Transaction on Circuits and Systems for Video Technology, vol. 16, issue 2, pages 286-290, Feb., 2006.
3. W. Yang, K. Ngan, J. Lim and K. Sohn, “Joint motion and disparity fields estimation for stereoscopic video sequences”, Signal Processing: Image Communication, vol. 20, no. 3, pages 265-276 , March 2005.
4. J. Lim, K. Ngan, W. Yang and K. Sohn, “Multiview sequences CODEC with view scalability”, Signal Processing: Image Communication, vol. 19, no. 3, pages 239-256, March 2004.

Conference Papers

1. W. Yang, F. Wu, Y. Lu, J. Cai and K. Ngan, “Scalable multi-view video coding using wavelet”, IEEE International Symposium on Circuits and Systems (ISCAS’05), pages 6078-6081, 23-26 May.
2. W. Yang, K. Ngan and J. Cai, “Efficient multiview video coding based on MPEG-4”, Pacific-Rim Conference on Multimedia, (PCM’04) vol. 3, Pages 167-174, Nov. 30 - Dec. 3, 2004, Japan.
3. W. Yang, K. Ngan and J. Cai, “MPEG-4 based stereoscopic and multiview video coding”, International Symposium on Intelligent Multimedia, Video and Speech Processing (ISIMP’04), pages 61-64, October 20-22, 2004, Hong Kong.

4. W. Yang and K. Ngan, "MPEG-4 based stereoscopic video sequences encoder", Acoustics, Speech and Signal Processing, 2004 Proceedings (ICASSP'04), IEEE International Conference on., vol. 3, pages 741-744, May 17-21, 2004, Canada.
5. W. Yang, K. Ngan and K. Sohn, "Edge-preserving regularization of Displacement fields and joint optimization of disparity and motion", International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS'03), pages 629-634, 7-10 Dec. 2003, Japan.
6. W. Yang, K. Ngan, J. Lim and K. Sohn, "Edge-preserving regularization of disparity and motion fields", 4th EURASIP Conference on Video/Image Processing and Multimedia Communications (EC-VIP-MC'03), pages 71-76, 2-4 July 2003, Zagreb, Croatia.
7. W. Yang and K. Ngan, "Object-based disparity estimation for stereoscopic images", Seventh International Conference on Control, Automation, Robotics and Vision (ICARCV'02), pages 1105-1109, Dec. 2002, Singapore.

Vita

Wenxian Yang was born in Hunan, China on August 23, 1979. After completing her study at Yali Middle School, Hunan, in 1997, she entered Zhejiang University. She received her B.E. degree in Computer Science and Technology from Zhejiang University in June 2001. Since July 2001, she began her research towards Ph.D. under the guidance of Professor King Ngi Ngan and Assistant Professor Jianfei Cai. She was with Microsoft Research Asia, Beijing, from July to November in 2004.

Her research interests include 3D computer vision, image and video processing, image and video coding, multiview video etc. She has authored over 10 technical journal and conference papers. She is a student member of IEEE.