Reduced-Complexity Disparity Estimation
for Efficient Multiview Imagery Encoding

Dispariteitsschatting met verlaagde complexiteit
voor de efficiënte codering van beeldmateriaal met meerdere gezichtspunten

Aykut Avci

Promotoren: prof. dr. ir. H. De Smet, dr. ir. J. De Cock
Proefschrift ingediend tot het behalen van de graad van
Doctor in de Ingenieurswetenschappen: Computerwetenschappen

Vakgroep Elektronica en Informatiesystemen
Voorzitter: prof. dr. ir. J. Van Campenhout
Faculteit Ingenieurswetenschappen en Architectuur
Academiejaar 2012 - 2013

UNIVERSITEIT
GENT

ISBN 978-90-8578-572-9
NUR 965
Wettelijk depot: D/2013/10.500/5

Ghent University
Faculty of Engineering and Architecture

Promoters:

Prof. dr. ir. Herbert De Smet
Department of Electronics and Information Systems,
Ghent University

Dr. ir. Jan De Cock (*)
Department of Electronics and Information Systems,
Ghent University

Other members of the examination committee:

Prof. dr. ir. Patrick De Baets (Chairman)
Department of Mechanical Construction and Production,
Ghent University

Prof. dr. Peter Lambert (Secretary)
Department of Electronics and Information Systems,
Ghent University

Prof. dr. ir. Ingrid Moerman (*)
Department of Electronics and Information Systems,
Ghent University

Prof. dr. ir. Peter Schelkens (*)
Department of Electronics and Informatics,
Vrije Universiteit Brussel

Dr. Youri Meuret
Brussels Photonics Team,
Vrije Universiteit Brussel

Ing. Christoph Stevens (*)
Alcatel-Lucent

(*) Reading committee.

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor in Engineering: Computer Science Engineering.

Academic year 2012-2013

For my family,

Special thanks

I would like to express my deepest gratitude and appreciation to everyone that has helped me in this stage of my life.

First and foremost, I would especially like to thank my promoter, Herbert De Smet, for his support, excellent guidance, limitless patience and encouragement during the course of this research. I would also like to thank my second promoter, Jan De Cock, for his invaluable inputs and efforts to make me enable to implement the ideas in this dissertation. I am feeling very lucky and blessed that we found Peter Lambert and you.

I had the chance to work together with Yeuri Meuret and Lawrence Bogaert at B-PHOT department in VUB (Vrije Universiteit Brussel). I learned a lot while I was helping programming. Thank you for everything.

I am so much grateful to all of my colleagues at CMST (Centre for Microsystems Technology) group, in particular to Jan D., Katrien, Dieter, Herbert D. P., Lothar, Geert V. S., Peter S., Sandeep, Dominique, Jeroen, Nadine, Sanjeev, Michal, Kamalpreet, Ann M., Diana, Erwin, Tom S., Fabrice, David, Bram, Pietro and Jindrich. I would like to sincerely thank to Amir for his endless patience towards my questions related to LaTeX. I would also like to thank to Marteen C. for his last-minute helps. I would like to single out my precious officemates. Kristof, Hüseyin, Roel, Jelle, Pankaj and Swarnakamal, thank you for all your supports and sincere talks. I will miss a lot our office atmosphere and fries-days.

I would like to express my special thanks to Günay, Doğan and Zelal for their warm and friendly conversations.

Thanks to all those others not being named here who gave me strength and motivation throughout.

I acknowledge the financial support of the Research Foundation-Flanders (FWO).

I thank to my mother, father and brother for their immeasurable love and support.

Finally and most importantly, I thank to my wife who has been the ultimate source of happiness of my life.

Aykut Avci
Gent, october 2012

Table of Contents

List of Figures

List of Tables

List of Acronyms

0-9

2D	Two dimensional
3D	Three dimensional

B

bpp	Bits per pixel

C

CABAC	Context Adaptive Binary Arithmetic Coding
CAVLC	Context Adaptive Variable Length Coding
CD	Compact disc
CIF	Common Intermediate Format
Codec	COmpressor/DECompressor

D

DCT	Discrete cosine transform
DPB	Decoded Picture Buffer
DVD	Digital Video Disc

F

FPS	Frame Per Second

G

GB	Gigabyte
Gbps	Gigabit per second
GOP	Group of pictures

H

HDTV	High-definition television
HDV	High-definition video

I

ISDN	Integrated Services Digital Network
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
ITU-T	International Telecommunication Union-Telecommunication

J

JSVM	Joint scalable video model
JMVM	joint multiview video model
JVT	Joint video team

L

LCD	Liquid crystal display

M

Mbps	Megabits per second
MPEG	Moving Picture Experts Group
MSE	Mean squared error
MVC	Multiview video coding
MVV	Multiview video

P

PSNR	Peak Signal to Noise Ratio

Q

QP	Quantization parameter
QCIF	Quarter CIF

R

RDO	Rate-Distortion Optimization

S

SDTV	Standard Definition television
SSIM	Structural Similarity Index
SVCD	Super Video CD
SVGA	Super Video Graphics Array

T

TV	Television

V

VCEG	ITU-T Video Coding Experts Group
VGA	Video Graphics Array

Summary

3D display technology has witnessed a rapid development in the past decades. Currently 3D displays are being widely used in different application areas such as education, broadcasting, entertainment, surgery, video conferencing, etc. However, this technology owes its success to the other complementary technologies such as image acquisition, compression and transmission. The compression is the main topic of this dissertation.

In multiview displays, the realism of the reproduced 3D scene is dependent on the number of available views that the displays can show. These view images can be captured from different viewpoints of a scene by using a camera array. A smoother transition between views can be obtained by increasing the number of cameras located in the camera array. However, this comes at the price of an increased amount of image data which needs to be encoded (compressed) to store and transmit the data efficiently.

The captured multiview videos for different views can be encoded separately by a state-of-the-art video codec like H.264/AVC, which is called simulcast coding. Although coding each video individually is an easy option to solve the problem, it is not the most efficient approach since the inter-view correlations between views are overlooked. In order to improve the efficiency of the encoder, the inter-view correspondences can be taken into account. However, in this case, the computational load of the encoder becomes very high, especially if the camera array is two dimensional, i.e. if vertically spaced views are also being captured. Limitations on processing power, memory requirement and the desirability of features like instant access to specific view frame in the multiview video may render this scheme unusable.

A periodic (2D) camera array results in a strong geometrical relationship among the captured view images. This fact forms the core of the methods I propose in this dissertation, which consequently reduces the complexity of the multiview encoder significantly.

The P and B frames are well-known frame types from the H.264/AVC video coding standard and are instrumental in the motion estimation pro-

cess that exploits the similarity between consecutive frames in the time domain. A similar process called disparity estimation can be used to exploit the similarity between views. Since the B frame offers bi-directional prediction, it shows a better coding performance than the P frame but brings high computational load to the encoder. The complexity efficient versions of these frame types, named the D_P and D_B frame respectively, are introduced in this dissertation.

The disparity estimation process is the most complex and time consuming part of the encoder. The D_P frame achieves a significant complexity reduction by skipping the disparity estimation process for some of its blocks. The skipping process is entirely based on the fact that the disparity vectors of the blocks in a D_P frame can for most blocks be derived from the previously encoded blocks in another frame due to the strong geometrical relationship between views. A derived disparity vector needs to be checked for its fidelity since the derivation process can fail in some blocks due to occlusions, anisotropic illumination effects or insufficient texture information. To do this, the rate-distortion cost value of the derived disparity vector is compared with a threshold value. The blocks whose RD cost value of the derived disparity vector is lower than the threshold value will be exempted from the disparity estimation process, which results in a net complexity gain for the encoder.

The threshold value plays a crucial role on the determination of convenience of the derived disparity vectors. Since the D_P frame is a modified standard-conforming P frame, the encoder shows the same coding performance as the P frame if the threshold value is set equal to zero. It means that none of the derived disparity vectors are used and the disparity estimation will be performed for all the blocks in the frame. As the threshold value increases, the complexity of the encoder decreases, while the quality and bit-rate of the encoded images degrade. Therefore, the threshold value is a parameter to adjust the trade-off between the complexity and the rate-distortion performance of the encoder.

I introduced five alternative prediction schemes to encode 5x3 view images taken from a 2D camera array, three of which were constructed with the D_P frame. It has been noticed that the different locations of the frames have influence on the rate-distortion performance of the encoder. The prediction scheme in which the I frame is placed in the middle gives the best performance. After investigating the impact of a wide range of threshold values on the encoding performance and the complexity, it has been realized that the complexity of the encoder can substantially be reduced without compromising the quality and bitrate. The optimum values of the

threshold should be calculated depending on the quantization parameter (QP) and the imagery content since the threshold value represents a point in the rate-distortion cost value scale.

The rate-distortion and the complexity performance of the multiview encoder are improved by applying individual threshold values for every block in a D_P frame. I propose a method which automatically calculates the optimum threshold values for blocks during the encoding, where the maximum complexity gain is achieved while maintaining the rate-distortion performance. In order to calculate the optimum threshold value of a block, the rate-distortion cost value of a previously encoded block from which the disparity vector is derived is utilized.

Basically, the B frame has a better coding efficiency than the P frame. When the B frame is employed in a prediction scheme to encode multiview images, the complexity of the encoder is much higher than the prediction schemes with only P frames. In this dissertation, the complexity efficient version of the B frame, called D_B frame, is also presented. Different prediction schemes constructed with the D_P and the D_B frames are proposed. With the help of the D_B frame, the computational load of the multiview encoder is reduced considerably. Automatic threshold values of the blocks in D_B frames are automatically generated during the encoding.

The proposed frame types allow us to encode multiview images effectively with a lower computational load while keeping the same quality and bitrate. All proposed prediction schemes are applied to different multiview image sets containing various real world objects. For this purpose, all ideas in this dissertation have been implemented in the JSVM reference software.

Samenvatting

De technologie van driedimensionale (3D) beeldschermen heeft de afgelopen decennia een snelle ontwikkeling gekend. Momenteel worden 3D-beeldschermen gebruikt in diverse toepassingen zoals onderwijs, videoconferenties, chirurgie, ontspanning en televisie-uitzendingen. Het succes van deze technologie is echter alleen maar mogelijk in combinatie met andere, complementaire, technologieën zoals beeldregistratie, datacompressie en –transmissie. Datacompressie vormt het belangrijkste onderwerp van dit proefschrift.

In de zogenaamde 'multiview' beeldschermen wordt het realisme van de gereproduceerde 3D-scène bepaald door het aantal verschillende gezichtshoeken ('views') dat het beeldscherm tegelijk kan weergeven. De beeldinformatie die bij al deze gezichtshoeken hoort, kan bijvoorbeeld opgenomen worden door middel van een rooster van camera's die op dezelfde scène gericht zijn. De overgangen tussen de verschillende gezichtshoeken worden vloeiender naarmate de camera's in het rooster minder ver uit elkaar staan en het aantal camera's dus hoger is. Dit gaat echter ten koste van een toegenomen datavolume dat moet gecodeerd, opgeslagen en getransporteerd worden.

In principe kan men de opgenomen multiview videos voor elke gezichtshoek apart coderen met een state-of-the-art video codec zoals H.264/AVC. Deze aanpak wordt simulcast-codering genoemd. Hoewel dit een eenvoudige optie is om het probleem op te lossen, is dit duidelijk niet de meest efficiënte aanpak aangezien de correlatie tussen de beelden geregistreerd vanuit verschillende gezichtshoeken over het hoofd worden gezien. Om de efficiëntie van de encoder te vergroten kunnen deze 'interview' gelijkenissen in rekening gebracht worden. Evenwel wordt in dit geval de rekenlast waarmee men de encoder opzadelt zeer hoog, in het bijzonder wanneer het camerarooster tweedimensionaal is en er dus ook verticaal gespatieerde gezichtshoeken worden geregistreerd. Beperkingen op de rekenkracht en het beschikbare geheugen alsook gewenste eigenschappen zoals het snel kunnen springen naar een willekeurig tijdstip in

de multiview video zouden deze aanpak onbruikbaar kunnen maken.

In een periodiek tweedimensionaal rooster van camera's ontstaat een sterk geometrisch verband tussen de beelden die worden geregistreerd door de individuele camera's. Deze vaststelling vormt de basis van de methodes die ik in dit proefschrift voorstel om de complexiteit van de multiview encoder aanzienlijk te reduceren.

In de H.264/AVC videocoderingsstandaard zijn bewegingsvectordetectie en 'P' en 'B'-beelden welbekende begrippen waarmee men de gelijkenissen tussen elkaar in de tijd opvolgende beelden benut om zo tot een belangrijke datareductie te komen. Dezelfde technieken kunnen gebruikt worden om de inter-view gelijkenissen te benutten. In plaats van bewegingsvectoren spreekt men hier van dispariteitsvectoren. Aangezien B beelden bi-directioneel mogelijk maken, laat dit toe een betere coderingsperformantie te bereiken dan met louter P beelden, ten koste van een hogere berekeningslast voor de encoder. In dit proefschrift introduceer ik nieuwe versies van deze beeldtypes, D_P respectievelijk D_B, die deze berekeningslast ('complexiteit') beperken.

Het proces dat de dispariteit tussen de verschillende views zoekt is het meest complexe en tijdrovende onderdeel van de multiview encoder. De introductie van het D_P-beeld realiseert een significante reductie van de rekenlast doordat het toelaat om de dispariteitsvectordetectie over te slaan voor sommige delen (pixelblokken) in ieder beeld. Dit is mogelijk doordat de dispariteitsvectoren voor de meeste pixelblokken in een D_P-beeld kunnen afgeleid worden uit de blokken die reeds in een ander beeld werden gecodeerd, door gebruik te maken van het geometrisch verband tussen de verschillende views. Dergelijke afgeleide dispariteitsvectoren moeten gecontroleerd worden op hun geschiktheid omdat het geometrisch afleidingsproces voor sommige blokken niet goed werkt als gevolg van occlusies, anisotrope belichtingseffecten of onvoldoende informatie over de textuur. Teneinde dit te doen wordt de zogenaamde 'rate-distortion' (RD) kost van de afgeleide dispariteitsvector vergeleken met een drempelwaarde. De blokken waarvoor de RD-kost lager is dan deze drempelwaarde worden uitgesloten van het dispariteitsvectordetectieproces, hetgeen leidt tot een vermindering van de complexiteit van de encoder.

De geïntroduceerde drempelwaarde speelt dus een cruciale rol bij het bepalen van de geschiktheid van de afgeleide dispariteitsvectoren. Aangezien een D_P-beeld een gemodificeerd standaard P-beeld is, vertoont de encoder net dezelfde performantie als de standaardencoder indien de drempelwaarde gelijk aan nul wordt gekozen. In dat geval wordt immers geen enkele van de geometrisch afgeleide dispariteitsvectoren gebruikt,

maar wordt het reguliere dispariteitsvectordetectieproces uitgevoerd voor alle pixelblokken. Naarmate de drempelwaarde stijgt, neemt de complexiteit van de encoder af, maar worden de kwaliteit en de bitrate slechter. Bijgevolg is de drempelwaarde een parameter die toelaat de afweging tussen complexiteit en rate-distortion van de encoder bij te sturen.

Ik heb vijf alternatieve voorspellingsschema's geïntroduceerd om een multiview registratie van 5x3 views, genomen door een tweedimensionaal camerarooster, te encoderen. Drie van deze schema's maken gebruik van D_P-beelden. Daarbij is gebleken dat de locatie van de verschillende beeldtypes in de schema's een invloed hebben op de RD performantie van de encoder. Het voorspellingsschema waarbij het I-beeld in het midden wordt geplaatst vertoont de beste performantie. Uit een grondige studie van de impact van een groot aantal drempelwaarden op de performantie en complexiteit van de encoder is gebleken dat de complexiteit van de encoder gevoelig kan gereduceerd worden zonder dat de kwaliteit en bitrate merkbaar worden aangetast. De optimum waarde voor de drempel hangt daarbij af van de kwantiseringsparameter (QP) alsook van de beeldinhoud vermits de drempelwaarde een punt vertegenwoordigt in de RD-kost schaal.

De rate-distortion en de complexiteit van de multiview encoder kunnen verder verbeterd worden door aangepaste drempelwaarden toe te passen op ieder pixelblok in een D_P-beeld. Ik heb een methode voorgesteld waarmee de optimale drempelwaarde automatisch kan berekend worden tijdens het coderingsproces. Daarmee kan de maximale winst in complexiteit behaald worden zonder dat de RD-performantie afneemt. Om de optimale drempelwaarde te berekenen wordt gebruik gemaakt van de RD kost waarde van het voordien geëncodeerde pixelblok waaruit ook de dispariteitsvector werd afgeleid.

Een B-beeld leidt tot een grotere coderingsefficiëntie. Introductie van het B-beeld in het voorspellingsschema voor multiview beelden leidt echter tot een gevoelige toename van de complexiteit van de encoder. In dit proefschrift wordt ook een variant geïntroduceerd op het B-beeld, D_B-beeld genoemd, waarbij de geometrische afleiding van dispariteitsvectoren leidt tot een reductie van de rekenlast. Verscheidene voorspellingsschema's die gebruik maken van deze D_P en D_B-beelden worden voorgesteld. Door de introductie van het D_B-beeld wordt de complexiteit van de multiview encoder aanzienlijk verminderd. Ook de drempelwaarden voor de D_B-blokken worden automatisch gegenereerd tijdens het coderingsproces.

De voorgestelde nieuwe beeldtypes laten ons toe om multiview beelden te coderen met een lagere rekenlast terwijl de bitrate en beeldkwaliteit

gelijk blijven. Alle voorgestelde voorspellingsschema's zijn toegepast op verscheidene multiview opnames van reële objecten. Daartoe werden alle in dit proefschrift geïntroduceerde concepten geïmplementeerd in de JSVM referentiesoftware.

1
General introduction

"Once a new technology starts rolling,
if you're not part of the steamroller,
you're part of the road."

—Stewart Brand

The current century is witnessing the booming popularity of 3D displays in a wide range of application areas from entertainment to medical. Although the displays are the last ring of a long essential process chain after the 3D data acquisition, compression and transmission, they are the most distinguishable and visible component for the users. However, all these steps are equally important and complement each other.

A 3D display offers a realistic depth perception of a captured scene to its viewers. The 3D display technologies can be categorized in different ways. A common feature of all 3D display technologies is to provide at least two views to the audience. While stereoscopic displays offer only two separated views, multiview displays can show a multitude of captured views at the same time to the viewers. The number of supported views in a multiview display is one of the biggest challenges among the researchers in 3D display technology domain.

In most multiview 3D displays, each individual view is actually a conventional 2D video. Computer generated multiview videos containing all

supported views are highly popular. One of the biggest reasons is that the methods to capture the real-world multiview videos require an expensive camera setup and do not allow to have a movie with imaginative creatures.

It is plausible that more viewing zones in a multiview 3D display result in a more realistic experience. Most 3D imaging systems only provide views taken from horizontally spaced viewpoints since vertically spaced views yield cumbersome problems such as complex optical design, difficult implementation, high system cost and huge data requirements. However, vertically spaced views are indispensable for any real 3D experience since they provide supplementary information of the scene.

One of the most straightforward and easy ways to capture real-world 3D videos is to use a 2D camera array in which the cameras are arranged in a plane. Increasing the number of cameras located in the camera array also increases the captured data size which is not suitable for efficient storage and transmission. Therefore, a high fidelity data compression, which is another ring in the process chain, is indispensable.

There are different ways of compressing (encoding) the multiview videos. The first and simplest method is to encode every view separately by a state-of-the-art video codec like H.264/AVC (Advanced Video Coding). Although this so called simulcast coding is an easy option to solve the problem, it is not the most efficient approach. The multiview videos contain high inter-view correlations since each view is taken from a slightly different perspective of the same scene. In order to improve the coding efficiency, the Joint Video Team (JVT) has developed a multiview video coding extension (MVC) of the H.264/AVC in which the correlations in the view domain as well as in the temporal domain are exploited. However, all these attempts to improve the coding efficiency bring extra computational load to the multiview encoder, which needs to be reduced due to practical reasons.

In this encoding, I focused on reducing the complexity of multiview video encoder by utilizing the strong geometrical relationship between captured views while keeping the quality and the bitrate the same. For this purpose, I propose complexity efficient versions of the frame types from the H.264/AVC video coding standard. I also introduce different prediction schemes and new frame types.

Outline

The contents of some of the chapters in this dissertation have been published in different journals and conference proceedings, which are listed

in the publication list at the end of this chapter. This work is supported by the Research Foundation-Flanders (FWO-Vlaanderen). The project is titled "Compact LCOS projection displays for high-quality 3D images with high spatial and angular resolution".

The dissertation starts with a general overview about the 3D display technologies. This chapter includes details about how the 3D perception occurs in a human brain. Different 3D displays technologies with which the viewers can have the 3D impression are presented. The chapter concludes with several data acquisition methods to capture real-world multiview videos.

In Chapter 3, the H.264/AVC video encoding standard is described in detail. In this chapter, the encoding process steps to compress a conventional 2D video will be explained. This process will further be used in Chapter 4 to explain the multiview video coding extension of H.264/AVC. The information presented in these two chapters will serve as a background for the entire PhD work.

The complexity problem in P frames will be dealt with in Chapter 5. The geometrical relationships between the view images will be explained and a complexity efficient version of the P frame utilizing this relationship will be proposed in this chapter. The influence of the different locations of I and P and the complexity efficient P frame on the coding performance will also be investigated. The threshold value is to adjust the trade-off between the complexity and the rate-distortion performance of the encoder is chosen as a predefined fixed value. It will be presented that the complexity and rate-distortion performance results can be improved by selecting the threshold value depending on the QP.

In Chapter 6, it will be explained that the complexity efficiency of the P frame can further be reduced without compromising the quality and bitrate by dynamically calculating the threshold values. Different weighting functions to estimate the threshold values from the extracted rate-distortion cost values will be introduced and their efficiencies will be compared. Moreover, it will be shown that the coding performance of the encoder can considerably be increased by calculating the threshold values for every block instead of one fixed value for whole frame.

In Chapter 7, a complexity efficient version of a B frame will be introduced. Novel prediction schemes constructed with this frame type will be demonstrated and their coding performances will be compared to each other in terms of rate-distortion and complexity. The overall coding performance of the prediction schemes from this chapter and from the previous chapters will also be compared to each other.

Chapter 8 will include the conclusions and some ideas for future work.

Research contributions

Here, I present a list of the research contributions within this dissertation:

- Design of the complexity efficient P and B frame types,
- Implementation of both complexity efficient frame types in JSVM reference software,
- Development of a master software to automatize the encoding process for whole prediction scheme,
- Investigate the influence of the different locations of the frame types on the coding performance,
- Design of different alternative prediction schemes constructed
  - with all frame types available from the video coding standards,
  - with the complexity efficient version of the P frame,
  - with the complexity efficient version of the B frame,
  - with all possible frame types both from the video coding standards and the complexity efficient frame types,
- Development of different weighting functions to calculate the optimum threshold values for the blocks,
  - Design, implementation and testing of two different weighting functions,
  - Investigate the optimum construction parameters of both weighting functions for both complexity efficient frame types,
  - Design of a new version of the software which is compatible with proposed weighting functions,
- In-depth analysis of the results.

Papers in international journals

(listed in the Science Citation Index [1])

1. **Avci A.,** De Cock J., De Smet J., Joshi P., Meuret Y., Lambert P. and De Smet H., "Novel complexity efficient prediction schemes for 2D camera arrays," *IEEE Trans. Consum. Electron.*:**In preparation**

2. **Avci A.,** De Cock J., De Smet J., Meuret Y., Lambert P. and De Smet H., "Reduced-complexity multiview prediction scheme with content-adaptive disparity vector estimation," *J. Electron. Imaging,* vol. 21, no. 3, 2012.

3. **Avci A.,** De Cock J., Lambert P., Beernaert R., De Smet J., Bogaert L., Meuret Y., Thienpont H. and De Smet H., "Efficient disparity vector prediction schemes with modified P frame for 2D camera arrays," *J. Vis. Commun. Image Represent.,* vol. 23, no. 2, pp. 287-292, 2012.

4. De Smet J., **Avci A.,** Beernaert R., Cuypers D. and De Smet H., "Design and wrinkling behavior of a contact lens with an integrated liquid crystal light modulator," *J. Disp. Technol.,* vol. 8, no. 5, pp. 299-305, 2012.

5. Beernaert R., **Avci A.,** De Smet J., De Smet H., De Coster J., Severi S. and Witvrouw A., "Novel analog pulse-width-modulated 15-μm SiGe micromirrors," *J. Soc. Inf. Display,* vol. 18, no. 10, pp. 855-861, 2010.

6. Beernaert R., Podprocky T., De Coster J., Witvrouw A., Haspeslagh L., **Avci A.,** De Smet J. and De Smet H., "Novel micromirror design with variable pull-in voltage," *Microelectron. Eng.,* vol. 87, no. 5-8, pp. 1248-1252, 2010.

7. Bogaert L., Meuret Y., Roelandt S., **Avci A.,** De Smet H. and Thienpont H., "Demonstration of a multiview projection display using decentered microlens arrays," *Opt. Express,* vol. 18, no. 25, pp. 26092-26106, 2010.

[1] The publications listed are recognized as 'A1 publications', according to the following definition used by Ghent University: A1 publications are articles listed in the Science Citation Index, the Social Science Citation Index or the Arts and Humanities Citation Index of the ISI Web of Science, restricted to contributions listed as article, review, letter, note or proceedings paper.

8. Avci N., Musschoot J., Smet P., Korthout K., **Avci A.,** Detavernier C. and Poelman D., "Microencapsulation of Moisture-Sensitive CaS:Eu2+ Particles with Aluminum Oxide," *J. Electrochem. Soc.,* vol. 156, no. 11, pp. J333-J337 , 2009.

9. Murat H., **Avci A.,** Beernaert R., Dhaenens K., De Smet H., Bogaert L., Meuret Y. and Thienpont H., "Two LCOS full color projector with efficient LED illumination engine," *Displays,* vol. 30, no. 4-5, pp. 155-163, 2009.s

Papers in international conferences

(listed in the Science Citation Index [2])

1. **Avci A.,** De Cock J., De Smet J., Meuret Y., Lambert P. and De Smet H., "A content-adaptive scheme for reduced-complexity, multi-view video coding," *Proc. SPIE Int. Soc. Opt. Eng.,* vol. 8290, no. 829014, 2012.

2. **Avci A.,** De Cock J., Beernaert R., De Smet J., Bogaert L., Meuret Y., Thienpont H., Lambert P. and De Smet H., "Reduced complexity multi-view video coding scheme for 2D camera arrays," *Proc. SPIE Int. Soc. Opt. Eng.,* vol. 8290, no. 829014, 2012.

3. **Avci A.,** Bogaert L., Beernaert R., De Smet J., Meuret Y., Thienpont H. and De Smet H., "Efficient disparity vector coding for multi-view 3-D displays," *Proc. SPIE Int. Soc. Opt. Eng.,* vol. 7526, no. 752609, 2010.

4. Beernaert R., De Coster J., Podprocky T., Witvrouw A., Severi S.,**Avci A.,** De Smet J. and De Smet H., "SiGe micromirrors for optical applications," *Photonics North 2010,* vol. 7750, no. 77501S, 2010.

5. Bogaert L., Meuret Y., Roelandt S., **Avci A.,** De Smet H. and Thienpont H., "Single projector multiview displays : directional illumination compared to beam steering," *Proc. SPIE Int. Soc. Opt. Eng.,* vol. 7524, no. 75241R, 2010.

[2]The publications listed are recognized as 'P1 publications', according to the following definition used by Ghent University: P1 publications are proceedings listed in the Conference Proceedings Citation Index - Science or Conference Proceedings Citation Index - Social Science and Humanities of the ISI Web of Science, restricted to contributions listed as article, review, letter, note or proceedings paper, except for publications that are classified as A1.

6. Meuret Y., Bogaert L., Roelandt S., Vanderheijden J., **Avci A.,** De Smet H. and Thienpont H., "LED projection architectures for stereoscopic and multiview 3D displays," *Proc. SPIE Int. Soc. Opt. Eng.,* vol. 7690, no. 769007, 2010.

7. Roelandt S., Bogaert L., Meuret Y., **Avci A.,** De Smet H. and Thienpont H., "Color uniformity in compact LED illumination for DMD projectors," *Proc. SPIE Int. Soc. Opt. Eng.,* vol. 7723, no. 77230X, 2010.

8. Beernaert R., Podprocky T., **Avci A.,** De Smet J. and De Smet H., "Micromirror with electromechanical pulse width modulation," *29th International Display Research Conference (IDRC 2009 / Eurodisplay 2009),* pp. 428-431, 2009.

Publications in other international conferences

1. De Smet J., **Avci A.,** Beernaert R., Cuypers D. and De Smet H., "Wrinkle formation in conformable liquid crystal cells for use in a contact lens display," *International Display Workshops 2011 (IDW '11),* pp. 1203-1206, 2011.

2. De Smet J., **Avci A.,** Beernaert R., Cuypers D. and De Smet H., "Spherically curved guest-host display for use in a contact lens," *17th International Display Workshops (IDW 2010),* pp. 1585-1586, 2010.

3. Beernaert R., Podprocky T., Witvrouw A., Haspeslagh L., **Avci A.,** De Smet J. and De Smet H., "Novel micromirror design with variable pull-in voltage," *35th International Conference on Micro- and Nano-Engineering (MNE 2009),* 2009.

2

3D display technologies

"It is in vain to say human beings ought to be satisfied with tranquillity: they must have action; and they will make it if they cannot find it."

—**Charlotte Bronte,** *Jane Eyre*

2.1 Introduction

The prevalence of the displays in our life is rapidly growing, comprising applications ranging from medicine to entertainment. In order to keep pace with this evolution, the researchers and technology developers are designing new displays for different application areas. As a recent example, an innovative spherical curved LCD display has been embedded into a contact lens. This technology can be used in many different application areas, since the whole surface of the lens can be used as a display [1–3].

The 3D displays add one dimension, the depth, to the traditional and well-known 2D displays. Although we are constantly aware of these three dimensions in our daily life, it is a hard and challenging task to visualize all dimensions artificially onto an electronic display device.

As a viewer relishes the 3D content being projected or displayed, many

underlying technical sophistications go unnoticed such as data acquisition, encoding, transmission, decoding and reconstruction. However, the display itself is the most significant aspect for the users since it is the only visible part of the process chain.

In this chapter, two constituents of the process chain, namely the different 3D display technologies and the data acquisition methods will be discussed.

2.2 Human 3D perception

The physical environment that we are living in is essentially three-dimensional, meaning that every object or room can be characterized as having a width, depth and height. We see the objects in 3D every day since we have 3D perception which is also known as depth perception. Humans have two eyes which are horizontally separated from each other. As we look around, the retina in each eye receives two slightly different two-dimensional images, which resulting in so-called stereoscopic or binocular vision. Our brain converts these two images together into a 3D visual experience, called 3D reconstruction. However, this is not the only way to see 3D. People who can only see with one eye, which is called monocular vision, can still perceive the scene in 3D [4, 5]. Basically, those people miss only one so-called 'depth cue' but can rely on the others to see the 3D without thinking about it. Some of the cues that humans use to perceive the 3D are:

- **Stereoscopic vision:** Both eyes register slightly different images. Objects closer to the observer appear to occupy different locations with respect to the background. The further away the objects, the less the apparent displacement of the objects between the left and the right view.

- **Accommodation:** The lenses of the eyes change their physical shape as the observer focuses on an object which is standing closer or further away. This provides information about the distance of the object in the scene.

- **Motion parallax:** Objects which are closer seem to move faster than the objects in the distance. The brain senses the depth information of an object by looking at its relative motion against a background object.

- **Size familiarity:** The distance information of an object can approximately be perceived by looking at its visible size with which the observer is familiar.
- **Occlusion:** The occlusion happens when the near surface overlaps the far surface, which provides information about distance.
- **Aerial perspective:** Due to its nature, vapour and the particles in the atmosphere causes light to scatter. Based on this fact, further objects have lower contrast and saturation than closer objects, which also gives information about the distance of the object.

2.3 3D display systems

Present technology enables us to have many degrees of freedom for designing the displays that are capable of facilitating a 3D experience to the viewers. The ultimate aim of all 3D displays is to create an optical illusion of the real scene so that the viewer feels fully to be in the real environment [6]. Yet, this objective is far from being realized in practice. Researchers have taken diverse routes to accomplish this ultimate 3D display. Each approach has its own characteristic advantages, promises and limitations. In this section, the major 3D display technologies will be discussed, but more detailed information can be found in [7].

2.3.1 Stereoscopic displays

Stereoscopic displays require the viewer to use a device enabling each eye to see correct images. Although wearing or being very close to a device to filter the correct images for each eye is considered a major drawback of these systems, they are reasonable and being widely used today for the applications where multiple observers are at present such as cinemas or large group activities.

The first stereoscope was invented in 1838, a couple of years before the invention of photography, by Sir Charles Wheatstone. He received his first patent for his device so-called reflecting mirror stereoscope, shown in Figure 2.1.a. In his experiments, he used simple stereoscopic pictures containing geometric 3D drawings. Wheatstone proved with his invention that the stereo perception occurs as a result of binocular vision. The technological development of 3D continued with the invention of the refracting stereoscope, called lenticular stereoscope, by Sir David Brewster in 1848, shown in Figure 2.1.b. Since he used lenses instead of mirrors or prisms,

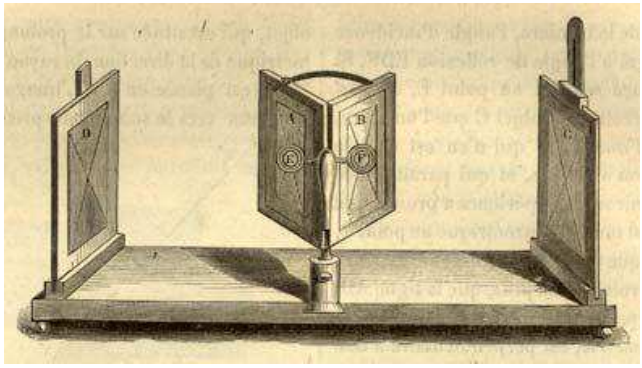
his device was much more compact than the other devices until that time due to the shorter focal length property of the lenses. Later on, Oliver Wendell Holmes invented the best-known version of the lenticular lens stereoscope in 1862, Figure 2.1.c.

In all above mentioned inventions, the viewing path for each eye was physically separated from each other, preventing one eye to see the image of the other eye. At the present day, this separation is realized by means of using special eyeglasses which ensure that each eye sees the correct image.

3D eyeglass technology can basically be grouped in two categories, active and passive eyeglasses. The active glasses contain a battery to power the integrated wireless system to establish synchronization between the eyeglass and the display and/or the LCD lenses of the eyeglass to open and shut at different times. One of the major drawbacks of the active shutter glasses is that they are more expensive and heavier than the passive ones. Another drawback is that both eyes of the observer can not see the image at the same time, only one eye sees an image while the other eye can not [11]. Passive glasses do not require any source of power and that makes them cheaper. There are two types of passive glasses, anaglyph and polarized. The anaglyph glasses contain most-commonly red and cyan coloured lenses which filter the incoming color coded images from the display. The downside of the anaglyph glasses is that some of the color information in the original image is lost during the production of the images. The polarized glasses allow the correct image to reach the eye since the direction of the polarisers for each lens is different. In linearly polarized glasses, one eye can see the image for the other eye if the observer tilts his head in one direction. This phenomena is called crosstalk or ghosting [12]. This problem has remarkably been solved by circularly polarized lenses. The viewers can tilt their head without seeing any artifacts.

Today, the stereoscopic displays are widely used in scientific visualizations, medical imaging, games, televisions, entertainments, communications, education, CAD/CAM applications, industrial inspections and advertisements [13–17].

There are other stereoscopic systems where the viewer can see 3D without wearing any special glasses, called autostereoscopic displays. Although they are multiview displays in most of the existing implementations, as it will be mentioned in the next section, there are some cases in which the autostereoscopic displays show only stereoscopic images to the viewer. These systems are designed in such a way that only one observer, mostly



(a)

(b)

(c)

Figure 2.1: (a) Wheatstone's stereoscope [8]. (b) Brewster-type stereoscope, 1849. [9]. (c) Holmes-type stereoscope [10].

speaking, can see the 3D either from a fixed distance or mobile with the aid of head-tracking systems [18–20].

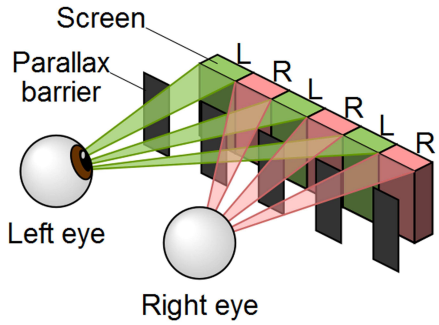
2.3.2 Multiview displays

Multiview displays are the displays which are capable of showing multiple images taken from different viewpoints of a scene to multiple zones [21]. They are built with special optics to direct the image information to the certain zones called viewing zones. Any observer can move freely within the allowed viewing range and see different perspectives of the scene. This makes it more realistic and is considered as the biggest advantage of the multiview displays. It is desired that the multiview display supports as many viewing zones as possible. As the number of viewing zones increases, the display exhibits more realistic 3D to the viewers. However, building the display becomes more difficult since the display needs to show the correct view images to each viewing zone simultaneously. Furthermore, the amount of necessary image information increases accordingly, which will be discussed in Section 2.4.

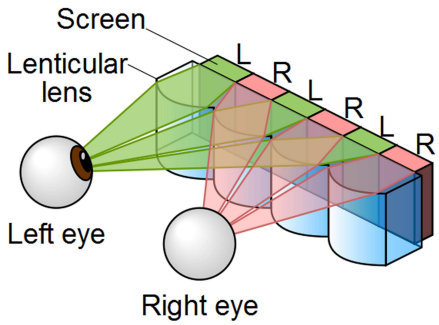
The two most common technologies for the multiview displays are:

- **Parallax-barrier displays:** The name is self-explanatory in that a barrier consisting of slits is placed in front of the image source, e.g. an LCD, so that each eye of the observer can see different pixels, see Figure 2.2.a [22–24]. The major drawback of this technology is that the viewer is restricted to a specific position, otherwise he can not see the 3D effects. Besides this, the screen resolution that each eye sees is half of the resolution of the image source. However, they are easily attachable and comparably cheaper. The parallax-barrier technology can be used for 2D/3D switchable displays [25, 26]. In this method, the barrier is constructed with an additional LCD panel and the slits are created by switching the appropriate pixel locations on the LCD on and off.

- **Lenticular displays:** Instead of a barrier, a sheet of cylindrical lenses (lenticulars) is located in front of the image source maintaining the fact that the image source is in the focal plane of the lenticulars, Figure 2.2.b. This construction makes it possible to see different pixels on the image source when viewed from different directions. Since the pixels which are encompassed by a lenticular constitute only one image pixel of a particular viewing zone, the spatial resolution of one view image is smaller than the native resolution of



(a)



(b)

Figure 2.2: Multi view display technologies [27](a) Parallax-barrier approach. (b) Lenticular lens approach.

the image source. The resolution decreases horizontally if vertically oriented cylindrical lenses are used, which is undesired. This issue can be overcome by slanting the lenticulars at a small angle so that the resolution reduction occurs both horizontally and vertically [28, 29].

As another technology, a multiview projection display is realized based on beam steering using decentered microlens arrays in the projection screen and time-sequential rear-projection of the view images. 27 viewing zones with XGA resolution in total and a horizontal field of view of 30° have been achieved with this technology [30, 31].

The view images are sequentially modulated by a single 2D display device in all above mentioned methods. There are some 3D visualization systems in which the view images for each viewing zone are projected by

individual projectors [32, 33]. By this way, while the resolution of the viewable images can have the same resolution of the projector, the overall system is bulky due to the number of projectors. As another method, a single imaging device can be used to show all view images on the display simultaneously. In this case, very fast imaging devices like digital micromirror arrays are essential [34–37].

2.3.3 Holographic displays

Holographic displays are the displays which are able to reconstruct the holograms that are basically the wavefront of a real object [38, 39]. This technology requires no eyeglasses and different perspectives of an object can be seen from different angles seamlessly.

The holograms which are used in holographic displays are not like view images that are used in the other display technologies and they contain the information regarding the light scattered from the object. Since the hologram is a record containing interference pattern of an object, typically lasers are used as a source of coherent light. The reflected light beam from the object and the reference beam are combined at the film and generate the hologram of the object, Figure 2.3a.

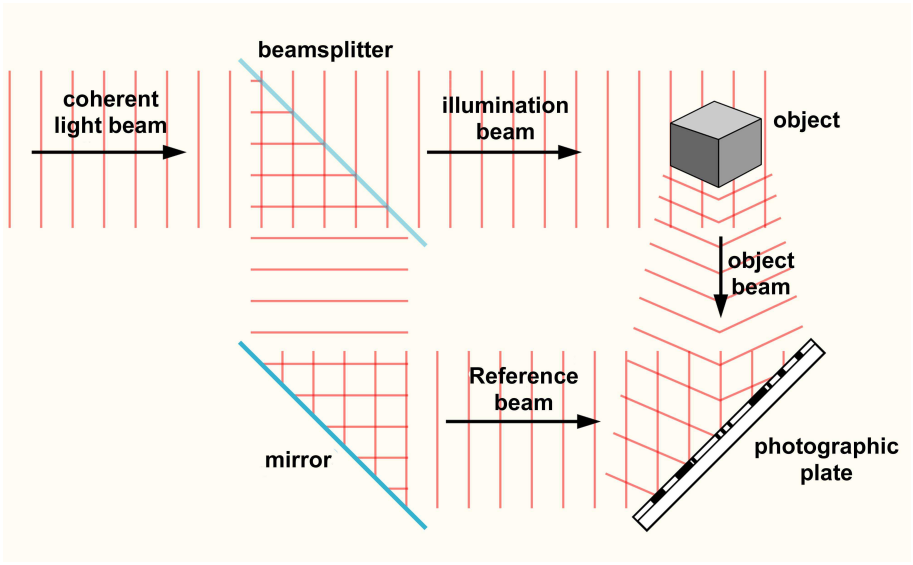
In order to reconstruct the hologram, a reference beam identical to the one that was used to record the hologram needs to be used. When the hologram is illuminated by this light source, there is no difference between the reconstructed hologram and the original scene. The reconstruction process is drawn in Fig 2.3b.

2.3.4 Volumetric displays

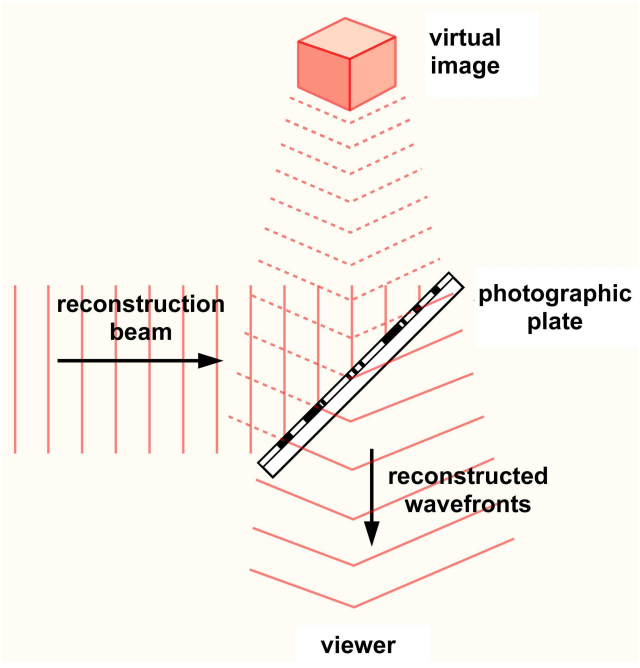
Volumetric displays are another type of 3D displays in which the realistic 3D reconstruction of an object can clearly be seen by the naked eye. In this technology, the image points are projected onto the physical volume of space whereby each volume element is called a 'voxel'.

Based on the technology, the volumetric displays are basically divided into two groups:

- **Swept volume technology:** The volumetric display based on swept volume technology generates the voxels on a rotating 2D screen at a very high rate in order to prevent visual aberration. However, this causes some problems such as noise, instability and short life time. An example of a swept volume display is shown in Figure 2.4.



(a)



(b)

Figure 2.3: (a) Recording process of a hologram [40]. (b) Reconstruction process of a hologram, a holographic display [40].

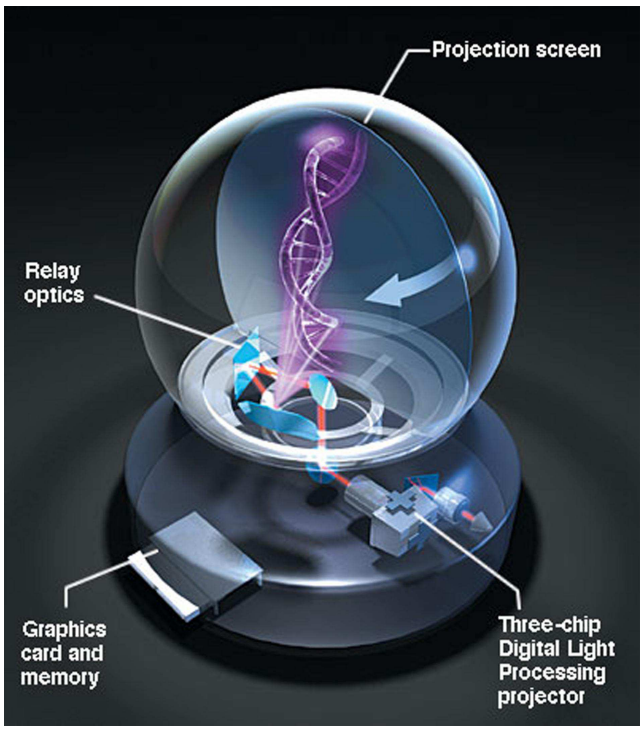


Figure 2.4: The Perspecta swept-volume display [41].

- **Static volume technology:** In static volume technology, there are no such moving parts like in the swept volume technology, which is the major advantage of this technology. One type of static volume display uses two laser sources of different wavelengths. The voxels are created at the intersection point of the beams coming from the two laser sources, thus emit light. Another type uses a stack of LCD panels. To create a certain voxel, the respective part of nineteen of the twenty LCD screens becomes transparent. Since this process is done for every voxels at a fast rate, the 3D object appears as it is shown in Figure 2.5.

2.4 Multi view data acquisition

All display technologies mentioned in the previous section are capable of showing at least two different views that are taken from various perspectives of a scene. These view images can be either captured by a set of

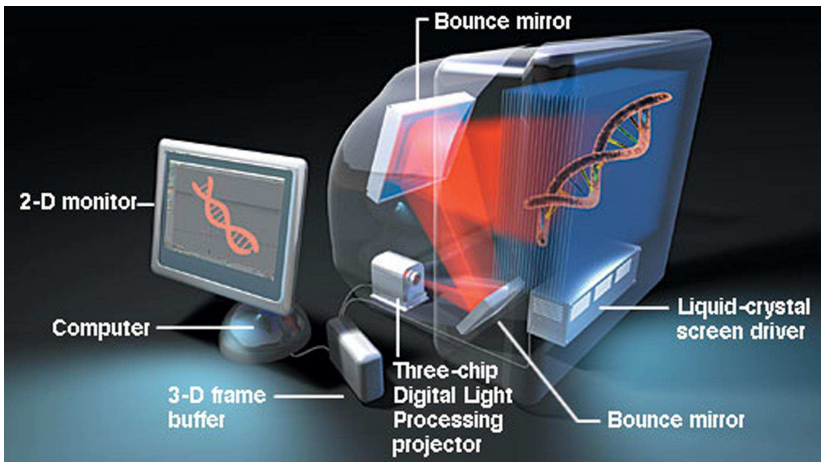


Figure 2.5: LightScape DepthCube volumetric display [41].

cameras or generated artificially by computer graphics. Nowadays, most of the 3D movies are computer animated mainly due to easier content creation. Once the content is created, the scene can be rendered for as many different camera positions as required. Different methods exist to capture the real-world multiview videos, and will be discussed separately in the next section.

2.4.1 Integral photography

This method was invented by Gabriel Lippmann in 1908 [42]. In this technology, a microlens array is used in front of the image sensor, Figure 2.6. Therefore, a captured image contains a collection of microimages each of which presents a different perspective of the scene. The number of captured perspectives is equal to the number of microlenses. As the size of the microlenses reduces, the number viewing angle increases, but the resolution of each decreases [43]. Since not all objects in the scene can be in focus simultaneously, the view images which are out-of-focus are blurred. This is called the image depth problem [44]. The major drawback of the technique is that it provides narrow viewing angles [45, 46].

The captured images can be visualised in the reverse order as the capturing mechanism, which is another type of 3D display that is not included in the previous section.

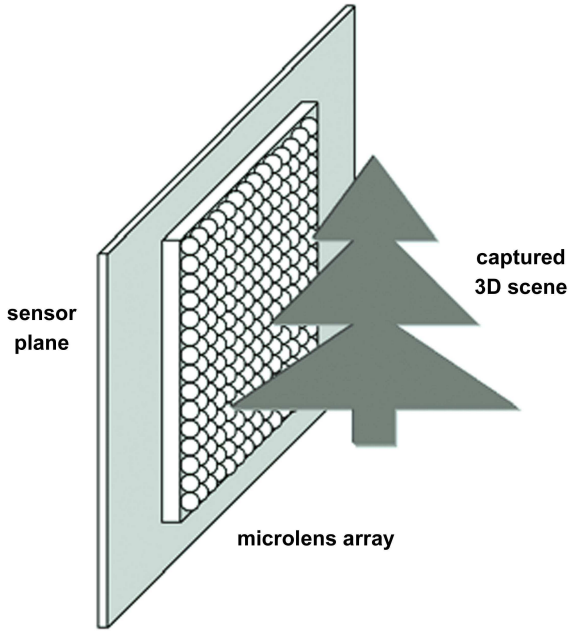


Figure 2.6: Capturing 3D images by the integral photography technique.

2.4.2 Multi view video plus depth

In this technique, different perspectives of a scene, called virtual views, can be created by processing together a 2D image and the depth map, which are already taken from a certain angle of the scene. The depth information can be obtained by specially designed range cameras or extracted from stereo pairs [47]. A sample image with its depth map is shown in Figure 2.7.

In 2007, MPEG, a working group of ISO/IEC in charge of the development of standards for coded representation of digital audio and video and related data, published ISO/IEC 23002-3 (also known as MPEG-C part 3) specification [49]. In this standard, the depth map is encoded like conventional 2D video streams. The new standard brings interoperability between different displays systems, utilization regardless of the display technology, backward compatibility with traditional 2D broadcasting and good compression performance [50].

The main disadvantage of the depth maps is that they cover only limited depth range and can not handle the occlusions [51].

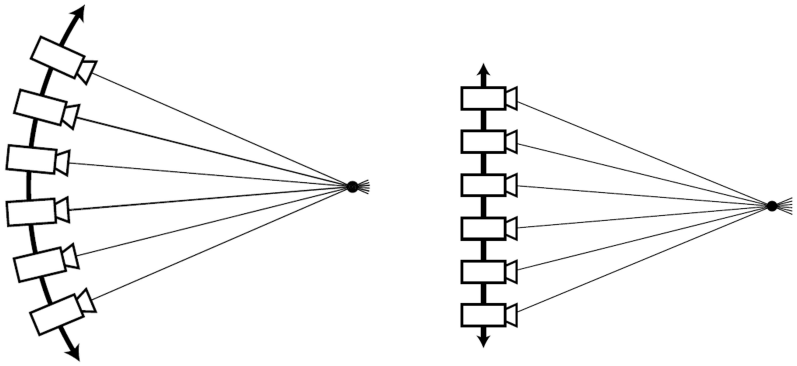
(a)

(b)

Figure 2.7: Multi view video plus depth example [48]. a) 2D image b) Depth map of the image in a.

2.4.3 Camera array

The camera arrays are suitable for recording the dynamic real-world scenes. The calibration process of the cameras is necessary in order to ensure the interoperability of the cameras [52]. The calibration method varies and becomes more challenging depending on the dimension of the array.



(a) (b)

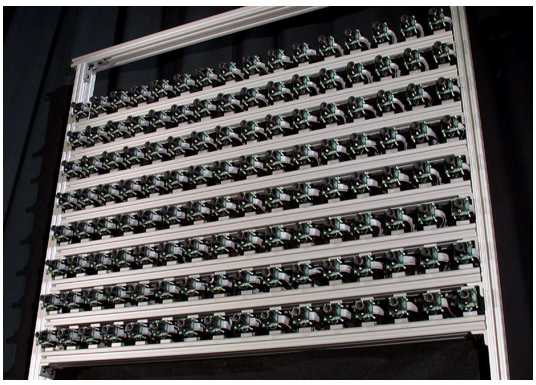
Figure 2.8: Different camera arrangements [54] a) Radial layout b) Parallel layout

Since the aligning of all cameras is a hard task, the captured view images may contain geometrical errors which needs to be considered before encoding [53].

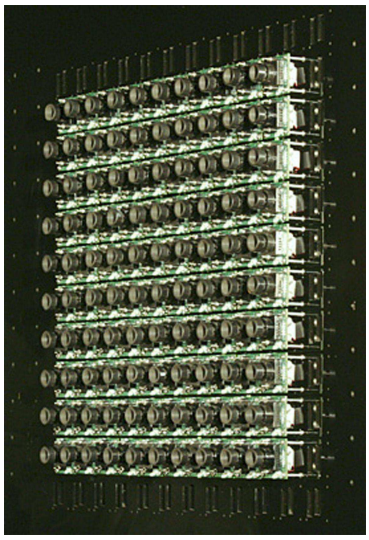
The camera arrays are divided into two groups according to their arrangements.

- **Radial Layout:** In radial layout, also called crossing or toed-in camera configuration, the cameras are arranged on a curve and their axes intersects at the same point, as shown in Figure 2.8.a. The major benefit of this layout is the presence of all the objects being captured in all the deployed cameras, which is not the case in parallel layout. Therefore, the issue that an object on the edge of one of the view images is out of view in the other view image is not the case in radially arranged cameras. On the other hand the acquired depth of the scene is limited since the view paths of the cameras are intersecting. One of the primary disadvantages of this layout is that the captured view images contain perspective distortions called keystone distortion [55].

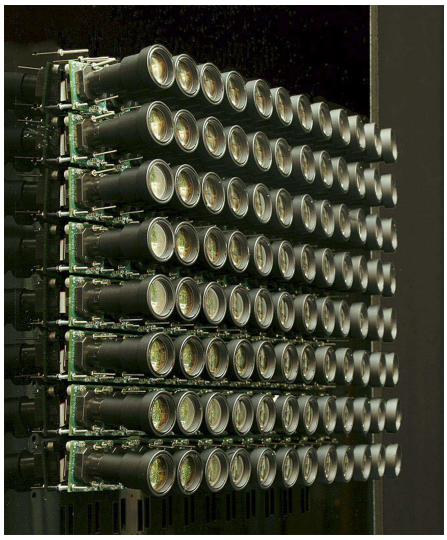
- **Parallel Layout:** In parallel layout, the optical axes of all the cameras are parallel to each other, as shown in Figure 2.8.b. This type of arrangement does not suffer from the distortion that is mentioned



(a)



(b)



(c)

Figure 2.9: Different camera array setups in parallel layout [56]. The optical axes of each camera located in this type of camera array are parallel to each other.

for the radial layout. For this reason, it is considered the correct way of capturing the multi view images [54]. In Figure 2.9 different camera setups in parallel layout are shown.

The cameras are positioned only horizontally in both arrangements in Figure 2.8. However, the arrangement can be extended through the vertical axis in order to capture vertical perspectives of a scene.

Table 2.1: Approximate uncompressed data rates of single images and movies captured by a 17x17 camera array assuming 24 bpp.

Spatial resolution	Single image	Single view movie (24 FPS.)	Multi view movie (24 FPS.) (17x17 views)
1920x1080	5.9 MB.	1139.1 Mbps.	321.2 Gbps.

One common feature of the camera arrays is that the file size of the captured image data increases proportional to the number of cameras. For example, one of the camera arrays developed by Stanford University consists of 289 cameras in a 17x17 grid and each camera can capture high resolution images like 1920 x 1080. Approximate raw data file sizes of the images and movies captured by this camera array are given in Table 2.1. Note that the audio data is excluded during this calculation. It is not feasible to broadcast such a huge amount of data and therefore it needs to be encoded. The encoding standards and more specifically the encoding of multi view images captured by 2D camera arrays will be discussed in detail in the next chapter.

References

[1] J. De Smet, A. Avci, R. Beernaert, D. Cuypers, and H. De Smet, "Design and wrinkling behavior of a contact lens with an integrated liquid crystal light modulator", *Display Technology, Journal of*, vol. 8, no. 5, pp. 299–305, 2012 (cit. on p. 9).

[2] J. De Smet, A. Avci, R. Beernaert, D. Cuypers, and H. De Smet, "Wrinkle formation in conformable liquid crystal cells for use in a contact lens display", in *IDW : proceedings of the international display workshops*, Society for Information Display (SID), 2011, pp. 1203–1206 (cit. on p. 9).

[3] J. De Smet, A. Avci, R. Beernaert, D. Cuypers, and H. De Smet, "Spherically curved guest-host display for use in a contact lens", in *Proceedings of the international display workshops*, vol. 17, Society for Information Display (SID), 2010, pp. 1585–1586 (cit. on p. 9).

[4] UCL, 2012, http://www.psychol.ucl.ac.uk (cit. on p. 10).

[5] Wikipedia, 2012, http://en.wikipedia.org (cit. on p. 10).

[6] L. Onural, T. Sikora, J. Ostermann, A. Smolic, R. Civanlar, and J. Watson, "An assessment of 3DTV technologies", in *NAB Broadcast Engineering Conference Proceedings*, Las Vegas, 2006 (cit. on p. 11).

[7] E. Lueder, *3D Displays*. Wiley, 2012 (cit. on p. 11).

[8] University of Antwerp, 2012, http://www.ua.ac.be (cit. on p. 13).

[9] Brewster-type stereoscope, 2012, http://www.nationalmediamuseum.org.uk (cit. on p. 13).

[10] Holmes-type stereoscope, 2012, http://www.nationalmediamuseum.org.uk (cit. on p. 13).

[11] Types of 3D Glasses, 2012, http://www.3dmovienews.org (cit. on p. 12).

[12] J. Konrad, B. Lacotte, and E. Dubois, "Cancellation of image crosstalk in time-sequential displays of stereoscopic video", *IEEE Transactions on Image Processing*, vol. 9, no. 5, pp. 897–908, 2000 (cit. on p. 12).

[13] S. T. Jones, S. E. Parker, and C. C. Kim, "Low-cost high-performance scientific visualization", *Computing in Science and Engineering*, vol. 3, no. 4, pp. 12–17, 2001 (cit. on p. 12).

[14] R. T. Held and T. T. Hui, "A guide to stereoscopic 3D displays in medicine", *Adademic Radiology*, vol. 18, no. 8, pp. 1035–1048, 2011 (cit. on p. 12).

[15] M. Laakso, "Stereoscopic display in a slot machine", in *Stereoscopic Displays and Applications XXIII, Proceedings of SPIE*, San Francisco, California, 2012 (cit. on p. 12).

[16] N. Hur, H. Lee, G. S. Lee, S. J. Lee, A. Gotchev, and S. I. Park, "3DTV broadcasting and distribution systems", *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 395–407, 2011 (cit. on p. 12).

[17] M. H. P. H. van Beurden, H. van Hoey, H. Hatzakis, and W. A. Ijsselsteijn, "Stereoscopic displays in medical domains: a review of perception and performance effects", in *Human Vision and Electronic Imaging XIV, Proceedings of SPIE*, San Jose, California, 2009 (cit. on p. 12).

[18] R. Borner, "Four autostereoscopic monitors on the level of industrial prototypes", *Displays*, vol. 20, no. 2, pp. 57–64, 1999 (cit. on p. 14).

[19] D. H. Kang, B. S. Oh, J. H. Oh, M. K. Park, H. J. Kim, S. M. Hong, J. H. Hur, J. Jang, S. J. Lee, K. H. Lee, and K. H. Park, "Auto-stereoscopic TFT-LCD with LC parallax barrier on wire grid polarizer", in *Digest of Technical Papers, SID International Symposium*, San Antonio, Texas, 2009 (cit. on p. 14).

[20] H. Kaneko, T. Ohshima, O. Ebina, and A. Arimoto, "Desktop autostereoscopic display using compact LED projector", in *Stereoscopic display and virtual reality systems X, Proceedings of SPIE*, Santa Clara, California, 2003 (cit. on p. 14).

[21] N. A. Dodgson, J. R. Moore, and S. R. Lang, "Multi-view autostereoscopic 3D display", *IEEE Computer*, vol. 38, pp. 31–36, 1999 (cit. on p. 14).

[22] H. Yamamoto, M. Kouno, S. Muguruma, Y. Hayasaki, Y. Nagai, Y. Shimizu, and N. Nishida, "Enlargement of viewing area of stereoscopic full-color led display by use of a parallax barrier", *Applied Optics*, vol. 41, no. 32, pp. 6907-6919, 2002 (cit. on p. 14).

[23] Y. H. Tao, Q. H. Wang, J. Gu, W. X. Zhao, and D. H. Li, "Autostereoscopic three-dimensional projector based on two parallax barriers", *Optics Letters*, vol. 34, no. 20, pp. 3220–3222, 2009 (cit. on p. 14).

[24] H. Yamamoto, S. Muguruma, T. Sato, K. Ono, Y. Hayasaki, Y. Nagai, Y. Shimizu, and N. Nishida, "Optimum parameters and viewing areas of stereoscopic full-color LED display using parallax barrier", *IEICE Transactions on Electronics*, vol. E83C, no. 10, pp. 1632–1639, 2000 (cit. on p. 14).

[25] B. Lee and J. H. Park, "Overview of 3D/2D switchable liquid crystal display technologies", in *Emerging Liquid crystal technologies V, Proceedings of SPIE*, San Francisco, California, 2010 (cit. on p. 14).

[26] D. Ezra, G. J. Woodgate, B. A. Omar, N. S. Holliman, J. Harrold, and L. S. Shapiro, "New autostereoscopic display system", 1995 (cit. on p. 14).

[27] Autostereoscopy, 2012, http://en.wikipedia.org (cit. on p. 15).

[28] B. Javidi, F. Okano, and J. Y. Son, *Three-Dimensional Imaging, Visualization, and Display*. Springer, 2008 (cit. on p. 15).

[29] C. van Berkel, D. W. Parker, and A. R. Franklin, "Multiview 3D LCD", San Jose, California, 1996 (cit. on p. 15).

[30] L. Bogaert, Y. Meuret, S. Roelandt, A. Avci, H. De Smet, and H. Thienpont, "Demonstration of a multiview projection display using decentered microlens arrays", *Optics express*, vol. 18, no. 25, pp. 26 092–26 106, 2010 (cit. on p. 15).

[31] L. Bogaert, Y. Meuret, S. Roelandt, A. Avci, H. De Smet, and H. Thienpont, "Single projector multiview displays: directional illumination compared to beam steering", in *Proceedings of the society of photo-optical instrumentation engineers (SPIE)*, vol. 7524, SPIE, the International Society for Optical Engineering, 2010, p. 10 (cit. on p. 15).

[32] Z. Megyesi, A. Barsi, and T. Balogh, "3D video visualization on the holovizio system", in *3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video*, Istanbul, Turkey, 2008 (cit. on p. 16).

[33] Y. Takaki, "High-density directional display for generating natural three-dimensional images", *Proceedings of the IEEE*, vol. 94, no. 3, pp. 654–663, 2006 (cit. on p. 16).

[34] R. Beernaert, A. Avci, J. De Smet, H. De Smet, J. De Coster, S. Severi, and A. Witvrouw, "Novel analog pulse-width-modulated 15-um sige micromirrors", *Journal of the society for information display*, vol. 18, no. 10, pp. 855–861, 2010 (cit. on p. 16).

[35] R. Beernaert, T. Podprocky, J. De Coster, A. Witvrouw, L. Haspeslagh, A. Avci, J. De Smet, and H. De Smet, "Novel micromirror design with variable pull-in voltage", *Microelectronic Engineering*, vol. 87, no. 5-8, pp. 1248–1252, 2010 (cit. on p. 16).

[36] R. Beernaert, J. De Coster, T. Podprocky, A. Witvrouw, S. Severi, A. Avci, J. De Smet, and H. De Smet, "Sige micromirrors for optical applications", in *Proceedings of SPIE, the International Society for Optical Engineering*, vol. 7750, SPIE, the International Society for Optical Engineering, 2010, p. 6 (cit. on p. 16).

[37] R. Beernaert, T. Podprocky, A. Witvrouw, L. Haspeslagh, A. Avci, J. De Smet, and H. De Smet, "Novel micromirror design with variable pull-in voltage", in *International Conference on Micro- and Nano-Engineering, 35th, Abstracts*, 2009, p. 2 (cit. on p. 16).

[38] S. A. Benton and V. M. Bove, *Holographic Imaging*. Wiley, 2007 (cit. on p. 16).

[39] P. A. Blanche, S. Tay, R. Voorakaranam, P. Saint-Hilaire, C. Christenson, T. Gu, W. Lin, D. Flores, P. Wang, M. Yamamoto, J. Thomas, R. Norwood, and N. Peyghambarian, "An updatable holographic display for 3D visualization", *Journal of Display Technology*, vol. 4, no. 4, pp. 424–430, 2008 (cit. on p. 16).

[40] Holography, 2012, http://en.wikipedia.org (cit. on p. 17).

[41] 3 Deep, 2005, http://spectrum.ieee.org (cit. on pp. 18–19).

[42] Lippmann, G., "Épreuves réversibles donnant la sensation du relief", *Journal de Physique Théorique et Appliquée*, vol. 7, no. 1, pp. 821–825, 1908 (cit. on p. 19).

[43] C. B. Burckhardt, "Optimum parameters and resolution limitation of integral photography", *Journal of the Optical Society of America*, vol. 58, no. 1, pp. 71–74, 1968 (cit. on p. 19).

[44] T. Okoshi, "Optimum design and depth resolution of lens-sheet and projection type three-dimensional displays", *Applied Optics*, vol. 10, no. 10, pp. 2284–2291, 1971 (cit. on p. 19).

[45] S. Min, B. Javidi, and B. Lee, "Enhanced three-dimensional integral imaging system by use of double display devices", *Applied Optics*, vol. 42, no. 20, pp. 4186–4195, 2003 (cit. on p. 19).

[46] S. W. Min, B. Javidi, and B. Lee, "Viewing-angle-enhanced integral imaging system using a curved lens array", *Optics Express*, vol. 12, no. 3, pp. 421–429, 2004 (cit. on p. 19).

[47] Y. K. Park, K. Jung, Y. Oh, S. Lee, J. K. Kim, G. Lee, H. Lee, N. Yun K.and Hur, and J. Kim, "Depth-image-based rendering for 3DTV service over T-DMB", *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 122–136, 2009 (cit. on p. 20).

[48] M. Dongbo, K. Donghyun, S. Y., and K. S., "2D/3D freeview video generation for 3DTV system", *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 31–48, 2009 (cit. on p. 21).

[49] I. J. 1. 2. 11, *Committee draft of ISO/IEC 23002-3 auxiliary video data representations*. WG 11 Doc. N8038. Montreux, Switzerland, 2006 (cit. on p. 20).

[50] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV", in *Stereoscopic Displays and Virtual Reality Systems XI, Proceedings of the SPIE*, San Jose, California, 2004 (cit. on p. 20).

[51] A. Vetro, A. Tourapis, K. Muller, and T. Chen, "3D-TV content storage and transmission", *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 384–394, 2011 (cit. on p. 20).

[52] T. Gandhi and M. M. Trivedi, "Calibration of a reconfigurable array of omnidirectional cameras using a moving person", in *Proceedings of the ACM 2nd international workshop on Video surveillance and sensor networks*, New York, 2004 (cit. on p. 21).

[53] Y. S. Kang and Y. S. Ho, "Geometrical compensation for multi-view video in multiple camera array", in *ELMAR, 50th International Symposium*, Zadar, Croatia, 2008 (cit. on p. 22).

[54] N. A. Dodgson, "Resampling radially captured images for perspectively correct stereoscopic display", in *SPIE Symposium on Stereoscopic Displays and Applications IX*, San Jose, California, 1998 (cit. on pp. 22–23).

[55] A. Woods, T. Docherty, and K. R., "Image distortions in stereoscopic video systems", in *Stereoscopic Displays and Applications IV, Proceedings of the SPIE*, San Jose, California, 1993 (cit. on p. 22).

[56] Stanford University, Computer Graphics Laboratory, 2012, http://lightfield.stanford.edu (cit. on p. 23).

3

Video encoding and standards

"Thinking something does not make it true.
Wanting something does not make it real."

—**Michelle Hodkin,** *The Unbecoming of Maya Dyer*

3.1 Introduction

A video is a collection of images of a scene captured at successive instances of time. Video technology finds numerous applications in our day to day life. With the rapid evolution of digital technology, video data nowadays almost invariably comes in digital format. In its raw format (uncompressed), the data file size is large and is not suitable for efficient storage and transmission. Therefore, the video data needs to be compressed (encoded).

Fundamentally, there are two types of data compression: lossless and lossy. In lossless data compression, as the name implies, the reconstructed video is identical, in all aspects, to the video that was compressed. Therefore, all information can be restored without any loss in the data. Compared to the lossless data compression, a significant file size reduction can be achieved at the cost of some visual quality in lossy data compression. Today, in most applications, the codecs are lossy. Elimination of

Table 3.1: Some typical digital video formats and their uncompressed data rates assuming 24 fps and 24 bpp.

Format	Spatial Resolution	Data Rate (Mbps)
QCIF	176x144	14.4
CIF	352x240	48.8
SDTV	720x576	239.2
HDTV (720p)	1280x720	531.2
HDTV (1080p)	1920x1080	1194.4

the excessive information results in a vast reduction of the file size of the compressed video data without noticeable degradation in the quality of the compressed file.

In this chapter, state-of-the-art lossy video coding standards and their working principles for single view video will be introduced. Then the concept of encoding multi view videos taken from a camera array will be discussed.

3.2 Video coding concept

Not more than a decade ago, analog recording video-tapes were widely used by the consumers. But, they were rapidly replaced by their digital counterparts, which bring several advantages. First of all, the duplication process is easy and unlike the copying process of analog recordings, the copies are identical to the original. Analog video tapes are vulnerable to degradation over time while digital recordings can be stored for years without any data loss. Also, video editing is much more easy in digital format than it is in analog format.

Besides recording, video broadcasting has also changed from analog to digital format over the last few years. In Table 3.1, some typical digital video formats with their uncompressed data rates are given. It is clear from this table that in the realistic case of a TV channel with a bandwidth of about 20 Mbps, uncompressed broadcasting of the raw video data is not possible except for unacceptably low resolutions.

Not only transmission but also efficient storage of video data is only

possible if the data is compressed. For example, a two hours HDV (720p) video at 24 fps and 24 bpp requires about 445 GB storage space in its raw format. This means that 18 single layer Blu-Ray Discs or 95 DVD Discs are necessary to cache this video. Nowadays, however, one of the ways for film makers to reach their customers is by means of Blu-Rays or DVDs. This has been made feasible by the invention of highly efficient compression and decompression methods. Compressing and decompressing requires computational power, which can be seen as a drawback of the digital formats. However, as a big plus, the best possible compressed video can be obtained for any given data rate.

3.2.1 Video coding standards

The process of encoding (compressing) and decoding (decompressing) a video is generally called video coding. The main goal of the video coding is to reproduce the original video after storing and/or transmission using fewer bits while maintaining its visual quality to a sufficient degree.

Since the video coding is widely used in various applications and different systems, interoperability is highly important. Therefore, standardization of encoding is indispensable to ensure seamless portability. Standardization is an evolving process and new improved standards keep replacing the outdated ones. Two highly desired features of encoding standards are forward and backward compatibility. The product designed with the new set of standards should be capable of interacting with the previous versions and vice versa. One should bear in mind that only a limited forward compatibility at best can be ensured. Forward and backward compatibility warrant convenience to the end user simultaneously imposing restrictions on the development of newer standards.

3.2.2 History of the video coding standards

There are two major standardization bodies working on the development of the video coding standards: VCEG (Video Coding Experts Group) of the ITU-T (International Telecommunication Union-Telecommunication) and MPEG (Moving Picture Experts Group) of the ISO/IEC (International Organization for Standardization/International Electrotechnical Commission).

The first digital video coding standard, called H.120 [1], was released in 1984 by ITU. This standard was primarily used in transmitting the data and the graphics during video conferences. In 1990, ITU introduced the H.261 standard [2], which is considered as the basis of modern video compression.

The aim of the H.261 standard was to transmit video over ISDN lines. The macroblock, a square group of pixels, was firstly introduced in this standard. In H.261, the hybrid-coding scheme is employed, which is still in use in recent video coding standards. According to the hybrid-coding, a frame to be encoded is predicted from the previously encoded frames based on estimating and compensating the motion. Then the residual data which contains the difference between the original and the compensated frame is transformed, quantized and coded. In early 1990s, MPEG released its first standard called MPEG-1 [3]. It was built on the H.261 standard and features a more complex encoding processes to improve the performance of the encoder. As it comes out from its title, which is "Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s", MPEG-1 was initially developed for the CD-ROMs in which the data rate is up to 1.5 Mbps. Afterwards MPEG-2 [4], still the most frequently used video coding standard, was released. Since it was jointly developed by both MPEG and ITU standard bodies, MPEG-2 is also known as H.262. It is capable of handling higher data rates such as digital TV signals. Unlike MPEG-1, which supports only non-interlaced video, MPEG-2 can support both interlaced [1] and progressive [2] videos. Since it is primarily targeted at high bitrates, the MPEG-2 standard is less suitable for video streaming applications. Nowadays, this standard is widely used in digital TVs, DVDs and SVCDs. Although the MPEG-3 standard was developed for HDTV signals, it was later integrated into MPEG-2 since it was realized that MPEG-2 can also accommodate HDTV. On the other hand, ITU released the H.263 standard in 1992 [5]. It was developed mainly for low-bitrate communication systems over the internet such as video-conferencing and telephony. Most of the video sharing websites use this standard. H.264, also known as AVC (Advanced Video Coding, MPEG-4 Part 10), was released in 2003 [6]. H.264/AVC is the most commonly used format for high definition video like in Blu-ray Discs. H.264/AVC outperforms all prior video coding standards due to its significantly improved coding efficiency [7]. Detailed working principles of the H.264/AVC video coding standard are elaborated in Section 3.4.

[1] Progressive video: All horizontal lines in a video frame are sequentially scanned onto the display. This technique is widely used in computer monitors and digital televisions.

[2] Interlaced video: Either odd or even numbered lines of a video frame are scanned onto the display alternatingly at a time instant. An advantage of interlaced scan is that a high refresh rate (50 or 60 Hz) can be achieved by using only half the bandwidth. While this technique has been very useful in the era of cathode ray tubes (CRTs), it is not directly compatible with flat-panel displays and a filter needs to be applied on the video to eliminate artefacts such as 'combing'.

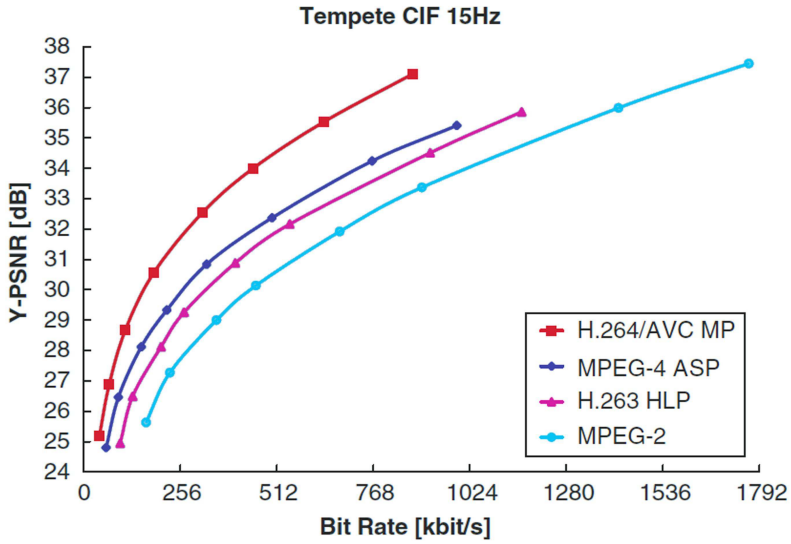


Figure 3.1: Rate-Distortion curves of *Tempete* test sequence generated by different video coding standards.[7].

To make a comparison, the so-called 'rate-distortion' (RD) curves of the video coding standards are plotted in Figure 3.1. While the abscissa indicates the bitrate, the ordinate shows the distortion obtained for a specific rate correspondingly. Peak Signal to Noise Ratio (PSNR) is the typical measure used to represent compressed image quality and will be explained in Section 3.3.2. Figure 3.1 clearly shows that the encoded test sequence has a lower bitrate and a higher quality when encoded with the H.264/AVC video coding standard.

3.3 Basics of Video Coding

In this section, some important processes and features of the H.264/AVC video coding standard will be outlined. The concepts discussed in this section form the foundation on which the rest of ideas discussed in this dissertation are based.

3.3.1 Color space

In display systems the color information is mostly represented by means of numbers. For example, only one number which gives the luminance or

brightness level is enough for a pixel in a monochrome image. However, in the case of colored images, at least three numbers are required to reconstruct a full-color[3] pixel. The value of these three numbers are subject to change depending on the selection of the color space.

Red, Green and Blue are the basic set of colors in the most used RGB color space. Different colors of a pixel can be realized by linear combination of individual color components. However, the human visual system is more sensitive to the luminance (brightness) rather than the color. Remembering the fact that the ultimate aim is to compress the data, the RGB color space can be converted into another color space (YCbCr color space) where the luminance information is represented with higher resolution than the color information. The component Y, representing the luminance or luma, is calculated as:

$$Y = k_r R + k_g G + k_b B \tag{3.1}$$

where k_r, k_g and k_b are the red, green and blue weighting coefficients respectively and:

$$\begin{aligned} k_r + k_g + k_b &= 1 \\ k_r, k_g, k_b &\geq 0 \end{aligned} \tag{3.2}$$

Cb, Cr and Cg are the color differences, called chrominance, between the Y value and each color component in the RGB color space, which can be formulated as ;

$$\begin{aligned} Cb &= B - Y \\ Cr &= R - Y \\ Cg &= G - Y \end{aligned} \tag{3.3}$$

Since any one of the chrominance values can be calculated by means of the other two, one of them is unnecessary to regenerate the color value of a pixel. Therefore, only the blue and red chrominance values are included in the YCbCr color space.

The k coefficients are defined by the standards [8] as follows:

[3]With a full-color pixel I mean a pixel that can represent any color in the color space I am working with.

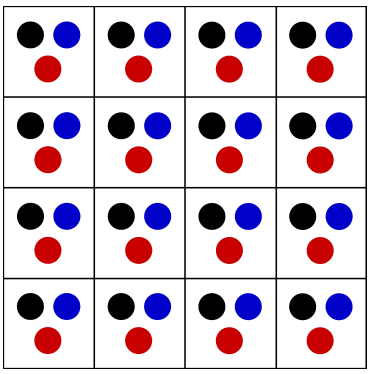


Figure 3.2: 4:4:4 color sub-sampling for a 4x4 pixel group. Black, blue and red circles represent the Y, C_b and C_r components respectively.

$$\begin{aligned} k_b &= 0.114 \\ k_r &= 0.299 \\ k_g &= 0.587 \end{aligned} \tag{3.4}$$

There are three sampling formats that are employed by H.264/AVC, where the luminance color component is represented with equal or higher resolution than the chrominance components based on the sensitivity of the human visual system.

- **4:4:4 sampling format:** In this sampling format all three components have equal resolution as shown in Figure 3.2. As Y, Cb and Cr color components are explicitly specified for each pixel, this format requires the same number of bits as required by the RGB color format.

- **4:2:2 sampling format:** In 4:2:2 sampling format, two Cb and Cr bits are present for four luminance bits as drawn in Figure 3.3. This format is frequently referred as YUY2 and is used for high quality color reproduction.

- **4:2:0 sampling format:** This very popular sampling format consists of only one Cb and Cr components. One of the possible ways for 4:2:0 sampling is demonstrated in Figure 3.4. Since each chroma component constitutes one-fourth of the luminance component, the

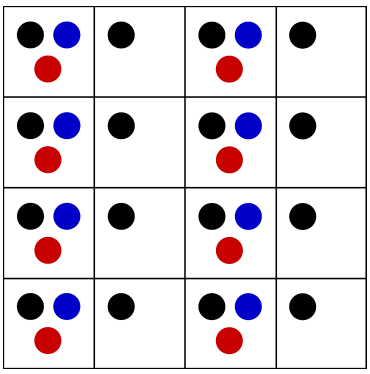


Figure 3.3: 4:2:2 color sub-sampling for a 4x4 pixel group.

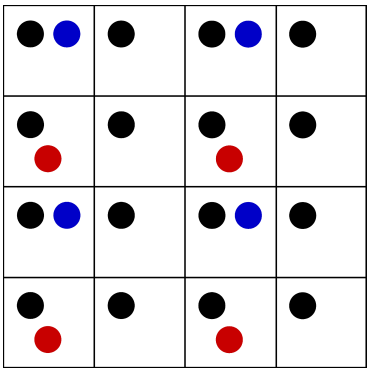


Figure 3.4: 4:2:0 color sub-sampling for a 4x4 pixel group.

total number of bits required is exactly half of what is used in 4:4:4 sampling format or in the RGB format.

3.3.2 Video Quality Measurement

Since H.264/AVC is a standard for lossy video encoding, quality evaluation of the encoded videos that are displayed to the viewers is very important. There are basically two types of video quality measurement methods.

- **Subjective quality measurement:** In subjective assessment, a group of people with difference perception of quality, is asked to review the content. In order to have a reliable assessment, a large number of users needs to be involved, which makes it costly and time consuming. Test conditions and procedures have been defined by ITU such

as the parameters of the experiment, ambient light in the test room, the features of the display, the viewing distance and the method to analyse the data [9].

- **Objective quality measurement:** The quality of the video is measured by using algorithms in objective quality measurement. On of the most widely used objective quality measurement techniques is PSNR (Peak Signal Noise Ratio).

  PSNR represents the distortion of the video at the pixel level on a logarithmic scale. In its calculation, the sequence Y_{ij} whose distortion is to be measured is compared with the reference sequence X_{ij}, resulting in the mean squared error (MSE):

$$MSE = \frac{1}{MN} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (X_{ij} - Y_{ij})^2 \tag{3.5}$$

  where M, N refers to the frame dimensions and i, j shows the pixel coordinates.

  PSNR then follows from:

$$PSNR_{dB} = 10log_{10} \frac{(2^n-1)^2}{MSE} \tag{3.6}$$

  where n is the number of bits used in the image.

  The structural similarity index (SSIM) is another widely used method to compare the similarity between two images [10]. Since the human visual system is taken into account, SSIM generates more reliable results compared to the PSNR. The SSIM measures the change in luminance ($l(x, y)$), contrast ($c(x, y)$) and the structure ($s(x, y)$) in an image with the following equation.

$$\begin{aligned} SSIMx, y &= l(x, y).c(x, y).s(x, y) \\ &= \frac{2\mu_x\mu_y}{\mu_x^2 + \mu_y^2} . \frac{2\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2} . \frac{2\sigma_{xy}}{\sigma_x\sigma_y}, \end{aligned} \tag{3.7}$$

  where μ and σ represent the mean and the standard deviation respectively while $\sigma_{x,y}$ gives the sample cross correlation of x, y.

Although the objective quality assessment is less reliable than its subjective counterpart, it is more practical since the subjective quality measurement is costly and time consuming.

3.4 H.264/AVC video coding standard

The H.264/AVC video coding standard is jointly developed by experts from VCEG and MPEG. This joint team is named JVT. The H.264/AVC video coding standard brings new features yielding high compression performance which is typically half the bitrate at the same quality compared to the other video coding standards.

A typical H.264/AVC encoder block diagram is shown in Figure 3.5. In this section, all main procedures drawn in Figure 3.5 will be explained step by step.

3.4.1 Frame types

The aim of video coding is to achieve a high compression ratio by eliminating redundancy existing in the raw video. There are mainly two types of redundancies,

- **Spatial redundancy:** Spatial redundancy refers to the correlation of the neighbouring pixels within a frame. Intra coding was introduced to remove this type of redundancy. The spatial redundancy can further be eliminated by using transform techniques.
- **Temporal redundancy:** A video is a collection of consecutive frames which contain high correlation at sufficient frame rate. Video can be compressed by removing such temporal redundancy and the coding technique which exploits this redundancy is given the name Inter coding. Temporal redundancy can be eliminated by motion estimation and compensation techniques.

Since H.264/AVC is a hybrid type of encoder, it is aimed to remove both temporal and spatial redundancy existing in the video. For this purpose there are three frame types available in the H.264 video coding standard.

- **I (Intra-coded) Frame:** The I frame is a frame which will be coded independently from the other frames, thus the decoder does not need any extra information to decode the I frame. Since the temporal

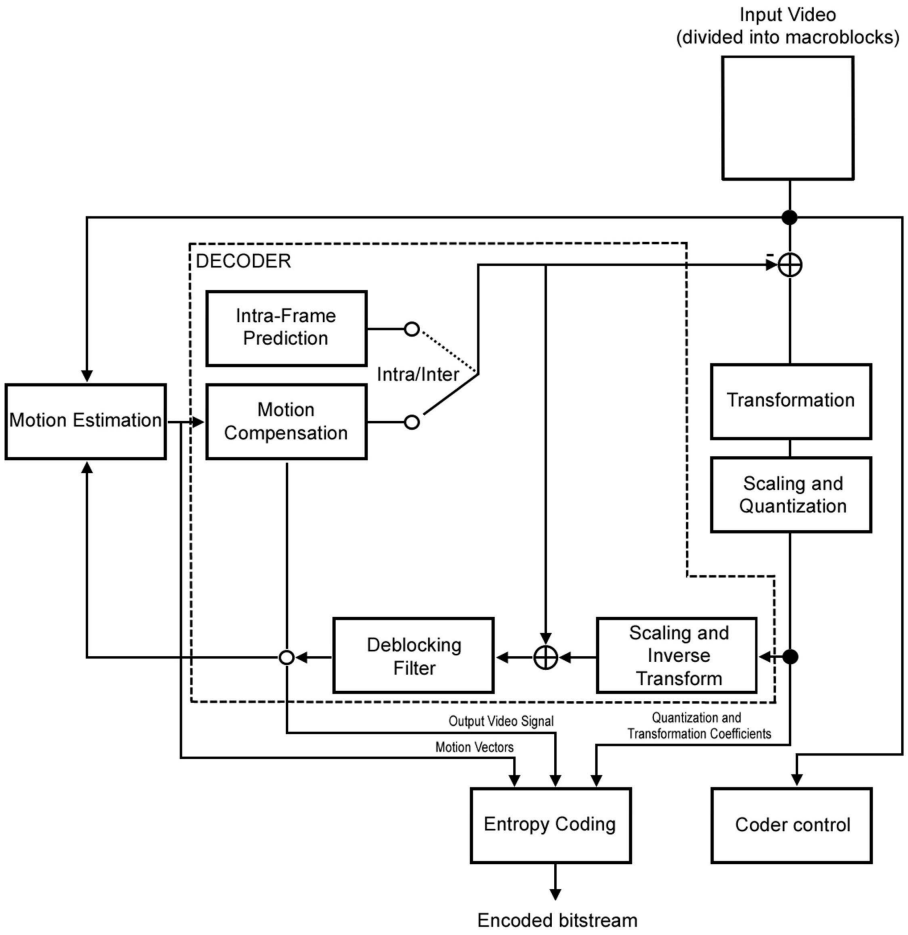


Figure 3.5: H.264/AVC encoder block diagram.

redundancy is not exploited in this frame type, an I frame always requires more bits than the other frame types.

A video sequence always starts with an I frame. The number of I frames in entire video sequence is completely dependent on the encoder configuration. Employing more I frames in an entire sequence ensures that the video can more easily be recovered in case of damage in the bitstream. On the other hand, however, it comes at the expense of an increase in the file size of the encoded bitstream. I frames enable us to implement fast forward, rewind or

random access to any part of the video.

- **P (Predicted) Frame:** Since the frames in a video sequence often contain repetitive information, a high compression ratio can be achieved by finding the corresponding parts (motion) of a frame in another frame. Once a good match is found for a block, that block can then be represented by a motion vector which points to the best match in the other frame and the transform coefficients (see Section 3.4.6) which give the predicted correction. The best case is that the encoder finds the exact copy of the candidate block in its reference frame. In this situation, the block is represented by a single motion vector that can be stored in much fewer bits than the image information of the block.

  P frames can take previous I or other P frames as a reference. Typically a frame which is encoded as a P frame requires fewer bits than an I frame does. The major drawback of the P frame is that the encoder needs to run the motion estimation process for each macroblock and its partitions, which is a highly time consuming and complex process for the encoder.

- **B (Bi-Predictive) Frame:** Structurally, a B frame is very similar to a P frame. The main difference is that B frames can take not only the previous but also the future I or P frames (in display order) as a reference. This increases the probability of finding a good match, thus improving the compression performance of the encoder. However, since the motion estimation process needs to be run for each reference frame, the complexity of the B frame is very high compared to I or P frames.

A typical encoding sequence of a video is given in Figure 3.6. According to this configuration, every ninth frame will be coded as an I frame and P and B frames are sandwiched in between these I frames. Each subencoding sequence is called a group of pictures (GOP). After encoding the first I frame, the P frame which is indicated by the arrow in the figure is to be coded. Since the B frame takes two reference frames, in this case the I and P frame (shown by arrows), these frames need to be coded before the coding for the B frame can start. The encoding process continues with the next B frame and then the other P frame in the GOP. The number of B or P frames in between two I frames can be increased optionally.

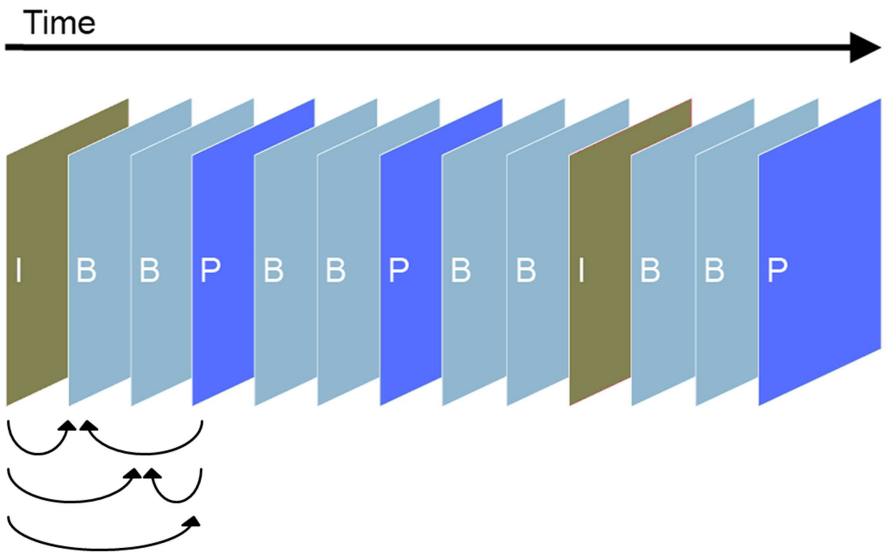


Figure 3.6: A typical GOP sequence.

3.4.2 Macroblocks and partitions

Since H.264/AVC is a block based video coding standard, an entire frame is divided into non-overlapping square groups of pixels called macroblocks. In the H.264/AVC video coding standard, the macroblock size is fixed to 16x16 pixels. In order to increase the coding efficiency, each macroblock can further be divided into sub-blocks, which is one of the unique features of H.264/AVC. Each division is called submacroblock or partition. In the H.264/AVC video coding standard, possible partitions of a 16x16 luma macroblock can be 16x16, 16x8, 8x16 and 8x8. On the other hand the sub-macroblock partitions can be 8x8, 8x4, 4x8 and 4x4 as drawn in Figure 3.7.

3.4.3 Motion estimation

If the frame is to be inter coded (either P or B frame), the motion estimation process needs to be run for each macroblock in the frame. The aim of the motion estimation process is to find the best possible match for a macroblock. In order to achieve this, a block is searched in a window of predefined size called search window, in the reference frame as shown in Figure 3.8. A larger search window on the one hand increases the

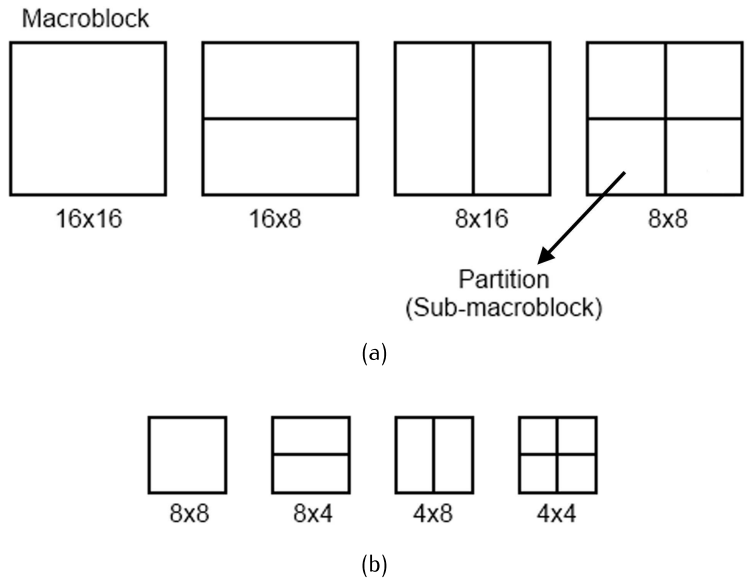


Figure 3.7: The macroblock and possible partitions. a) Possible partitions of 16x16 macroblock b) Possible partitions of 8x8 sub-macroblock.

probability of finding a better match, but on the other hand adversely affects the required number of calculations, also called the complexity.

The motion estimation process is the most complex and time consuming step of the encoding procedure. To reduce this burden, many different searching methods have been developed and are present in the H.264/AVC video coding standard [11–16]. These methods are aiming to find the best match without trying all the possible positions in the search window, thus reducing the complexity of the encoder by decreasing the number of evaluations.

One of novelties of the H.264/AVC video coding standard is that the motion estimation can be completed at ¼ pixel accuracy, which is known as *qpel motion estimation*. This enables the encoder to search finely in the proximity of the closest match in order to enhance the quality of estimation.

3.4.4 Motion compensation

Once the motion estimation process is realized, the current block is subtracted from the best matching block found in the reference frame, to generate the *residual block*. If the best match is identical to the current block,

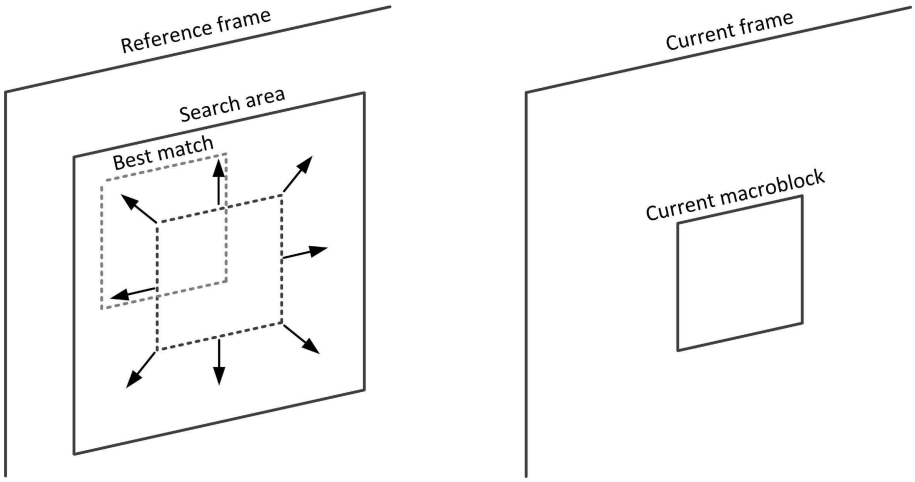


Figure 3.8: Block searching mechanism in motion estimation [17].

the residual information would contain no information. This case results in the best compression, since only the motion vector needs to transmitted for that particular block. If the encoder cannot find a good temporal match, then the block will be coded as an intra macroblock, which is less efficient. Therefore the compression efficiency of the encoder is directly dependent on the efficacy of the motion vectors.

3.4.5 Transform

The outcome of the motion estimation process is the motion vector information and the corresponding residual data. In the transformation process, the residual data is transformed to the frequency domain, where repetitive information in the residual data is removed by performing a frequency analysis.

Typical block-based video coding standard before H.264/AVC used an 8x8 block size for the transformation while H.264/AVC allows both 4x4 and 8x8 transformations, which helps to reduce ringing effects [4] in the encoded bitstream. The transformation process is based on DCT (Discrete Cosine Transformation) with some extra features. One of the unique features of the H.264/AVC video coding standard is its use of integer transforms where only integer arithmetic is used instead of floating point arithmetic, thereby

[4]Ringing effect is an artefact that appears as bands near the edges of objects in an image [18].

reducing drift to the calculations. The integer transformation has almost the same compression performance as the floating point DCT.

The forward integer transform of a 4x4 residual data block X is

$$Y = (C\ X\ C^T) \otimes E \tag{3.8}$$

where C is the 4x4 transformation matrix:

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \tag{3.9}$$

and CXC^T is called the core 2D transform, $\otimes$ represents the Hadamard matrix multiplication [5] and E is the weighting matrix which is described as

$$E = \begin{bmatrix} a^2 & \frac{ab}{2} & a^2 & \frac{ab}{2} \\ \frac{ab}{2} & \frac{b^2}{2} & \frac{ab}{2} & b^2 \\ a^2 & \frac{ab}{2} & a^2 & \frac{ab}{2} \\ \frac{ab}{2} & \frac{b^2}{2} & \frac{ab}{2} & b^2 \end{bmatrix} \tag{3.10}$$

where $a = \frac{1}{2}, b = \sqrt{\frac{1}{2}}.cos(\frac{\pi}{8})$ and $c = \sqrt{\frac{1}{2}}.cos(\frac{3\pi}{8})$.

The inverse integer transform can be calculated as,

$$X' = C^T\ Y'\ C. \tag{3.11}$$

[5] In Hadamard matrix multiplication, the elements of the two same sized matrices are multiplied each other element by element as follows:

$$A \otimes B = \begin{bmatrix} A_{11} & A_{12} & . & . & . & A_{1n} \\ A_{21} & A_{22} & . & . & . & A_{2n} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ A_{m1} & A_{m2} & . & . & . & A_{mn} \end{bmatrix} \otimes \begin{bmatrix} B_{11} & B_{12} & . & . & . & B_{1n} \\ B_{21} & B_{22} & . & . & . & B_{2n} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ B_{m1} & B_{m2} & . & . & . & B_{mn} \end{bmatrix}$$

$$= \begin{bmatrix} A_{11}B_{11} & A_{12}B_{12} & . & . & . & A_{1n}B_{1n} \\ A_{21}B_{21} & A_{22}B_{22} & . & . & . & A_{2n}B_{2n} \\ . & . & . & & & . \\ . & . & & . & & . \\ . & . & & & . & . \\ A_{m1}B_{m1} & A_{m2}B_{m2} & . & . & . & A_{mn}B_{mn} \end{bmatrix}$$

Table 3.2: Quantization step sizes in H.264/AVC video coding standard.

QP	0	1	2	3	4	5
Qstep	0.625	0.6875	0.8125	0.875	1	1.125
QP	6	7	8	9	10	11
Qstep	1.25	1.375	1.625	1.75	2	2.25
QP	12	...	20	...	26	...
Qstep	2.5	...	6.5	...	13	...
QP	31	...	48	...	51	...
Qstep	22	...	160	...	224	...

3.4.6 Quantization

After the transform a quantization process takes place, which is executed on the Y_{ij} values (cf. Eq. 3.8). The parameter which determines the level of the quantization is called quantization parameter (QP) which is an integer number ranging from 0 to 51.

The H.264/AVC video coding standard uses a predefined scalar quantizer called Q_{step}, which is directly related to the QP. Therefore the number of Q_{step} values is the same as the number of QP values. Q_{step} can directly be derived from the QP as shown in Table 3.2. Basically the value of the Q_{step} doubles for every increment of 6 in QP.

The basic forward quantizer equation is

$$W_{ij} = round(\frac{Y_{ij}}{Q_{step}}), \tag{3.12}$$

where Y_{ij} is a transform coefficient and i,j indicates the position, Q_{step} is the quantization step size and W_{ij} is the quantized coefficient.

The simplified and preferred way of applying the forward quantization in the reference model software [19] is

$$\begin{cases} |Z_{ij}| = (|W_{ij}|.MF + f) >> qbits \\ sign(Z_{ij}) = sign(W_{ij}) \end{cases}, \tag{3.13}$$

and

$$qbits = 15 + floor(\frac{QP}{6}) \tag{3.14}$$

where MF is multiplication factor, $>>$ denotes a binary right shift and

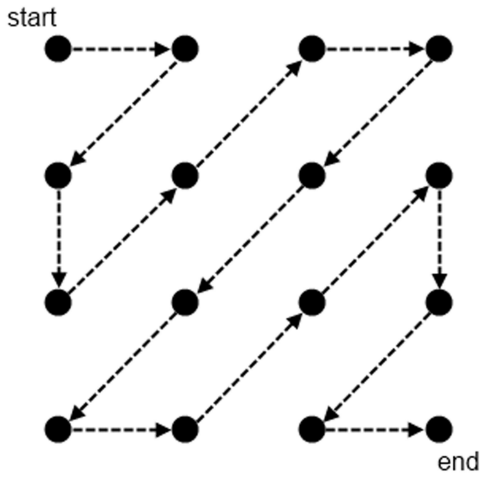


Figure 3.9: Zigzag scan for a 4x4 block.

f is equal to either $2^{qbits}/3$ or $2^{qbits}/6$ depending if the block is Intra or Inter coded respectively.

The equation used for the inverse quantizer is

$$Y'_{ij} = Z_{ij}.Qstep \tag{3.15}$$

.

3.4.7 Entropy Coding

Entropy coding is a lossless data compression method. After transformation and quantization, significant low frequency coefficients are typically accumulated on the left top part of the block. As shown in Figure 3.9, these coefficients are scanned in a zigzag fashion in order to maximize the number of consecutive zeros in the value chain, which increases the efficiency of the entropy coding.

In the H.264/AVC video coding standard two different entropy coding methods are provided.

- **Context Adaptive Variable Length Coding (CAVLC) method:** In this method, data is compressed by looking at the probability of occurrence of the elements for which different variable length codes are assigned. Shorter codes are chosen for more frequent elements

in the data and the length of the code is inversely proportioned to the frequency of the element.

- **Context Adaptive Binary Arithmetic Coding (CABAC) method:** The CABAC method achieves high redundancy reduction since it combines the adaptive binary arithmetic coding method with context modelling. In the CABAC method, the non-binary elements are converted into their binary equivalents [6], which increases the probability of their occurrences and keeps the complexity of the method low. The probability model for each bit is determined and updated based on previous coding statistics. The CABAC shows 9%-14% better compression efficiency over CAVLC at the cost of higher complexity [20]

3.4.8 De-blocking filter

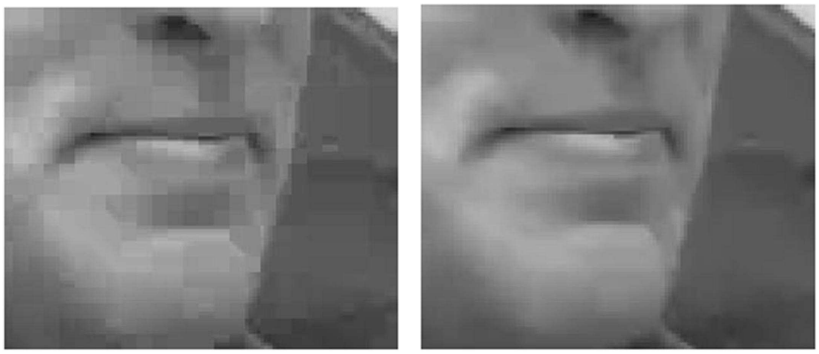
Encoded pictures can contain some "blocky" artefacts, especially at the lower bitrates, due to the transformation process. The de-blocking filter is applied only on the edges of the transform blocks in each picture. With the de-blocking filter, significant subjective and objective quality improvement can be achieved [21]. An example of an encoded picture with and without a de-blocking filter is shown in Figure 3.10.

3.4.9 Rate Control

The bitrate of the compressed video can vary due to several facts. First of all the resulting bitrate depends on the specified encoding parameters. If they are kept the same for different videos containing different contents, the output rate naturally varies. Moreover, the bitrate is dependent on the frame types that are defined in the prediction scheme. Obviously, the bitrate of an encoded video in which all frames are configured to be coded as I frame will be higher than the one in which the typical I-B-B-P configuration is used.

Since most of the time the network bandwidth and the buffer capacities of the systems are limited, variation in the bitrate is undesired. Therefore,

[6] Binarization: It is the name of the process where all the syntax elements such as transformation coefficients and motion vector information are converted into their binary equivalents.



(a) (b)

Figure 3.10: Result of adaptive de-blocking filter [21] a) without filter b) with filter

there is a the rate control process which ensures the encoded bitstream will satisfy the bandwidth and buffering constraints.

In the H.264/AVC video coding standard, the rate control mechanism plays an important role although its operation is not specified in the standard. The output rate is maintained by changing the encoding parameters, mainly QP. The task of the rate control mechanism is to determine a suitable quantization step size for each macroblock, with which the encoder fulfils the required constraints. There are many approaches to achieve this goal [22–26], among which the Lagrangian bit-allocation techniques are most widely used [27].

3.4.10 Mode Decision and Rate-Distortion Optimization

As mentioned before, the H.264/AVC video coding standard outperforms any of the other prior video coding standards. Among the main contributions to its success are seven different motion-compensation block sizes, two intra prediction modes and a SKIP mode which indicates that the block can be represented by a motion vector without having any residual data. The encoder must decide the best macroblock mode which yields the maximum coding efficiency for that macroblock.

In the H.264/AVC video coding standard, the Lagrangian Rate-Distortion Optimization (RDO) method is used [27]. The general formula of the RDO employed in H.264/AVC is

$$RDcost_{mode} = D + \lambda_{mode}.R \tag{3.16}$$

where the $RDcost$ is the rate-distortion cost, λ_{mode} is the Lagrange multiplier, D is the distortion of the reconstructed information and R indicates the number of bits required to store all necessary information such as the integer transformed residual data, motion vectors, macroblock mode and QP. In the reference software [19] the λ_{mode} is defined as a function of the QP:

$$\lambda_{mode}(QP) = 0.85 \, . \, 2^{\frac{(QP-12)}{3}} \tag{3.17}$$

The mode decision algorithm runs step by step as follows,

1. The encoder calculates the RD cost value of the current block in SKIP mode and sets the best mode to SKIP and the best RD cost to the calculated RD cost.

2. For the every other mode, the encoder runs the motion estimation process to find a suitable motion vector which is pointing to the best possible match among all possible matches and yielding a lower RD cost value than that from the SKIP mode in step 1.

3. The mode which produces the minimum RD cost value among the possible modes is selected as the best mode, which will be used to encode that particular block.

As can be seen from the algorithm steps, the encoder needs to calculate and compare the RD cost values of all possible modes. This process is highly efficient but complex and time consuming.

References

[1] CCITT Recommendation H.120, Codecs for videoconferencing using primary digital group transmission, CCITT (currently ITU-T) , Geneva, 1989. (cit. on p. 33).

[2] CCITT Recommendation H.261, Video codec for audiovisual services at p x 64kb/s, CCITT (currently ITU-T) , Geneva, 1990. (cit. on p. 33).

[3] ISO/IEC 11172-2, Information Technology - Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1,5 Mbit/s-video, Geneva, 1993. (cit. on p. 34).

[4] ISO/IEC IS 13818-2, Generic coding of moving pictures and associated audio, 1994. (cit. on p. 34).

[5] Draft ITU-T Recommendation H.263, Video coding for low bitrate communication, 1995. (cit. on p. 34).

[6] ITU-T Rec.H.264|ISO/IEC IS 14496-10, Advanced Video Coding Generic Audiovisual Services, 5th edition, 2010. (cit. on p. 34).

[7] J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video coding with H.264/AVC: tools, performance, and complexity", *Circuits and Systems Magazine,IEEE*, vol. 4, no. 1, pp. 7–28, 2004 (cit. on pp. 34–35).

[8] Recommendation ITU-R BT.601-5, Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios, ITU-T, 1995. (cit. on p. 36).

[9] Recommendation ITU-T BT.500-11, Methodology for the subjective assessment of the quality of television pictures, ITU-T, 2002. (cit. on p. 39).

[10] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity", *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600–612, 2004 (cit. on p. 39).

[11] L. M. Po and W. C. Ma, "A novel four-step search algorithm for fast block motion estimation", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 6, no. 3, pp. 313–317, 1996 (cit. on p. 44).

[12] B. Liu and A. Zaccarin, "New fast algorithms for the estimation of block motion vectors", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 3, no. 2, pp. 148–157, 1993 (cit. on p. 44).

[13] C. Zhu, X. Lin, and L. P. Chau, "Hexagon-based search pattern for fast block motion estimation", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 5, pp. 349–355, 2002 (cit. on p. 44).

[14] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding", *Image Processing, IEEE Transactions on*, vol. 9, no. 3, pp. 497–501, 2000 (cit. on p. 44).

[15] J. Chalidabhongse and C.-C. Kuo, "Fast motion vector estimation using multiresolution-spatio-temporal correlations", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 7, no. 3, pp. 477–488, 1997 (cit. on p. 44).

[16] C. H. Cheung and L. M. Po, "A novel cross-diamond search algorithm for fast block motion estimation", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 12, no. 12, pp. 1168–1177, 2002 (cit. on p. 44).

[17] E. Iain and G. Richardson, *H.264 and MPEG-4 Video Compression: Video Coding for Next Generation Multimedia*. Wiley, 2003 (cit. on p. 45).

[18] B. Juurlink, M. Alvarez-Mesa, C. C. Chi, A. Azevedo, C. Meenderinck, and A. Ramirez, *Scalable Parallel Programming Applied to H.264/AVC Decoding*. Springer, 2012 (cit. on p. 45).

[19] Joint Video Team, H.264 Reference Software, 2012, http://iphome.hhi.de/suehring/tml/index.htm (cit. on pp. 47, 51).

[20] D. Marpe, H. Schwarz, G. Blättermann, G. Heising, and T. Wieg, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 620–636, 2003 (cit. on p. 49).

[21] P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, "Adaptive deblocking filter", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 614–619, 2003 (cit. on pp. 49–50).

[22] T. 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, 1997 (cit. on p. 50).

[23] H. Hamdi, J. W. Roberts, and P. Rolin, "Rate control for VBR video coders in broad-band networks", *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 6, pp. 1040–1051, 1997 (cit. on p. 50).

[24] S. Ma, W. Gao, and Y. Lu, "Rate-distortion analysis for H.264/AVC video coding and its application to rate control", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 12, pp. 1533–1544, 2005 (cit. on p. 50).

[25] D. K. Kwon, M. Y. Shen, and C. C. J. Kuo, "Rate control for H.264 video with enhanced rate and distortion models", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 5, pp. 517–529, 2007 (cit. on p. 50).

[26] Y. Liu, Z. G. Li, and Y. C. Soh, "A novel rate control scheme for low delay video communication of H.264/AVC standard", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 1, pp. 68–78, 2007 (cit. on p. 50).

[27] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688–703, 2003 (cit. on p. 50).

4

Multiview video coding

"Almost everything you do will seem insignificant, but it is important that you do it."

—**Mahatma Gandhi**

4.1 Introduction

Nowadays, 3D applications are booming thanks to the simultaneous advancements in different research domains such as displays and cameras. It is obvious that there is a long process chain from the data acquisition to the display where the captured scene is reconstructed in 3D. In this long process chain, the multiview video coding is one of the most important steps.

Most of the time, the 3D displays require different views captured from different perspectives of the scene to reconstruct them in 3D. The video containing different views of a scene is called multiview video (MVV). The realism of the reconstructed 3D is closely related to the number of different views taken around the scene. However, this poses a serious problem of storing and transmitting a substantial amount of data to the receiver side. In order to prevent this from happening, the MVV needs

to be encoded which is called multi-view video coding (MVC). Since all individual video sequences of the MVV are taken from the same scene, they contain strong correlations. High compression performance can be reached if these similarities are exploited efficiently.

In this chapter, details are given on the principles of the multiview video coding extension of the H.264/AVC standard. Since it is an extension, only the differences with respect to the H.264*/AVC standard will be outlined.

4.2 Multiview video coding concept

An MVV contains a multiple of video sequences that are captured from different viewpoints of a scene. As it has already been mentioned in Section 2.4, such an MVV can be obtained in different ways and one of the most common ways is to use a camera array. Due to the increased number of views, the MVV contains an enormous amount of data compared to a typical single view video. This prohibits MVV to be stored or transmitted over networks including broadcasting. Therefore it needs to be compressed.

MVC is a key technology for applications such as free-viewpoint video, free-viewpoint television, immersive teleconferencing and 3D displays. A typical model is depicted in Figure 4.1. The MVC encoder receives multiple videos captured from different perspectives of a scene as an input and generates one encoded bitstream. This bitstream is transmitted to the receiver side that can be any sort of display such as a conventional 2D TV, a stereoscopic display or a multiview display. At the decoder side, the bitstream is decoded and the views are regenerated. Note that the bitstream containing the MVV should be decodable by a wide range of displays. At this point, the MVC standard plays an important role, which is to have backward compatibility (cf. Section 3.2.1).

4.3 Disparity estimation

The disparity is the displacement between the projection points of an object on image planes captured from a different view point.

Algorithmically, the disparity estimation process is the same as the motion estimation process. The only difference is that the candidate block is searched into the frames in the view domain instead of the frames in the temporal domain which is the case in motion estimation. The vector pointing to the to the displaced version of the candidate block is called the disparity vector.

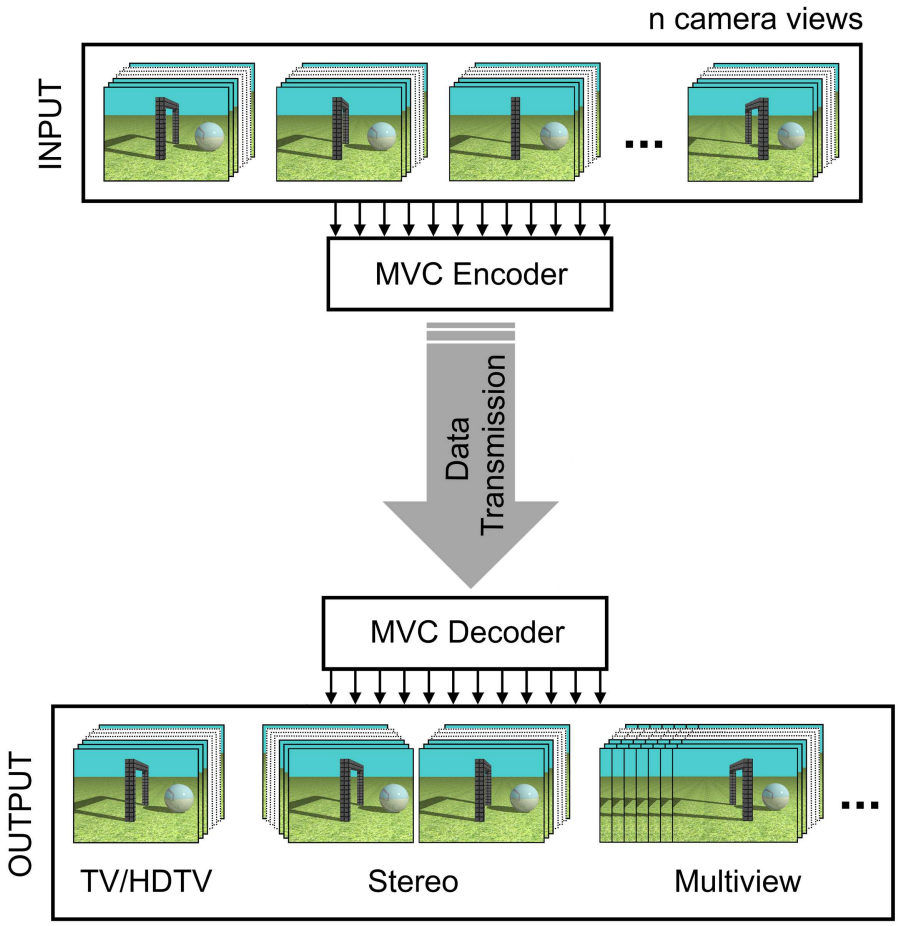


Figure 4.1: Typical multiview video coding model.

In a 3D scene, one object can hide another, more remote object, for a certain camera position. These so-called occlusions cause big problems for the matching process while the disparity vector is being estimated. Since there is no good reference block that can be found for these occluded areas, they can be poorly compressed and cause an increase in size of the encoded file. On the other hand, searching a reference block for an occluded block is an important issue. While one view may not have any good reference for the occluded block, any of the other views may contain the occluded block information. Moreover, in some cases, this information can be found in the same view but in another frame in the temporal domain.

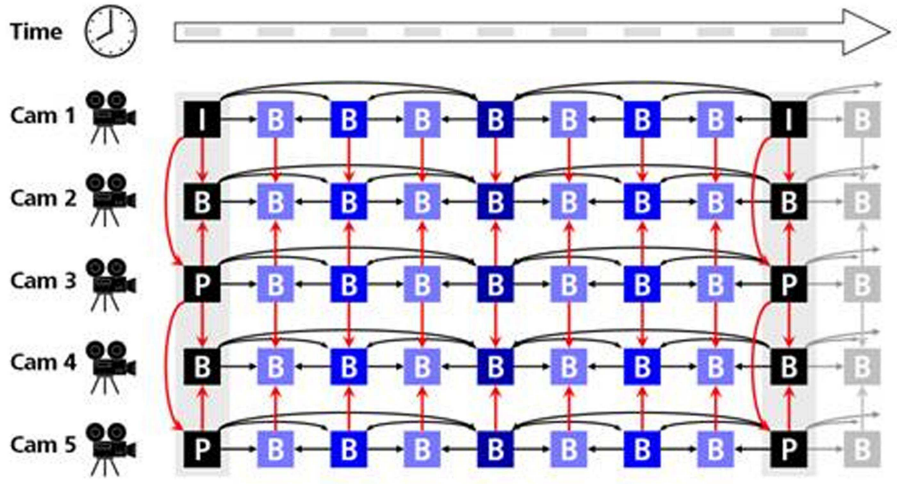


Figure 4.2: Temporal/inter-view prediction structure for MVC [1].

Therefore, it is important for the multi-view encoder to have the ability to issue different views and frames. Naturally, this increases the complexity of the multiview encoder, which needs to be limited.

4.4 Multiview video coding standard

One possible and straightforward solution of encoding an MVV is to encode each view independently from any other views, which is called simulcast coding. In this solution, the redundancy in the temporal domain can be removed by motion estimation as outlined in the previous chapter. However, there is also a high correlation between individual views in the MVV sequences. Therefore, more efficient compression performance can be achieved if the redundancies in the view domain can also be eliminated. Apart from the simulcast coding, several techniques and prediction structures to encode MVV efficiently have already been reported [3–6]. It has been experimentally proven that the prediction scheme shown in Figure 4.2 is the prediction scheme that shows the best RD performance [7–9]. A common conclusion from the conducted experiments is that the encoding of MVV by exploiting the temporal and inter-view dependencies yields much better rate-distortion performance than the simulcast coding technique. This is visible in Figure 4.3 in which sample RD curves from the simulcast and MVC coding for the Ballroom test sequence are shown.

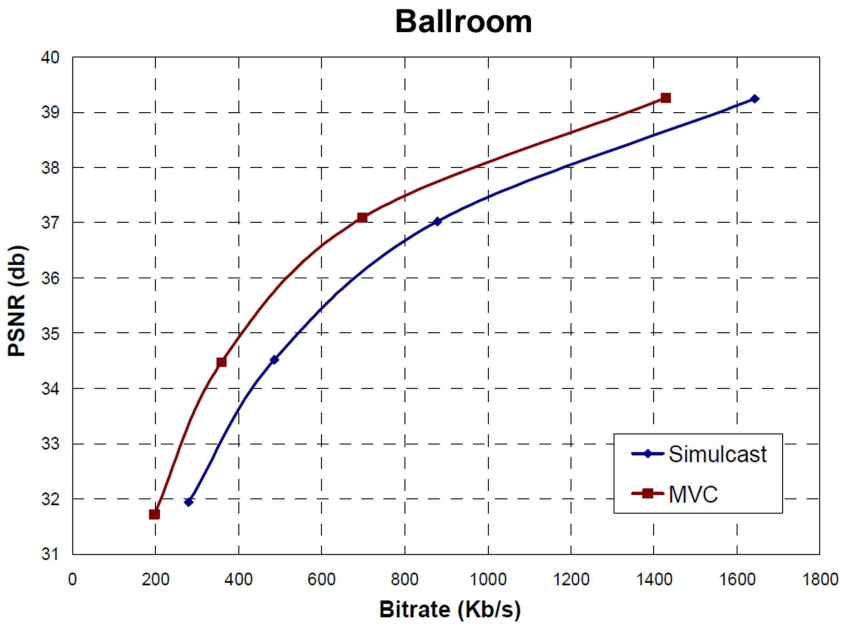


Figure 4.3: Rate-distortion results from simulcast and MVC for the Ballroom test sequence [2]

It has been also reported that the MVC shows parallel results for other test sequences [2, 10].

MPEG and VCEG developed 'H.264/AVC Multiview High Profile' as a standard on MVC in 2008. The reference software which came along with this standard is called Joint Multiview Video Coding Model (JMVC [1]) [11]. Since it is based on H.264/AVC, block based disparity and motion estimation is used and it supports all the frame types existing in H.264/AVC.

The unique feature of JMVC is that it supports multiple reference frame selection in either the temporal or the view domain. It promises high coding efficiency due to the fact that the correlation can be exploited in both domains. However, the frames to be searched for correlation need to be buffered during the encoding. Naturally, this increases the memory requirement which makes the implementation of MVC harder and expensive [12].

[1] As is the case in the reference software of the H.264 video coding standard, the JMVC reference software is written in C++ and its source code is publicly available.

The JMVC offers a high coding efficiency at the cost of a drastically increased computational load. The complexity of an MVC encoder is much higher than that of a single view encoder since the redundancy in the view domain is to be eliminated as well as in the temporal domain. On one hand, the possibility of finding more similar block information increases by searching in different frames in both time and view domain. On the other hand, the motion estimation process when combined with the disparity estimation causes a significant increase in the computation load of the encoder, which needs to be reduced for real time applications.

In the MVC standard, the same mode decision and rate-distortion optimization method explained in Section 3.4.10 for H.264/AVC is used. That is to say that, the best block mode is decided based on the RD cost value of the resultant motion or disparity vector.

4.5 Complexity issue in MVC

The complexity of the prediction scheme shown in Figure 4.2 is very high since the vector prediction and compensation need to be performed both in the temporal and in the view domain [13]. This increases the memory requirement and possible delay in the application.

As it appears in Figure 4.2, the cameras are oriented parallel and positioned equidistantly on one axis. In the case of 2D camera arrays in which the cameras are arranged both horizontally and vertically as explained in Section 2.4.3, the complexity of a customized similar prediction structure would be much higher and needs to be reduced for real time applications. Since there is a correlation between the complexity and the duration of the encoding time, the encoding process will take less time if its complexity is reduced. The less encoding time, the less processing time and the less energy that is required by the processing unit. Therefore reducing the complexity of the encoder is beneficial also for non-real time applications."

Exploiting the inter-view correlations between views in an MVV is a very challenging task due to the high computational complexity which makes it difficult to implement. Many methods that promise to reduce the complexity of the multiview encoder have been proposed. It is noteworthy to mention that only horizontally distributed views are employed in all following methods. Kannangara et al. reported that the complexity reduction can be obtained by decreasing or early termination of the number of prediction modes for a macroblock [14]. Choi et al. proposed an early SKIP mode decision and selective intra mode decision methods to reduce the computational complexity of the encoder [15]. Yin et al. employed an

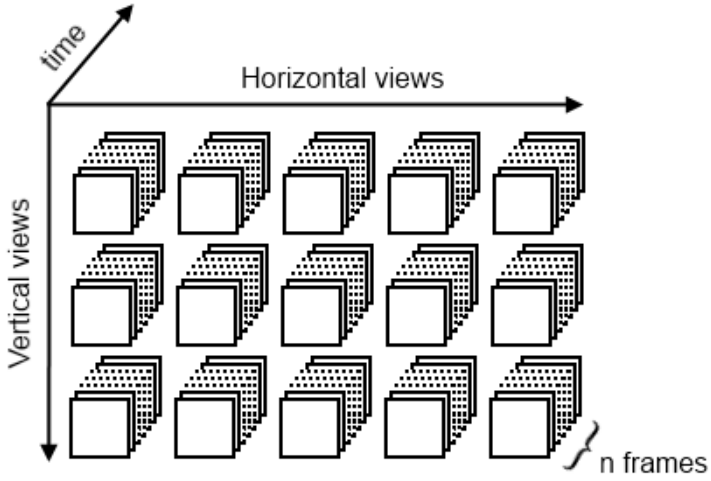


Figure 4.4: Illustration of the multiview video sequences taken from cameras located in both horizontal and vertical direction.

algorithm which jointly optimizes the motion estimation and the mode decision process [16]. Deng et al. introduced an iterative search method for motion and disparity estimation by utilizing the stereo-motion consistency constraints for the stereoscopic video coding [17]. Apart from above mentioned methods, different techniques to reduce the enormous computational load to encode multiview videos have been proposed in the literature. For example, Li et al. proposed a method to reduce the motion and disparity estimation computing by limiting the search region [18]. Zeng et al. described an algorithm to achieve the complexity reduction by terminating the mode decision process if the rate-distortion (RD) cost of a macroblock is lower than an adaptively calculated threshold value [19]. Ding et al. proposed a content-aware prediction algorithm with inter-view mode decision to reduce the high computational complexity of the multiview video encoder [20]. Shen et al. proposed a fast disparity and motion estimation algorithm based on the homogeneity which is estimated by looking at the motion vectors of neighbouring and previously encoded corresponding macroblocks [21, 22]. Zhu et al. described a fast disparity estimation algorithm which utilizes the spatio-temporal correlation and the temporal variation of the disparity field [23].

4.6 Multiview video imagery

All methods mentioned in the previous section are achieving the complexity reduction in MVV which contains either only horizontally distributed views or some selected views of view images captured by a 2D camera array. This is because most 3D imaging systems do not provide vertical parallaxes to the viewers due to cumbersome problems such as complex optical design, difficult implementation, high system cost and huge data requirements. However, vertically spaced views are indispensable for any real 3D experience since they provide supplementary information of the scene.

In Figure 4.4, an illustration of a multiview video sequences taken from a 2D camera array where the cameras are located both horizontally and vertically is shown. There are many different ways to encode such MVV besides the simulcast coding of each view individually. One of the best ways would be to encode this MVV by performing the disparity and the motion estimations on both time and view domain for each view frame. However, the computational load of the encoder will be very high especially with the existence of the vertical views and may not be feasible for practical reasons such as processing power, memory requirement and the instant access of specific view frames in the MVV. Another option is to encode one time instant view images from each view as a group. In this case, no estimation in time domain will be performed.

To encode multi-view image data captured by a 2D camera array at one time instant, different prediction schemes have already been proposed in the literature. Merkle et al. proposed prediction schemes shown in Figure 4.5 (MVCS-1 and MVCS-2) to encode an MVV [10]. Although the redundancies in the view domain can be successfully exploited by one of those schemes, the computational complexity of the encoder is still high due to the complex and time consuming disparity vector searches.

In these prediction schemes, each square with frame type letter indicates a view image captured by a camera in the 2D camera array. By one of these prediction schemes, the view images are aimed to be encoded only in the view domain, not in the temporal domain and the arrows show the prediction direction of the frames. These prediction schemes are designed for a 5x3 camera matrix but the method can be extended to larger matrices.

The location of the I frame is also playing an important role. For example, while the maximum GOP configuration of a coding chain is I-P-P-P-P-P-P-P in Figure 4.5.a, it is only I-P-P-P in Figure 4.5.b. The prediction scheme in Figure 4.5.c contains B frames which increases the RD efficiency.

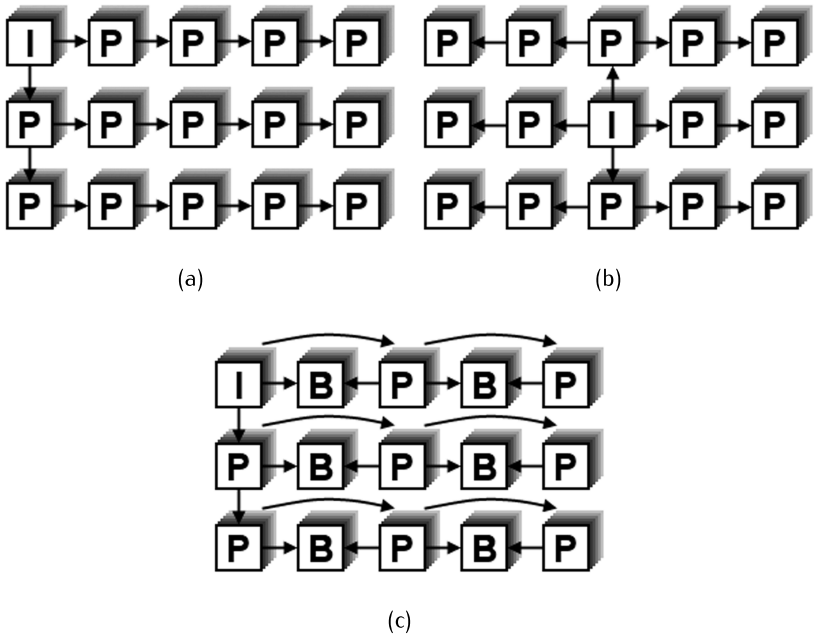


Figure 4.5: Prediction schemes to encode view images taken from a 2D camera array [24]. a) MVCS-1 b) MVCS-2 c) MVCS-3

Alternative prediction schemes to these ones can be designed. On the one hand, a prediction scheme can promise a good coding efficiency, but on the other hand it may generate a high computational load which may not be reasonable to use in systems which have limited process power capability. Therefore, when designing an MVC prediction scheme, the complexity is an important issue that needs to be monitored closely besides the RD performance of the encoder.

In the rest of the dissertation, I will give details on the possibilities of reducing the complexity of the prediction schemes given in Figure 4.5. I will propose several novel alternative prediction schemes constructed with the complexity efficient frame types that I introduced and their RD and complexity performances are experimentally compared.

References

[1] I. J. 1. 2. 11, *Introduction to Multiview Video Coding*. WG 11 Doc. N9580, Antalya, Turkey, 2008 (cit. on p. 60).

[2] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard", *Proceedings of the IEEE*, vol. 99, no. 4, pp. 626–642, 2011 (cit. on p. 61).

[3] H. Kalva, L. Christodoulou, L. Mayron, O. Marques, and B. Furht, "Challenges and opportunities in video coding for 3D TV", in *IEEE International Conference on Multimedia and Expo*, Tonto, Ontario, Canada, 2006 (cit. on p. 60).

[4] D. Socek, D. Culibrk, H. Kalva, O. Marques, and B. Furht, "Permutation-based low-complexity alternate coding in multi-view H.264/AVC", in *IEEE International Conference on Multimedia and Expo*, Tonto, Ontario, Canada, 2006 (cit. on p. 60).

[5] X. Cheng, L. Sun, and S. Yang, "A multiview video coding scheme using shared key frames for high interactive application", in *Picture Coding Symposium (PCS)*, Beijing, China, 2006 (cit. on p. 60).

[6] K. J. Oh and Y. S. Ho, "Multi-view video coding based on the lattice-like pyramid GOP structure", in *Picture Coding Symposium (PCS)*, Lisbon, Portugal, 2007 (cit. on p. 60).

[7] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1461–1473, 2007 (cit. on p. 60).

[8] P. Merkle, K. Muller, A. Smolic, and T. Wiegand, "Efficient compression of multi-view video exploiting inter-view dependencies based on H.264/MPEG4-AVC", in *IEEE International Conference on Multimedia and Expo*, Tonto, Ontario, Canada, 2006 (cit. on p. 60).

[9] M. Flierl, A. Mavlankar, and B. Girod, "Motion and disparity compensated coding for multiview video", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1474–1484, 2007 (cit. on p. 60).

[10] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1461–1473, 2007 (cit. on pp. 61, 64).

[11] Pandit, P. and Vetro, A. and Chen, Y., WD 1 Reference software for MVC, JVT-AA212, Geneva, CH, April, 2008. (cit. on p. 61).

[12] Y. Chen, Y. K. Wang, K. Ugur, M. Hannuksela, J. Lainema, and M. Gabbouj, "The emerging MVC standard for 3D video services", *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, p. 786 015, 2009 (cit. on p. 61).

[13] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. B. Akar, G. Triantafyllidis, and A. Koz, "Coding algorithms for 3DTV - a survey", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1606–1621, 2007 (cit. on p. 62).

[14] C. S. Kannangara, I. E. G. Richardson, M. Bystrom, J. R. Solera, Y. Zhao, A. MacLennan, and R. Cooney, "Low-complexity skip prediction for H.264 through lagrangian cost estimation", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 202–208, 2006 (cit. on p. 62).

[15] I. Choi, J. Lee, and B. Jeon, "Fast coding mode selection with rate-distortion optimization for MPEG-4 Part-10 AVC/H.264", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 12, pp. 1557–1561, 2006 (cit. on p. 62).

[16] Y. Peng, H. Y. C. Tourapis, A. M. Tourapis, and J. Boyce, "Fast mode decision and motion estimation for JVT/H.264", in *International Conference on Image Processing*, Barcelona, Catalonia, Spain, 2003 (cit. on p. 63).

[17] Y. L. Chan, K. B. Jia, C. H. Fu, and W. C. Siu, "Fast motion and disparity estimation with adaptive search range adjustment in stereoscopic video coding", *IEEE Trans. Broadcast.*, vol. 58, no. 1, pp. 24–33, 2012 (cit. on p. 63).

[18] X. Li, D. Zhao, S. Ma, and W. Gao, "Fast disparity and motion estimation based on correlations for multiview video coding", *IEEE Trans. on Cons. Elec.*, vol. 54, no. 4, pp. 2037–2044, 2008 (cit. on p. 63).

[19] H. Zeng, K. K. Ma, and C. Cai, "Mode-correlation-based early termination mode decision for multi-view video coding", in *17th IEEE International Conference on Image Processing (ICIP)*, Hong Kong, China, 2010 (cit. on p. 63).

[20] L. F. Ding, P. K. Tsung, S. Y. Chien, W. Y. Chen, and L. G. Chen, "Content-aware prediction algorithm with inter-view mode decision for multiview video coding", *IEEE Transactions on Multimedia*, vol. 10, no. 8, pp. 1553–1564, 2008 (cit. on p. 63).

[21] L. Shen, Z. Liu, S. Liu, Z. Zhang, and P. An, "Selective disparity estimation and variable size motion estimation based on motion homogeneity for multi-view coding", *IEEE Transactions on Broadcasting*, vol. 55, no. 4, pp. 761–766, 2009 (cit. on p. 63).

[22] L. Shen, Z. Liu, T. Yan, Z. Zhang, and P. An, "View-adaptive motion estimation and disparity estimation for low complexity multiview video coding", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 6, pp. 925–930, 2010 (cit. on p. 63).

[23] W. Zhu, X. Tian, F. Zhou, and Y. Chen, "Fast disparity estimation using spatio-temporal correlation of disparity field for multiview video coding", *IEEE Trans. on Cons. Elec.*, vol. 56, no. 2, pp. 957–964, 2010 (cit. on p. 63).

[24] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, pp. 1461–1473, 2007 (cit. on p. 65).

5

Complexity efficient P frame

"Fools ignore complexity. Pragmatists suffer it.
Some can avoid it. Geniuses remove it."

Alan Perlis

5.1 Introduction

The performance of the encoder is highly dependent of the processing power of the system, which is, however, rather limited in most implementations. There are many computationally intensive processes in the H.264/AVC video coding standard, among which the motion or disparity estimation process is the most complex and time consuming [1–4]. Every extra step to improve the coding efficiency of the encoder mostly results in an increase in complexity of the encoder. Typically, the complexity of MVC exceeds that of its single view counterpart by a factor equal to the number of views. As a result of this, the complexity of the MVC is an important issue which needs to be reduced for practical reasons besides the efficiency of the encoder in terms of quality and bitrate.

Substantial correlation is present among views in the MVV sequences [5, 6]. This can be exploited to encode the view images efficiently. This is

a common objective of all MVC techniques. In this work, this information is used to our advantage and is employed to reduce the complexity of the encoder. I will introduce a novel complexity-efficient version of standard P frame called D_P frame in this section. I will propose several novel prediction schemes constructed with D_P frame and I will compare their respective efficiencies in terms of quality, bitrate and the complexity.

5.2 Parallel geometry

As mentioned in Section 4.3, the disparity means the apparent positional difference of an object in the image plane when captured by separate cameras. As can be seen from Figure 5.1, the position coordinates of the image of the object in different image planes are different. A disparity vector is a measure of this displacement of image in different views. The magnitude of the disparity vector is directly related to the distance of the object from the cameras and also on the spacing between the cameras. Small object to camera distance implies a large disparity vector. Similarly, an increase in inter camera spacing results in enlargement of magnitude of disparity vector.

In Figure 5.1, the aforementioned relationship between the disparity vectors in the parallel camera geometry are shown. In this figure, the cameras are equidistant and a ray emanating from the object is projected on the corresponding projection plane. The point where the ray intersects the projection plane gives the true translation of the object point in the camera space.

By taking the tangent of each angle between the incoming ray and the projection plane, the following equations can be written;
For camera 1,

$$\frac{d}{x_1} = \frac{D-d}{c-t-x_1} \tag{5.1}$$

For camera 2,

$$\frac{d}{x_2} = \frac{D-d}{t-x_2} \tag{5.2}$$

For camera 3,

$$\frac{d}{x_3} = \frac{D-d}{c+t-x_3} \tag{5.3}$$

where x_1, x_2 and x_3 are the distance between the intersection point and the center of the projection plane, D is the distance between the camera plane and the object and d is the distance between the camera plane and the projection plane. After solving all above equations together, it follows

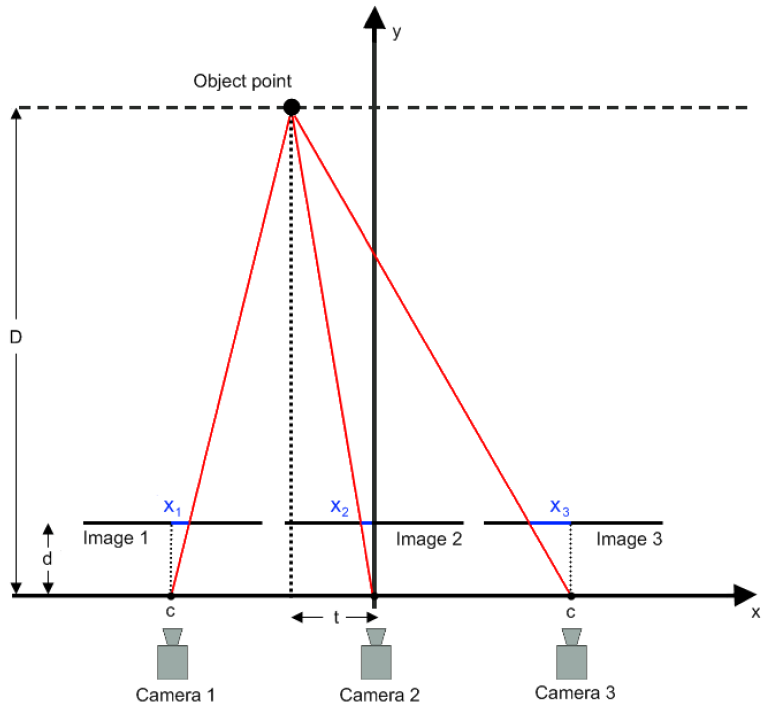


Figure 5.1: Parallel camera geometry and projection of an object point.

that:

$$x_2 = \frac{(x_3 - x_1)}{2} \tag{5.4}$$

$$|DV_{1-2}| = x_2 + x_1 \tag{5.5}$$

$$|DV_{2-3}| = x_2 - x_3 \tag{5.6}$$

where the DV_{1-2} and DV_{2-3} represent the disparity vectors from camera 1 to camera 2 and from camera 2 to camera 3 respectively.

From these equations, it can be concluded that the disparity vectors of points in different views are purely translational especially if the cameras are located equidistantly [7–9]. Moreover, it also follows that the disparities of a point in the scene photographed by cameras which are spaced equidistantly are equal.

This is also valid for vertical views, which is illustrated in Figure 5.2.

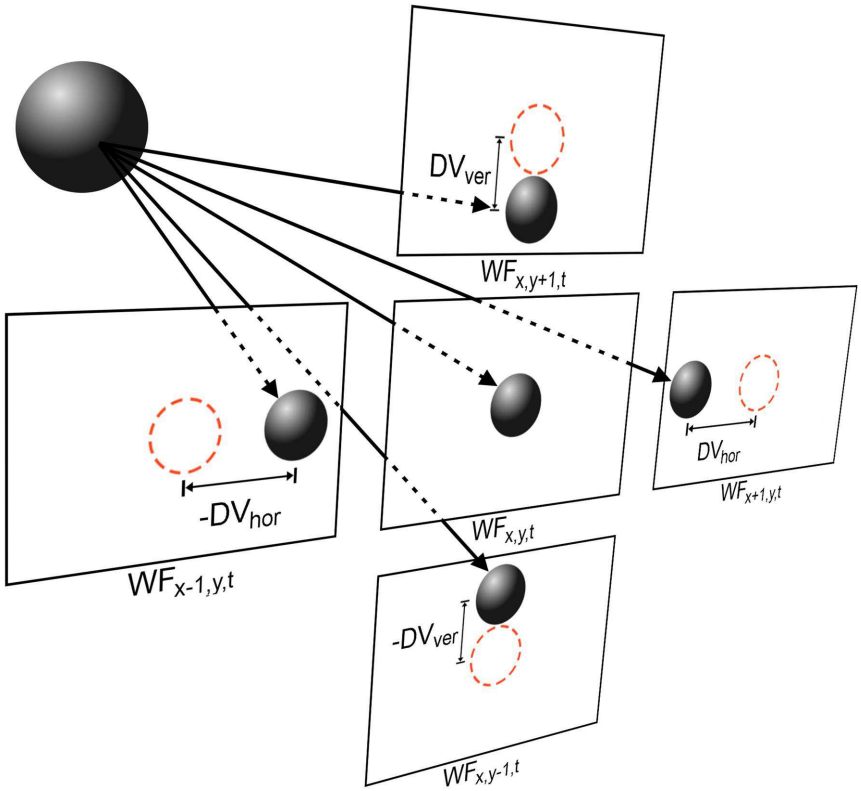


Figure 5.2: Captured view images of an object by equidistantly positioned cameras.

In this case, the size of the disparity vector from $WF_{x,y,t}$ to $WF_{x+1,y,t}$ is the same as the disparity vector from $WF_{x,y,t}$ to $WF_{x-1,y,t}$. Similarly, as for the vertical views, the disparity vector from $WF_{x,y,t}$ to $WF_{x,y+1,t}$ equals the disparity vector from $WF_{x,y,t}$ to $WF_{x,y-1,t}$. A minus sign preceding the disparity vector indicates that the disparity vector has the opposite direction.

This equilibrium proves that there is a strong geometrical relationship between view images captured by a 2D camera array in which the cameras are located equidistantly. This idea will later be employed in the algorithms to reduce the complexity of the MVC, which will be explained in detail in the next section.

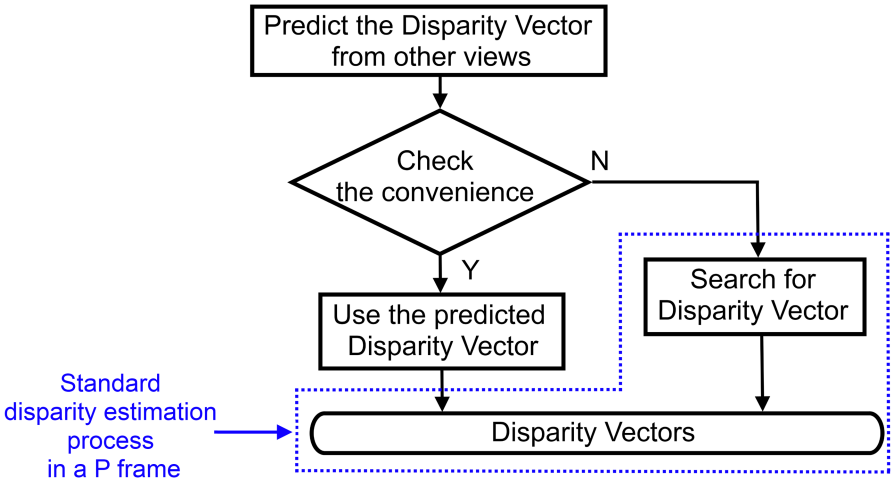


Figure 5.3: The flow chart of the disparity estimation process of the D frame.

5.3 Derived P frame (D_P frame)

To encode one time instant of view images, either of the prediction schemes, containing I or P frames can be used, as shown in Figure 4.5. Although these schemes are aiming to achieve high coding efficiency in terms of quality and bitrate, the overall complexity of the encoder remains high and needs to be reduced for practical reasons.

Although many fast disparity vector search methods have been proposed in the literature, it continues to be the most complex and hence most time consuming process of the encoder. This problem becomes even more severe for MVV sequences since several views need to be encoded at every time instance. For this purpose, I hereby introduce the D_P frame which is a standard-compliant and complexity efficient version of the P frame. The ultimate aim of the D_P is to reduce the complexity burden of the encoder when encoding one time instant view images captured by a 2D camera array.

5.3.1 Structure of the D_P frame

A D_P frame is a frame for which the encoder does not need to run the disparity estimation process for all of its macroblocks. The key idea behind the D_P frame is that most of the disparity vectors in a frame can success-

fully be derived from the already known disparity vectors of other views because of the existing geometrical correlation among views as discussed in the previous section.

To implement the D_P frame, the disparity estimation process of the P frame is modified and its block diagram is shown in Figure 5.3. The algorithm applied to a block in a D_P frame proceed as follows:

1. The disparity vector of the block is derived from the other views in a fashion determined by the prediction scheme that is selected for encoding the whole MVV.

2. The RD cost value of the block pointed by the derived disparity vector is calculated in the same way as it is computed in the reference software.

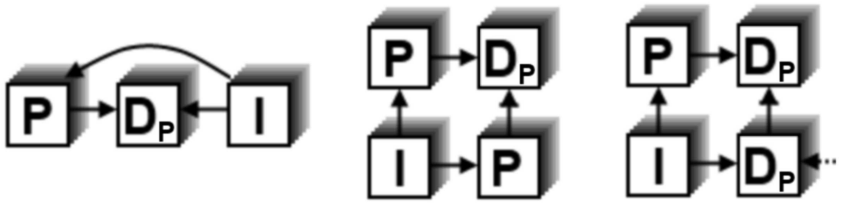
3. The calculated RD cost value is compared to a predefined threshold value (TH) in order to assess whether the derived disparity is convenient. This comparison is needed since the purely geometric derivation of disparity vectors can not cope with certain situations such as occlusions and viewing angle dependent illumination effects. A suitable choice of the TH value can lead to significant reduction in the number of disparity estimation processes that have to be conducted, resulting in a substantial decrease in the complexity of the encoder.

4. If the comparison in step 3 reveals that the RD cost value is lower, the derived disparity vector is saved and encoding continues without any disparity estimation process for that particular block.

5. If not, then the same disparity estimation process as in a P frame is performed only for that block.

6. The encoding process continues with the transformation and quantization process.

The derivation procedure can vary according to the nature of the prediction scheme. In Figure 5.4, three different ways of disparity vector predictions are given as an example. In Figure 5.4.a, the disparity vectors of the D_P frame are interpolated from already calculated disparity vectors of the P frame. In Figure 5.4.b, the disparity vectors of the D_P frame are the vectorial combination of those of the two neighbouring P frames, a P and a D_P frame (Figure 5.4.c), or even two D_P frames. While the arrow of the P frame indicates its reference frame, the arrows of the D_P frame show the frames whose disparity vectors are used for its derivation.



(a) (b) (c)

Figure 5.4: Three different disparity vector predictions.

5.3.2 Complexity efficient prediction schemes with D_P frames

Traditional prediction schemes to encode one time instant of 5x3 views are already depicted in Figure 4.5. Here, I propose different novel prediction schemes, which are shown in Figure 5.5. In all schemes, either the D_P or the P frame takes the I frame as a reference, implying that the GOP size is only two for each time.

The performance of the encoder is closely related to the chosen prediction scheme. Moreover, the performance of the prediction scheme completely depends on the location of the I and the P frame and the derivation direction of the disparity vectors for the D_P frames. As the distance (in view space) between the I and the P frames increases, the efficiency of the derivation decreases since the frames will contain more occluded blocks and a higher chance of illumination change. The proposed schemes in Figure 5.5 are designed to investigate following aspects:

- The influence of the different locations of I, P and D_P frames on the RD performance of the encoder,
- The success of the derivation process of disparity vectors,
- The effects of the derived disparity vectors on the RD performance and the complexity of the encoder.

While the disparity vectors for D_P frames are derived from disparity vectors of P frames which are diagonally located in scheme 1, disparity vectors of P frames which are located orthogonally are used in schemes 2 and 3 for the derivation process. Scheme 3 is designed to observe the effects of a long distance between I and P frames. All three schemes support quadruple parallel processing or fair load balancing where all D_P

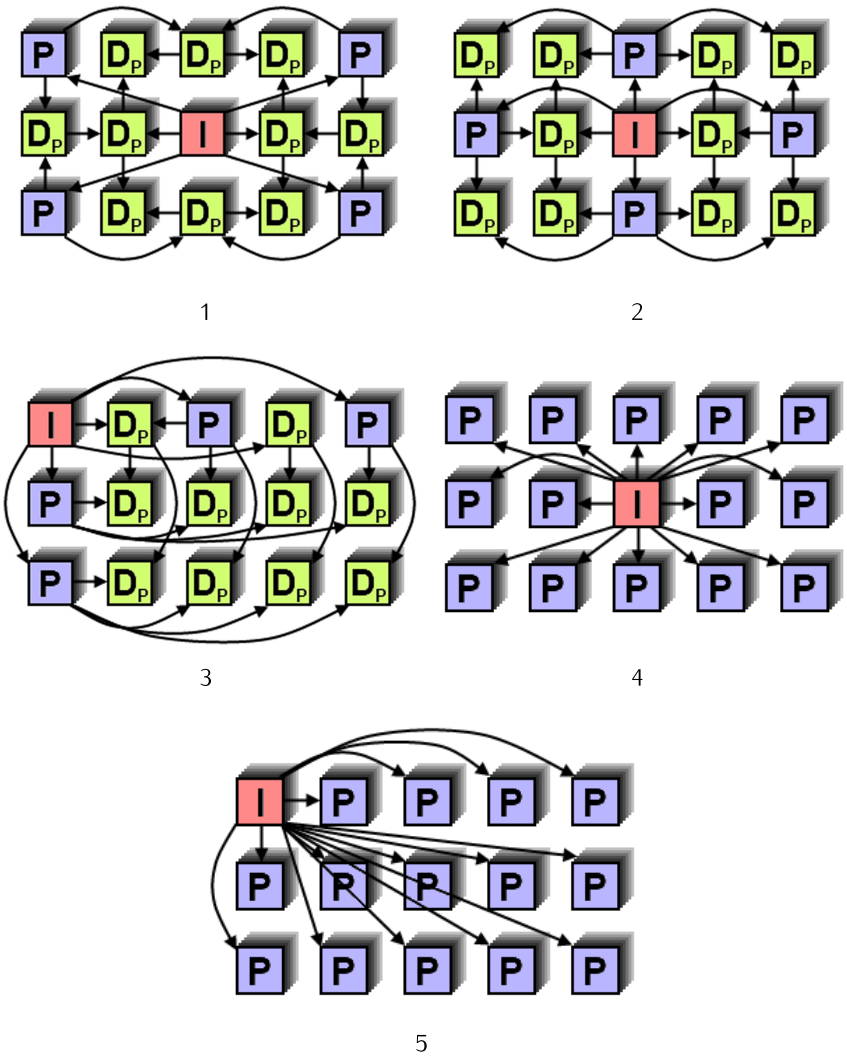


Figure 5.5: Alternative prediction schemes. Scheme 4 is the reference scheme for scheme 1 and 2, and scheme 5 is the reference scheme for scheme 3.

frames can be encoded after encoding the P frames. Schemes 4 and 5 are the reference schemes, where each D_P frame is replaced by a P frame, for the schemes 1, 2 and 3 respectively.

These proposed schemes can further be extended for bigger camera arrays. In such a case, the location of the P frames is of vital importance

Table 5.1: Encoder parameters.

Macroblock partition	8x8
Search Mode	Fast Mode
Search Range	32
Entropy coder	CABAC
QP	22, 27, 32, 37

for the encoder's performance as the accuracy with which the disparity vectors of the P frames can be determined has an important bearing on the success of the derivation of the disparity vectors of the D_P frames. In case of an excessive number of cameras in the camera array, the possibility of dividing the whole view images into non overlapping sub camera arrays should be considered.

5.4 Experimental setup

As mentioned in Section 4.2, the standard reference software for MVC is JMVC. In this work, I used standard-conforming JSVM reference software as encoder [10] and some of the important encoding parameters are listed in Table 5.1. The reason of using JSVM reference software instead of JMVC is that the JMVC reference software does not allow us to encode some of the prediction schemes proposed in this work. Since both software modules were derived from the same code base and the JSVM software supports more configuration options, I chose it as the encoder.

In order to verify and compare the efficiencies of the proposed prediction schemes against their reference schemes, different image sets taken from a 2D camera array are used. Although the camera array has 17x17 views, only the center 5x3 views are used in the experiments due to the dimensions of the proposed and traditional prediction structures. The cropped and rectified view images have been provided to us by Stanford University [11].

Since the PSNR objective quality measurement method (see Section 3.3.2) is widely used in almost all experimental results in the literature, it is employed as the quality metric in this work instead of SSIM method.

Table 5.2: The image sets used in the experiments.

Image Set	Resolution
Amethyst	768 x 1024
Chess	1392 x 800
The Stanford Bunny	1024 x 1024
Eucalyptus Flowers	1280 x 1536
Treasure Chest	1536 x 1280
Truck	1280 x 960

5.5 Performance evaluation of complexity efficient prediction schemes with D_P frames

5.5.1 Threshold analysis

As it appears from the flow chart of the D_P frame in Figure 5.3, the disparity estimation process can only be skipped if the RD cost value of the derived disparity vector is lower than the TH. The reduction in the complexity of the encoder depends directly on the extent of skipping of the disparity estimation processes, which in turn is decided by the TH.

In order to investigate the influence of the TH on the coding efficiency of all proposed schemes in Figure 5.5, these prediction schemes with different THs are applied on all image sets. To see the resulting difference, the Bjontegaard method is employed [12]. This method calculates the difference of two RD curves and produces two equivalent results, BD-PSNR and BD-Rate showing PSNR and bitrate differences respectively.

In Figure 5.6, the impact of the various THs on the BD-PSNR and BR-Rate is shown. As it can be interpreted, the D_P frame is exactly the same as the P frame if the TH is set equal to zero. This is clearly visible in Figure 5.6, where there is no BD-PSNR and BD-Bitrate difference between the prediction scheme and its reference scheme when the TH is set equal to zero. As expected, the PSNR and bit-rate degrades as the TH increases. This is because by increasing TH I relax the accuracy of the derived disparity vectors. Therefore, to keep the same image quality, more residual data will have to be added into the data stream. Conversely, if I keep the same bitrate, the image quality (expressed in the PSNR value) will decrease. All curves reach their asymptotic limit for high TH, meaning that under these circumstances no disparity estimation is performed at all

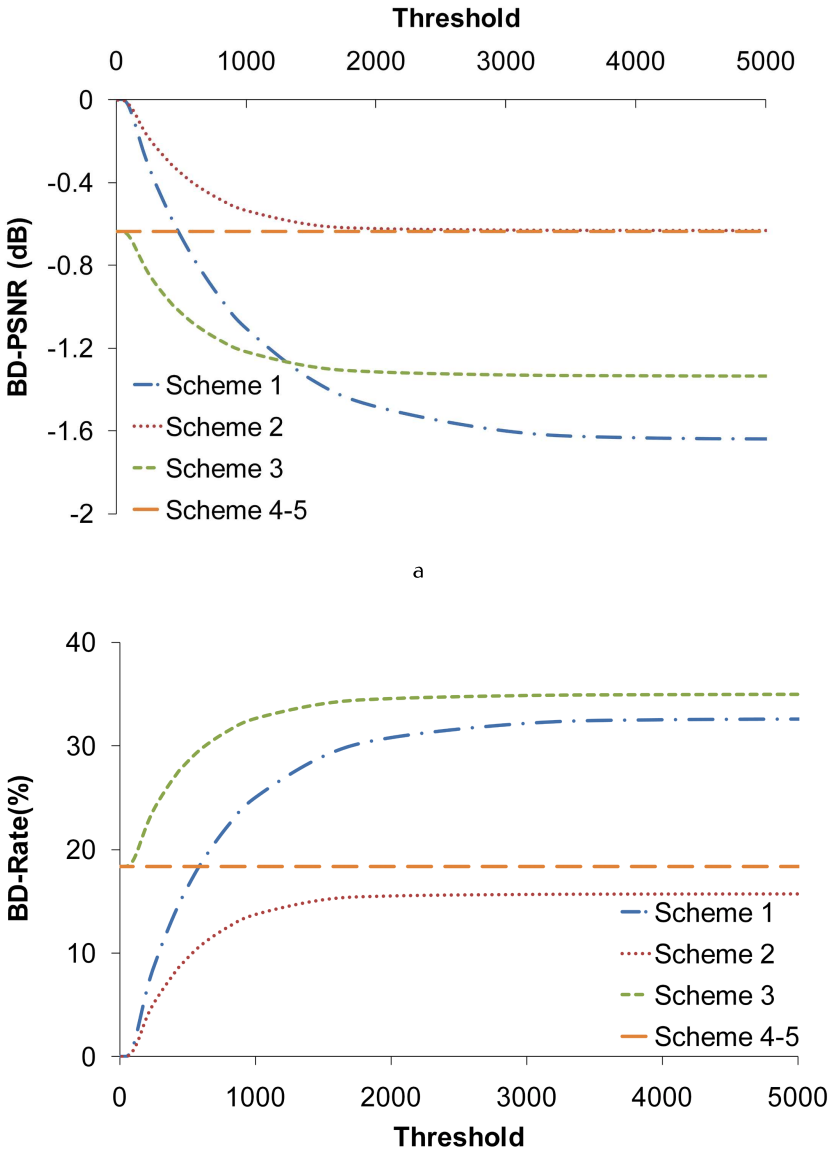


Figure 5.6: Relative quality (a) and bit-rate (b) difference between the schemes and their references. The graphs show the averaged results from all image sets.

and the geometrically derived disparity vectors are accepted. Scheme 3 is based on another reference scheme (scheme 5) and therefore, the relative difference between the reference schemes is taken as an offset point for this curve. It is obvious from Figure 5.6 that scheme 2 produces the lowest loss in terms of quality and bitrate compared to the other schemes. An encoder which is configured with scheme 2 can suffer a maximum loss of 0.6 dB in quality for equal bitrate or an increase of 15.7% in bit-rate for equal quality.

On the other hand, the benefit of using a TH is that the number of executed disparity estimation processes decreases and hence the complexity of the encoder is reduced. Figure 5.7 shows the average number of searched blocks of the first three schemes as a function of TH for different values of the QP. As expected, the number of searched blocks increases with QP. This is because the calculation of the RD cost value of a block is QP dependent (see Section 3.4.10). A higher QP yields a larger difference between the candidate block and the reference block. As a result, for a given TH value, more blocks will yield RD cost values above this TH and have to be searched using the conventional disparity estimation process, resulting in an increasing complexity. Comparing Figures 5.7 and 5.8, I notice that the complexity decreases significantly with rising TH, while the coding efficiency is only slightly affected. The effect is most prominent for the lowest values of TH, below the point where the curves corresponding to the schemes with D_P frames in Figure 5.6.a start going off x-axis, but stays valid even for the asymptotic limits for which all disparity vectors of the D_P views are derived geometrically from other views. This will be explained in detail in Section 6.2. It proves that this geometrical disparity vector derivation concept is a highly efficient way to reduce the encoder complexity at the expense of only a minor quality or bitrate loss. The lower limit of the number of searched blocks and hence the complexity depends solely on the number of P frames in the schemes. Since four P frames are used in all proposed schemes and disparity estimation has to be performed for every block in the P frames, this leads to a limit value of $4/14 = 28.5\%$.

5.5.2 Rate distortion analysis

The prediction schemes were applied to the image sets listed in Table 5.2 and the resulting average RD curves are plotted in Figure 5.8 (a fixed predefined TH value of 150 is used). As defined in Figure 5.5, the proposed complexity efficient prediction schemes 1, 2 and 3 have to be compared

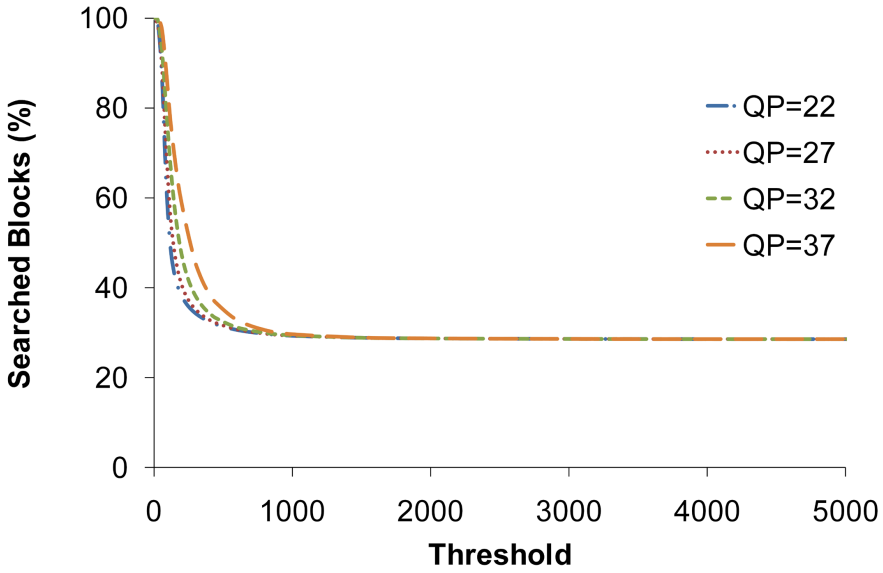


Figure 5.7: Percentage of searched blocks as a function of the threshold value for different quantization parameters. The presented data was averaged over all schemes.

to their corresponding reference schemes, either scheme 4 or scheme 5. Naturally, the reference schemes produce the best results since the TH is not selected as zero. If the TH is further increased, the gaps between schemes are expected to widen in accordance with the Bjontegaard metric results shown in Figure 5.6.

Since the QP has a strong effect on the calculation of the RD cost value and TH represents a point in the RD cost value scale, it is necessary to choose TH depending on the QP. Since TH is chosen as 150 for any QP in Figure 5.8, the curves diverge as QP is increasing and, similarly, the percentage of searched blocks is increasing in Figure 5.9.

The average percentage of searched blocks, which is an indicator for the complexity of the encoder, is measured using the same parameters as used in Figure 5.8 and is presented in Figure 5.9. Both figures (Figure 5.8 and Figure 5.9) need to be considered together because of the balance between quality, bitrate and complexity. If the scheme 2 is employed with QP = 22, the quality drop is 0.1 dB or the bit rate increase is 1.5% while the complexity gain is about 58%. As QP increases, this gain drops, as explained in the previous section. Figure 5.8 and 5.9 demonstrate that

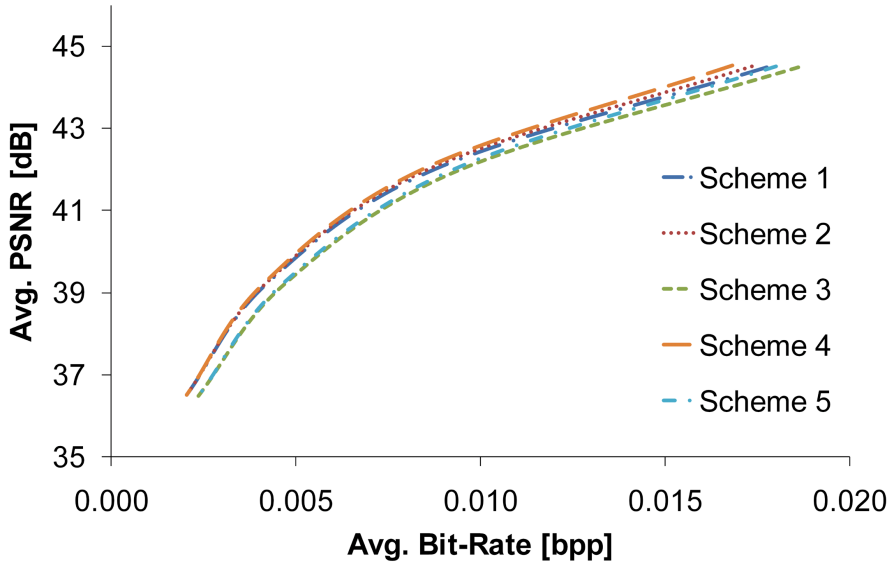


Figure 5.8: Rate–distortion performances of all schemes (TH = 150). The results from all image sets are averaged. dB and bpp stand for decibel and bit per pixel respectively.

Table 5.3: Performance of complexity efficient prediction schemes 1, 2 and 3 as a function of TH.

TH	Scheme 1			Scheme 2			Scheme 3		
	QL (dB)	BRI (%)	AVG (%)	QL (dB)	BRI (%)	AVG (%)	QL (dB)	BRI (%)	AVG (%)
800	0.9	27.6	30.6	0.4	11.6	29.6	0.5	12.3	29.9
450	0.6	15.8	34	0.3	7.2	32.8	0.3	7.9	33.6
270	0.3	8.4	40.8	0.2	4.1	39.3	0.2	4.5	40.7
180	0.2	4.4	49.6	0.1	2.2	48	0.1	2.4	49.9
150	0.1	2.9	55	0.1	1.5	53.7	0.1	1.6	55.5
120	0.1	1.6	63.1	0	0.8	61.5	0	0.9	63.9
60	0	0.1	91.9	0	0.1	90.9	0	0.1	92
0	0	0	100	0	0	100	0	0	100

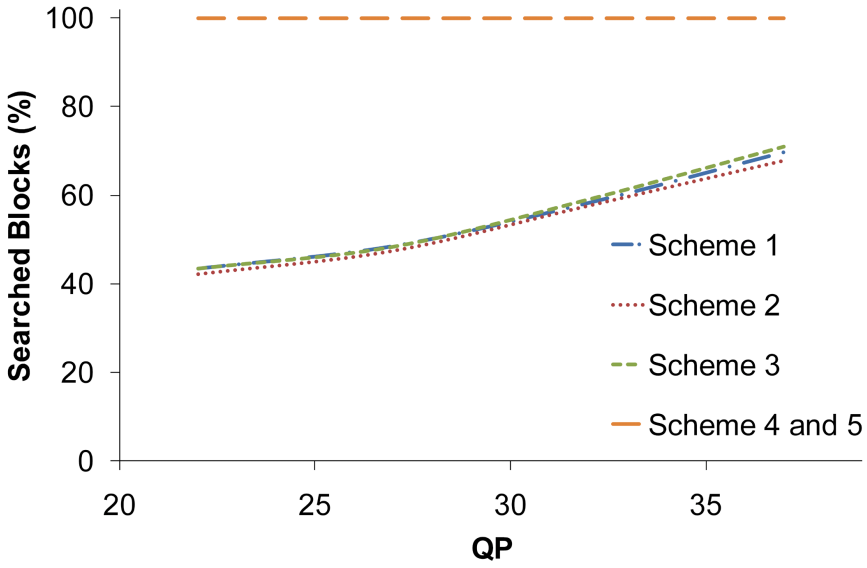


Figure 5.9: Complexity performance of all schemes (TH = 150). The results from all image sets are averaged.

scheme 2 again yields the optimum result among the proposed schemes both what coding efficiency and complexity is concerned. Moreover, visual inspection did not show any differences in subjective quality between the reference scheme and scheme 2.

In Table 5.3, the relative quality loss (QL) and bit-rate increase (BRI) of the schemes as well as the percentage of searched blocks (AVG) averaged over QP values ranging from 22 to 37 are presented as a function of TH. The best value of the TH can be chosen depending on the application, user requirements and processing power limitations. While the TH should be chosen high enough for applications where the complexity has higher priority than the bit-rate and quality, it is wise to choose a TH below the inflection point of the curve in order to have the maximum complexity gain without having any deformation in RD performance. However, Table 5.3 clearly confirms the superiority of scheme 2 over the other proposed schemes.

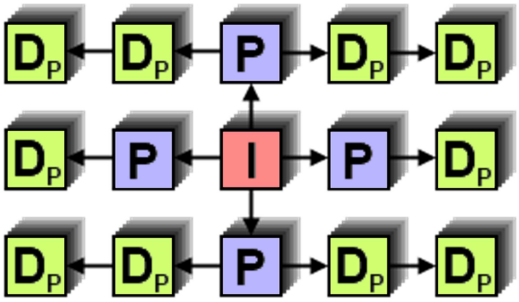


Figure 5.10: Modified version of traditional multiview encoding scheme (CR-MVCS). Some of the P frames are replaced with D_P frames.

5.6 Complexity efficient prediction scheme with variable threshold

In the prediction schemes given in Figure 5.5, each view image was encoded by having the I frame as reference, meaning that the GOP size was two (I-P or I-D_P) for each iteration. In Figure 5.10, an alternative prediction scheme in which longer GOP sizes are employed (I-P-D_P and I-P-D_P-D_P) is proposed. By this way the RD and complexity performance of the proposed prediction scheme can fairly be compared with the traditional prediction schemes given in Figure 4.5 (MVCS-1 and MVCS-2).

In the prediction scheme in Figure 5.5, the P frames are the first frames to be encoded after the I frame. Therefore, the disparity vectors of the blocks of the P frame are already available before encoding the D_P frames starts. Thus, the disparity vectors for the D_P frames can be derived from the disparity vectors of the P frames.

The arrows drawn in the prediction scheme show the prediction direction of the disparity vectors. The disparity vectors of the D_P frames situated in the middle row are extrapolated from the disparity vectors of adjacent P frame in the same row. The disparity vectors of the D_P frames in the other row are derived from the disparity vectors of the corresponding middle row frame.

In previous section (Section 5.5.1), I have already shown the RD chart of a prediction scheme with fixed TH for all the QPs. In this section, it was mentioned that the curves diverge as the QP increases. This problem occurs due to the fixed TH for all QPs. Since the TH is a value to be compared with the RD cost value and the RD cost value is QP dependent

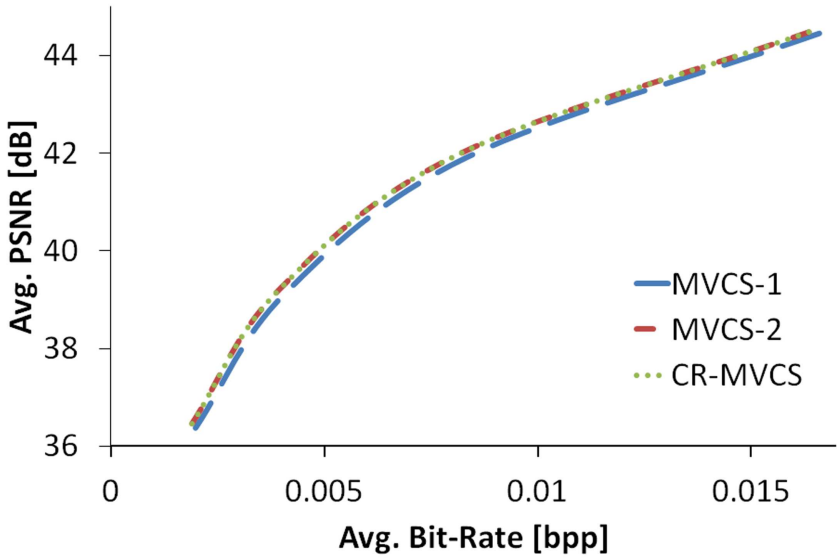


Figure 5.11: Average results of all image sets. The curve of CR-MVCS is on top of the curve of MVCS-2.

by its definition, different THs for different QPs need to be determined.

The RD graph of averaged results from all image sets is plotted in Figure 5.11. The ideal values for TH were experimentally determined as a function of QP for every image set so that the CR-MVCS produces the same RD results as MVCS-2. The resulting THs are summarized in Table 5.4. As it can be concluded from Figure 5.11, the MVCS-2 yields better results than MVCS-1. The major reason is that five view frames are cascade encoded in MVCS-1, which reduces the efficiency of the encoder. On the other hand, although identical results are achieved from CR-MVCS and MVCS-2, the number of block positions that were evaluated during the search for the appropriate disparity vectors is lower for CR-MVCS than for MVCS-2. The percentage complexity gain is dependent on the QP and on the content of the image set. The lower the value of QP, the lower the allowed value of TH to obtain the same RD results and consequently the higher the number of disparity estimation that is required. On the other hand, the complexity can be reduced more by choosing a larger value for TH. In this case, the quality and bitrate result deteriorates. Nevertheless, this property may be a desirable solution for systems where limiting the process power is a greater concern than the preservation of the quality

Table 5.4: Experimental results from all image sets. The different THs are selected for different QPs and image sets. The percentage of searched blocks shows the remaining percentage of the original amount of disparity estimation calculation effort.

Image Set	QP=22	QP=27	QP=32	QP=37
	TH / Searched Blocks [%]			
Amethyst	22 / 99.1	55 / 86.7	87 / 71.1	145 / 62.2
The S. Bunny	29 / 99.3	43 / 96.9	86 / 76.8	149 / 62.0
Euc. Flowers	27 / 97.9	40 / 92.0	80 / 81.5	140 / 64.6
Tre. Chest	38 / 94.9	46 / 89.2	80 / 74.2	138 / 68.5
Truck	32 / 99.7	50 / 98.5	82 / 91.2	136 / 66.0

Table 5.5: BD difference results when the TH parameter is set to its experimented values as shown in Table 5.4. BD-Rate-1, BD-PSNR-1 and BD-Rate-2, BD-PSNR-2 show the Bjontegaard comparison results from CR-MVCS vs. MVCS-1 and CR-MVCS vs. MVCS-2 respectively.

Image Set	BD-Rate-1/ BD-Rate-2 [%]	BD-PSNR-1/ BD-PSNR-2 [dB]
Amethyst	-5.45 / 0	0.20 / 0
The S. Bunny	-4.45 / 0	0.14 / 0
Euc. Flowers	-3.51 / 0	0.13 / 0
Tre. Chest	-4.35 / 0	0.22 / 0
Truck	-6.80 / 0	0.22 / 0

and bit-rate.

The RD curves of the CR-MVCS and MVCS-2 are superimposing each other in Figure 5.11, meaning that there is no quality and bit-rate difference between them. This can also be seen in the Table 5.5 where the BD-Rate-2 and BD-PSNR-2 values are zero. Note that BD-Rate-1, BD-PSNR-1 and BD-Rate-2, BD-PSNR-2 show the Bjontegaard comparison results from CR-MVCS vs. MVCS-1 and CR-MVCS vs. MVCS-2 respectively. In fact, this is the result one would expect if the TH was chosen to be zero, resulting in the full disparity estimation process for all blocks as it is done in traditional P frames. However, thanks to the success of

the disparity vector derivation, I am able to obtain Bjontegaard neutrality using the TH values listed in Table 5.4 and hence a reduction of the number of block searches. If it is averaged for all QPs in the range 22-37, the percentage of searched blocks for disparity vector is 79.8% for the Amethyst image set. Therefore, the complexity of the encoder is reduced by 20.2% for this image set without compromising the quality or bitrate.

5.7 Conclusions

MVC is an essential technology to compress the huge amount of data generated. The MVV can be encoded efficiently if the inter-view redundancy is exploited, however, in this case, the complexity of the encoder increases drastically. This problem becomes more severe if there are vertical views along with the horizontal views present in the MVV. The most complex and time consuming process of the encoder is the disparity estimation process. In this chapter, I have proposed a new frame type called D_P frame with which a complexity reduction can be achieved by skipping the disparity estimation processes for some of its blocks. The skipping process is based on the fact that the disparity vector of a block can be derived from a previously encoded block in another frame due to the strong geometrical relationship between views.

Different novel prediction schemes constructed with the proposed D_P frames and P frames have been introduced. The relative position of the I frame with respect to the D_P and P frames has been experimentally investigated. The best performance is obtained with the prediction scheme in which the I frame is symmetrically positioned with respect to the D_P and P frames.

The complexity of the D_P frame can be adjusted by changing the TH parameter which is used to determine if the derived disparity vector is to be accepted. The TH can be chosen any value, however, the RD performance needs to be monitored since there is a trade-off between the complexity and the RD performance of the encoder. First, a predefined fixed TH is employed. The experimental results showed that the TH should be chosen depending on the QP and the imagery content. There is no RD and complexity difference between the P frame and the D_P frame if TH is set equal to zero. Maximum complexity gain is achieved when TH is set equal to a very high number, where the disparity vector of the blocks are all derived and no disparity estimation process is run for any block. In this case, the RD performance reaches its asymptotic limit. Second, different TH values are experimentally calculated with respect to the QP

and globally applied on all the image sets. It is seen that the complexity reduction is feasible without having deterioration in the RD performance of the encoder. That is to say that, the same quality and bit-rate results can be obtained with significantly reduced complexity gain.

Since no change has been made to the syntax and the encoded bit-stream is fully compliant to the H.264/AVC specification, the decoder does not need any extra information to decode the encoded file, and all disparity vectors in one view can be derived independently from the other views.

It is clear that TH is a very important parameter in the D_P frame. In the experiments in this chapter, always a predefined fixed TH is used. In the next chapter, an algorithm to find the optimal TH value automatically will be explained. Moreover, the TH will be estimated block by block instead of applying a TH globally, which is an important factor to increase the efficiency of the algorithm.

References

[1] S. Na and C. M. Kyung, "A multi-layer motion estimation scheme for spatial scalability in H.264/AVC scalable extension", in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009 (cit. on p. 71).

[2] X. Wu, W. Xu, N. Zhu, and Z. Yang, "A fast motion estimation algorithm for H.264", in *Signal Acquisition and Processing, 2010. ICSAP '10. International Conference on*, 2010, pp. 112–116 (cit. on p. 71).

[3] J. Y. Tham, S. Ranganath, M. Ranganath, and A. A. Kassim, "A novel unrestricted center-biased diamond search algorithm for block motion estimation", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 4, pp. 369–377, 1998 (cit. on p. 71).

[4] D. H. Gerard, P. W. A. C. Biezen, H. Huijgen, and O. A. Ojo, "True-motion estimation with 3-D recursive search block matching", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 3, no. 5, pp. 368+, 1993 (cit. on p. 71).

[5] K. Song, T. Chung, Y. Oh, and C. S. Kim, "Error concealment of multi-view video sequences using inter-view and intra-view correlations", *Journal of Visual Communication and Image Representation*, vol. 20, no. 4, pp. 281–292, 2009 (cit. on p. 71).

[6] M. Zamarin, S. Milani, P. Zanuttigh, and G. M. Cortelazzo, "A novel multi-view image coding scheme based on view-warping and 3D-DCT", *Journal of Visual Communication and Image Representation*, vol. 21, no. 5-6, SI, pp. 462–473, 2010 (cit. on p. 71).

[7] Y. Kim, S. Choi, S. Cho, and K. Sohn, "Efficient disparity vector coding for multiview sequences", *Signal Processing: Image Communication*, vol. 19, no. 6, pp. 539–553, 2004 (cit. on p. 73).

[8] Y. Kim, J. Kim, and K. Sohn, "Fast disparity and motion estimation for multi-view video coding", *Consumer Electronics, IEEE Transactions on*, vol. 53, no. 2, pp. 712–719, 2007 (cit. on p. 73).

[9] T. Y. Chung, I. L. Jung, K. Song, and K. C. S., "Multi-view video coding with view interpolation prediction for 2D camera arrays", *Journal of Visual Communication and Image Representation*, vol. 21, no. 5â€"6, pp. 474–486, 2010 (cit. on p. 73).

[10] ITU-T Recommendation H.264, Advanced Video Coding for Generic Audiovisual Services, ISO/IEC 14496-10 Advanced Video Coding, 2009. (cit. on p. 79).

[11] Stanford University, Computer Graphics Laboratory, 2012, http://lightfield.stanford.edu (cit. on p. 79).

[12] G. Bjontegaard, Calculation of average PSNR differences between RD-curves, ITU-T SG16/Q.6 Doc. VCEG-M33, 2001 (cit. on p. 80).

6 Dynamic threshold value calculation

"Simplicity does not precede complexity, but follows it."

Alan Perlis

6.1 Introduction

In the previous chapter, the D_P frame and different novel complexity efficient multiview prediction schemes with D_P frames were proposed. According to the experimental results, significant complexity gain can be achieved by employing the D_P frames in place of some of the P frames in the prediction scheme. As explained, there is a trade-off between the quality, bitrate and the complexity of the prediction scheme with D_P frames, which can be adjusted by varying the TH. Since my aim is to achieve the same quality and bitrate results as the reference scheme while reducing its complexity as much as possible, the determination of the optimal value of the TH is an important issue.

In this chapter, I propose a method to find the TH automatically by analysing the RD cost values of the blocks in the reference frames. The TH

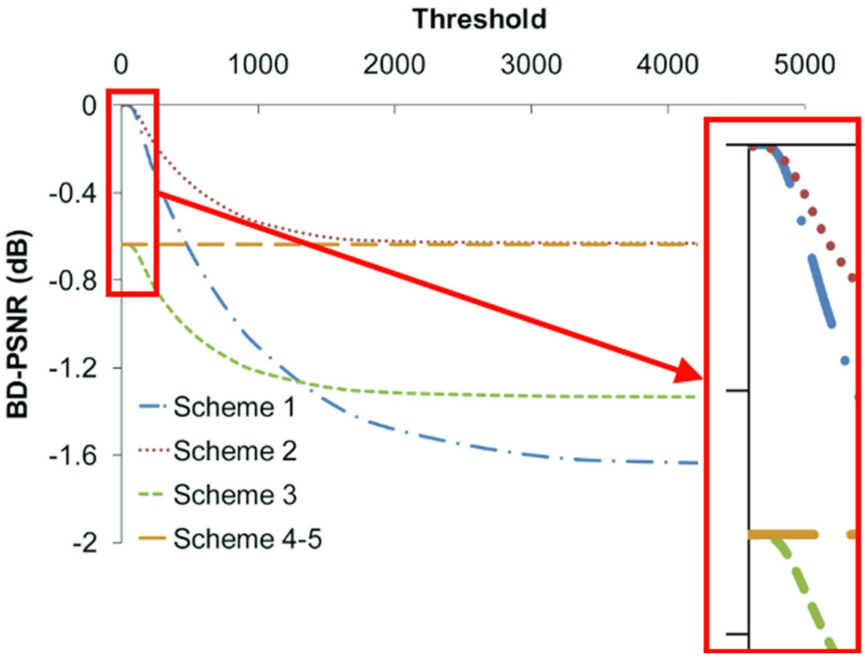


Figure 6.1: Critical value of the TH. The graph shows the relative quality difference between the prediction schemes and their references. It reflects the averaged results from all image sets.

had a fixed value for the whole prediction scheme in the previous chapter [1]. In the new method I propose, the THs are calculated locally, meaning that they are calculated for every block independently during the encoding process.

6.2 Automatic threshold value calculation

The TH plays a crucial role on the determination whether a derived disparity vector is convenient as explained in the previous chapter, or not. It was observed that the RD performance of the encoder starts to degrade as the TH increases beyond a certain value where the accuracy of the derived disparity vectors begins relaxing [2]. This can clearly be seen in Figure 6.1 in which a part of the Figure 5.6.a is zoomed. It is this value of TH with which maximum complexity gain can be achieved without compromising the quality and bitrate. The aim of the dynamic TH value calculation

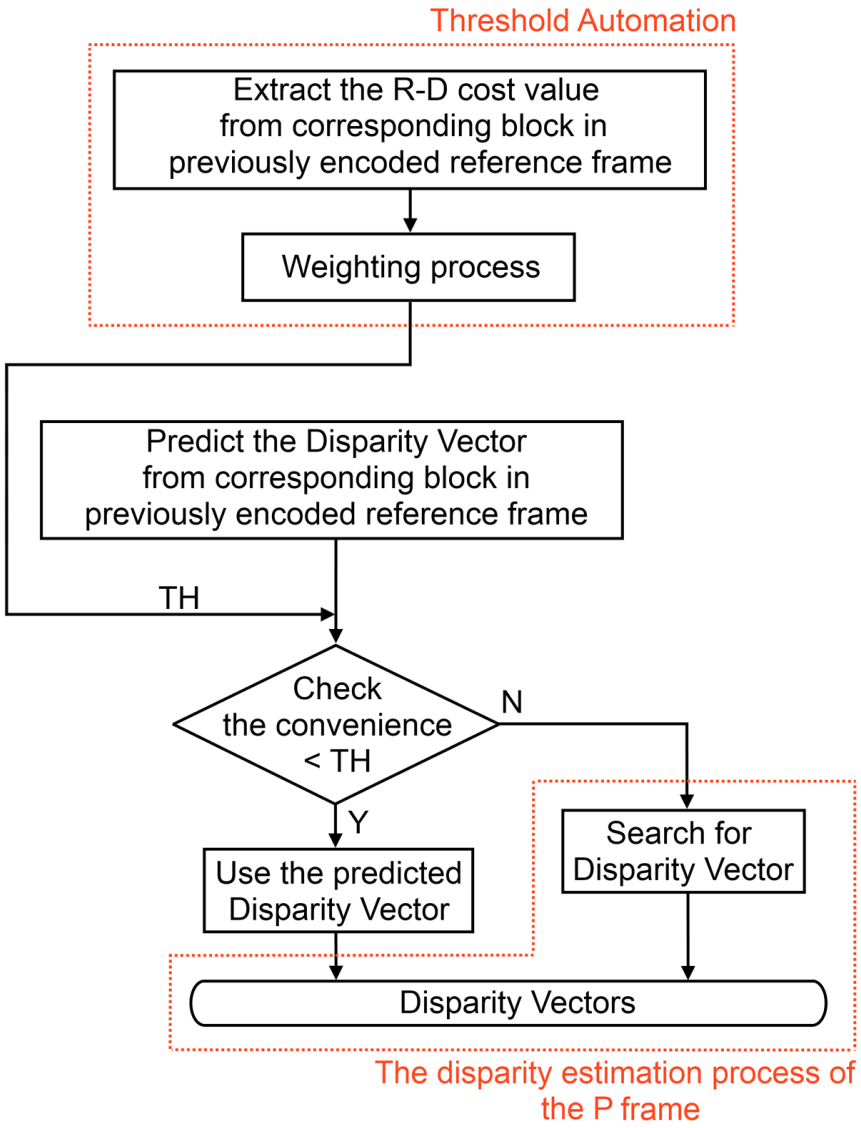


Figure 6.2: The flow chart of the disparity estimation process of the D_P frame with dynamic threshold value calculation.

process is to automatically find this certain TH which varies depending on the nature of the image set and the QP.

In order to increase the accuracy and the efficiency of the encoding

process, the TH should be calculated block by block instead of applying a single TH to the whole prediction scheme, which was the case in the previous chapter. Since the D_P frame is utilizing the disparity vector information from the blocks of previously encoded frames, the reference frames must be encoded before we can begin to encode the D_P frame.

6.2.1 Proposed solution

As stated before, the encoder finds the most suitable mode by comparing the RD cost values. Once the mode is decided, the disparity vector for the block is set and encoding continues with the transformation process. At this point, the RD cost value of the block is only used for mode decision. However, I discovered that it can be used as a parameter to find the TH since it is QP dependent and gives information about the encoding performance of that particular block. For this purpose, the RD cost values of the blocks in the reference frame are extracted besides the disparity vectors. These values are to be used as THs for the corresponding blocks in the D_P frame after a certain weighting process as it is shown in the flow chart of the disparity estimation process of the D_P frame in Figure 6.2.

The weighting process is a necessary step to derive the THs from the extracted RD cost values [3]. A small RD cost value of a block in a P frame, which for instance typically occurs in backgrounds, indicates a good match between the reference block and the candidate block, confirming the success of the encoding. Therefore, this small cost value should be weighted by a large coefficient because the encoder is expected to find also a good match between the corresponding block in the D_P frame and its reference block in another frame. On the other hand, blocks with large RD cost values, which for instance occur in occlusions, need to be weighted by a small coefficient to have smaller THs, in order to restrict possible deterioration caused by derived disparity vectors that point to irrelevant reference blocks.

6.2.2 Weighting Functions

The number of different possible weighting functions aiming to generate higher THs for lower RD cost values and lower THs for higher RD cost values is practically unlimited. Here, I propose two different weighting functions underlining the fact that there is an inverse relation between the RD cost values and the weighting coefficients. The proposed weighting functions, called linear weighting function (LWF) and quadratic weighting function (QWF), are shown in Figure 6.3 and their equations are as follows;

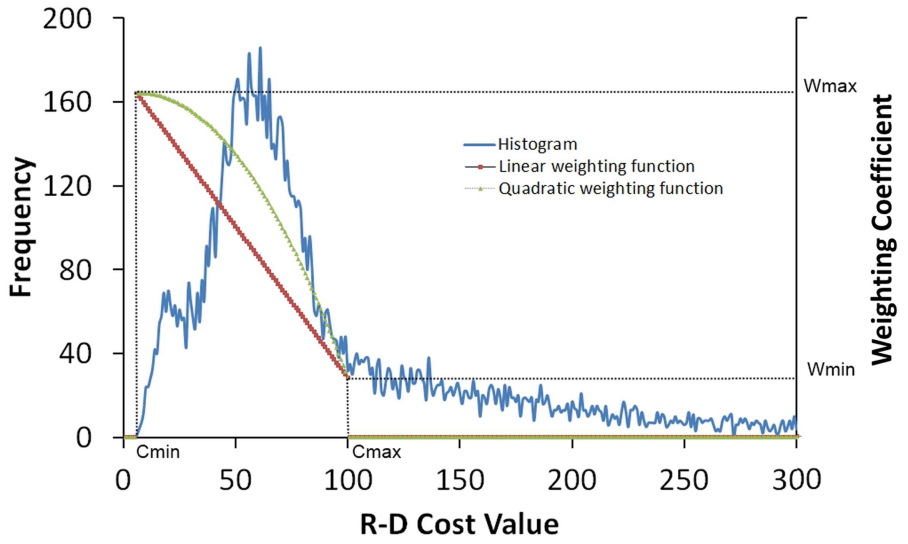


Figure 6.3: Histogram graph of the RD cost values extracted from a P frame in the prediction scheme (left ordinate) and two different weighting functions (right ordinate).

LWF can be described as;

$$f(x) = \begin{cases} 0 & , \quad x < C_{min} \text{ or } x > C_{max} \\ \frac{W_{max}(C_{max}-x)+W_{min}(x-C_{min})}{(C_{max}-C_{min})} & , \quad C_{min} \leq x \leq C_{max} \end{cases} \tag{6.1}$$

QWF can be described as;

$$f(x) = \begin{cases} 0 & , \quad x < C_{min} \text{ or } x > C_{max} \\ W_{max} + \frac{W_{min}-W_{max}}{(C_{max}-C_{min})^2}(x - C_{min})^2 & , \quad C_{min} \leq x \leq C_{max} \end{cases} \tag{6.2}$$

where x is the extracted RD cost value of a block from which the disparity vector is derived.

The TH can then be calculated by,

$$TH = x.f(x) \tag{6.3}$$

The histogram graph of the extracted RD cost values from the P frame is also plotted to explain the weighting process in a better way in Figure

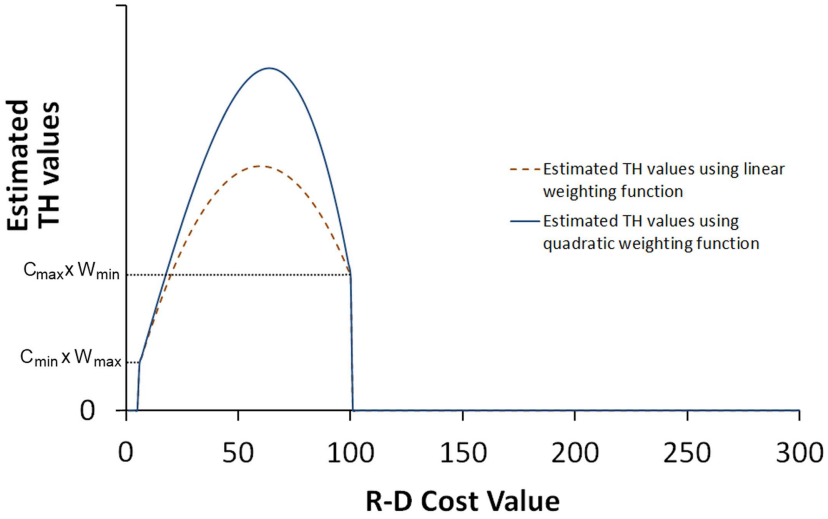


Figure 6.4: Estimated threshold values corresponding to the histogram shown in Figure 6.3.

6.3. While the left ordinate presents the frequency of the histogram graph, the right ordinate indicates the weighting coefficient. The starting and finishing points where the weighting functions are applied are marked as C_{min} and C_{max} in the x-axis which represents the RD cost values. As can be seen from the graph, both weighting functions produce the maximum weighting coefficient for the lowest RD cost value and the value of the coefficient lowers as the RD cost value increases.

Both proposed weighting functions differ from each other in the RD cost values ranging between C_{max} and C_{min}. While the LWF generates gradually decreasing weighting coefficients as the extracted RD cost value increases,the QWF generates higher weighting coefficients for especially lower extracted RD cost values compared to the LWF. This can clearly be seen in Figure 6.4 where the estimated THs corresponding to the histogram graph shown in Figure 6.3 is plotted.

The encoder runs the disparity estimation process for those blocks where the weighting function yields zero, which means that the RD cost value of those blocks does not fall in between C_{min} and C_{max}. For the blocks whose RD cost values lie between C_{min} and C_{max}, both weighting functions are applied and a suitable TH is calculated. Note that the proposed weighting functions may generate different THs for a given RD

Table 6.1: W_{max} parameter values for different QPs.

QP	W_{max}	
	Linear Weighting Function	Quadratic Weighting Function
22	0.9	0.7
27	1.6	1.25
32	1.8	1.5
37	2.3	1.7

cost value. This can be seen in Figure 6.4.

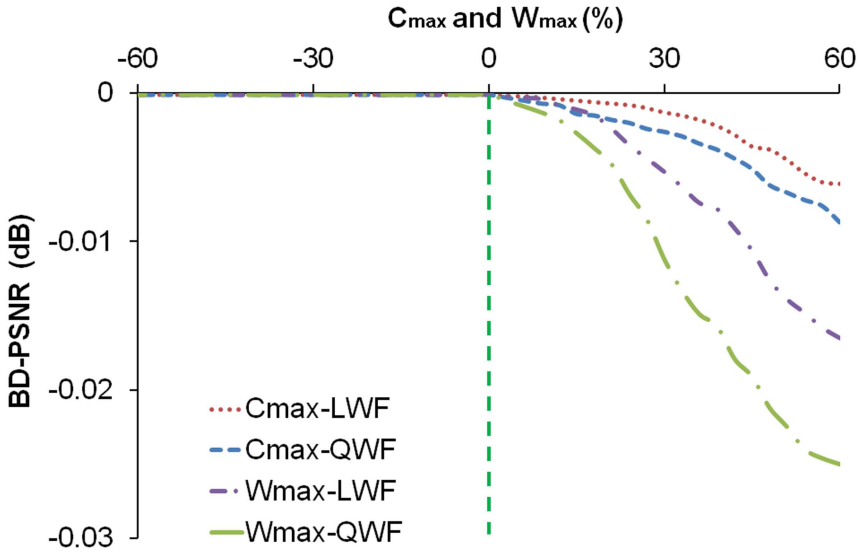
6.3 Experimental results with dynamic threshold calculation

The prediction scheme in Figure 5.10 (CR-MVCS) and the traditional prediction scheme in Figure 4.5 (MVCS-2) are applied to different image sets listed in Table 5.2 with the parameters listed in Table 5.1. Since it has already been proven in the previous section that MVCS-2 is exhibiting better RD performance than MVCS-1 for the current image sets, MVCS-1 is not included in the experiments in this section.

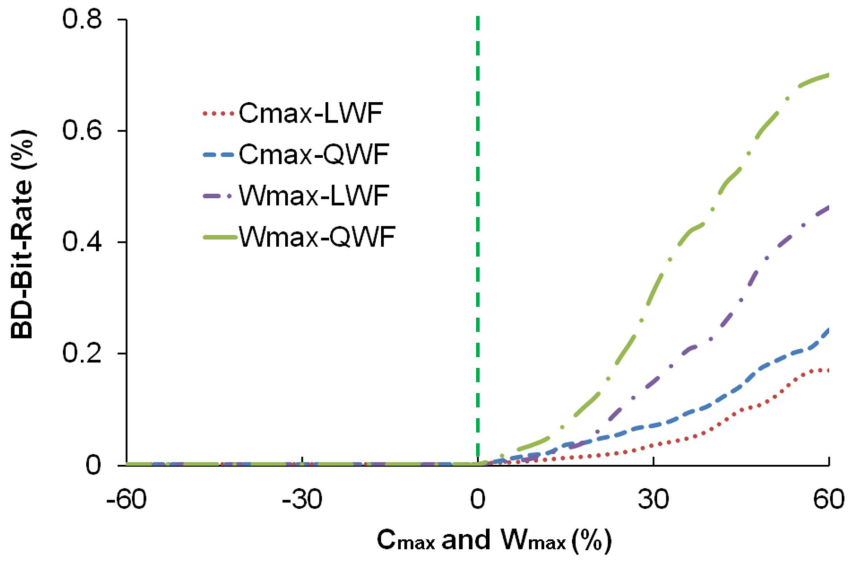
6.3.1 C_{max}, C_{min}, W_{max} and W_{min} value determination

The C_{max} and C_{min} parameter values used to construct the weighting functions are dynamically calculated during the encoding process of the D_P frame by using the extracted RD cost values of the blocks from its reference frame from which the disparity vectors are derived. While the C_{min} value corresponds with the minimum RD cost value, C_{max} is chosen equal to the average of all the RD cost values from the reference frame, which is found experimentally. The parameter corresponding to the minimum weighting coefficient, denoted W_{min}, is set equal to zero in order to have a continuous function. The optimum value of W_{max} is determined for different QPs as listed in Table 6.1, which are calculated by observing the BD difference and the complexity results while changing the W_{max} parameter value.

In order to verify the accuracy of the experimentally determined values



a



b

Figure 6.5: Impact of different C_{max} and W_{max} values on the quality and bitrate of the encoder. In the abscissa of these graphs, the coefficients are varied over ±60% with respect to the experimentally determined values.

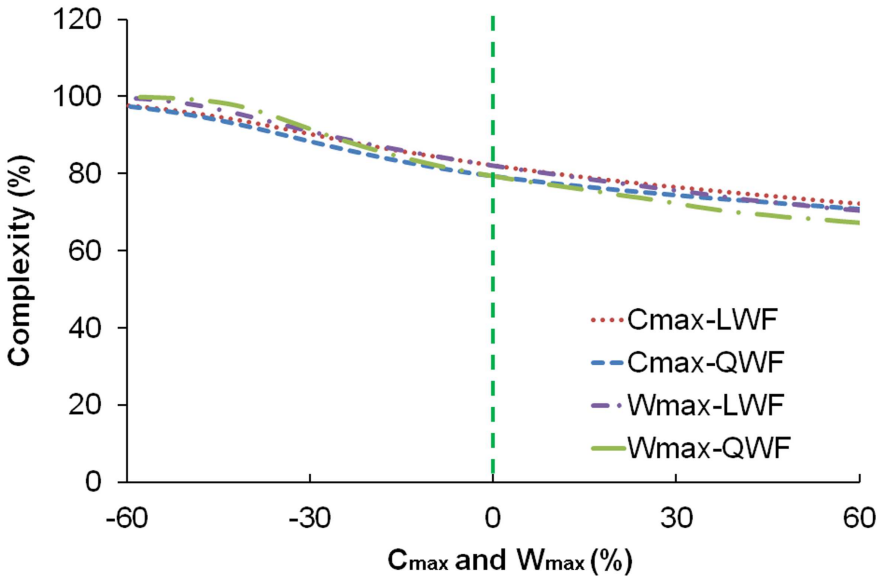


Figure 6.6: Impact of different C_{max} and W_{max} values on the complexity of the encoder.

of C_{max} and W_{max} as mentioned above, various data points lying within ±60% of the respective experimental values are tested on the Amethyst image set. Figure 6.5.a and 6.5.b show the BD-PSNR and BD-Bit-Rate performances between MVCS-2 and CR-MVCS where C_{max} and W_{max} parameters are used to construct the weighting functions in terms of quality and bit-rate respectively.

The complexity results which are obtained while generating the results in Figure 6.5.a and 6.5.b are plotted in Figure 6.6. In order to draw a conclusion on the effectiveness of C_{max} and W_{max}, all three figures from Figure 6.5 and Figure 6.6 need to be interpreted together. On the one hand, it can be concluded from Figure 6.5.a and 6.5.b that the encoded bitstream shows no RD difference (up to four decimal places) compared to the reference prediction scheme if the C_{max} and W_{max} parameters are set equal or lower than their experimentally determined values which correspond to zero in the x-axis in Figure 6.5. On the other hand, however, the complexity of the encoder increases as the C_{max} and W_{max} parameter values decrease (Figure 6.6). Therefore, experimentally determined values of C_{max} and W_{max} yield maximum complexity gain while preserving RD performance of the encoder.

Table 6.2: The complexity and BD performance results of the LWF and QWF for different image sets.

Image Set	Avg. Searched Blocks [%]		BD-Rate [%]		BD-PSNR [dB]	
	Linear	Quad.	Linear	Quad.	Linear	Quad.
Amethyst	82.9	79.4	0.0	0.0	0.0	0.0
The S. Bunny	90.5	86.8	0.0	0.0	0.0	0.0
Euc. Flowers	88.4	84.5	0.0	0.0	0.0	0.0
Tre. Chest	81.3	80.0	0.0	0.0	0.0	0.0
Truck	92.2	88.4	0.0	0.0	0.0	0.0

The experiment shows similar results when the method is applied to any of the other image sets. Note that the results depicted in Figure 6.5 where the x-axis is equal to zero can also be observed in Table 6.2 where the comparative results of all image sets with experimentally determined values of C_{max}, C_{min}, W_{max} and W_{min} are presented.

6.3.2 Comparative Results

The image sets are encoded with the MVCS-2 and the CR-MVCS prediction schemes, utilizing either the LWF or the QWF for several QP values. In addition to the bitrate and quality results of encoding, the number of blocks for which the disparity estimation process is run is also recorded in order to compare the prediction schemes and the weighting functions in terms of complexity. Since the disparity estimation process has to be performed for all blocks in the P frames in the MVCS-2 prediction scheme, the complexity of the MVCS-2 is 100% regardless of the QP. However, the complexity varies for the CR-MVCS prediction scheme because of skipped disparity estimation processes. In order to show the efficiency of the CR-MVCS with weighting functions, the number of searched blocks of all the frames are calculated for different QPs and averaged. As it can be concluded from the results shown in Table 6.2, the CR-MVCS prediction scheme utilizing either LWF or QWF is showing the same RD performance as the MVCS-2 prediction scheme since the BD-Bit-Rate and BD-PSNR values are equal to zero. However, the CR-MVCS prediction scheme requires lower complexity to achieve the same RD results, which

is the main advantage of having the D_P frames instead of P frames in the prediction scheme. As can been seen from the Table 6.2 that in case of the Amethyst image set, only 79.4% of all the blocks were searched for disparity vector, which implies the complexity gain of 20.6% for the overall disparity estimation process of the encoder.

It is also clear from the Table 6.2 that the QWF yields better complexity results compared to the LWF. The QWF is more efficient because the generated weighting coefficients are greater than or equal to the one obtained from a LWF. This results in a higher TH which enables us to skip a relatively larger number of blocks yielding higher complexity efficiency. On the other hand, there exists a trade-off between the complexity gain and RD performance. For a very high value of the TH the encoder will run the disparity estimation process for fewer blocks but the RD performance will degrade. However, it should be noted that the weighting functions are not limited to the two implementations which are presented in this work and different weighting functions can be designed. In this case, appropriate parameters to construct the weighting function should also be investigated.

View images encoded with the CR-MVCS and MVCS-2 prediction schemes were also compared visually and did not show any differences in quality. In Figure 6.7, the images coming from the top right camera view of all image sets are shown after encoding and subsequent decoding. The MVCS-2 prediction scheme, which comprises only I and P frames, is compared with the CR-MVCS prediction scheme for both the LWF and QWF, constructed using the experimentally determined values of C_{min} and W_{max}.

6.4 Conclusions

The TH plays an important role in adjusting the trade-off between the RD performance and the complexity of the encoder. My primary aim in this work is to generate the same bitstream as the reference scheme without having any difference in the quality while reducing the complexity of the encoder as much as possible. For this purpose, I have proposed an algorithm to estimate the optimum THs for the blocks in a D_P frame for different QPs automatically.

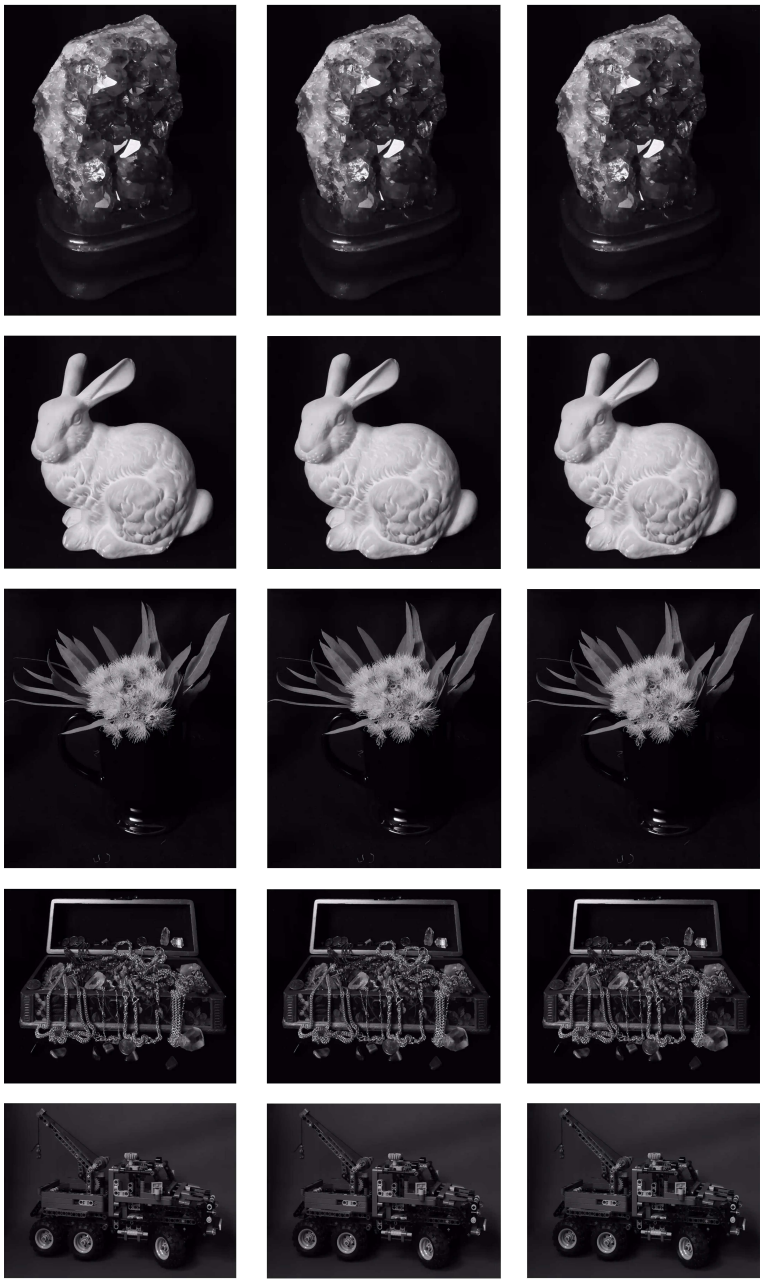
It has been realized that the RD performance of the encoded data starts deteriorating if the TH rises above a certain value. This value is subject to change depending on the imagery content and the QP. I have noticed that the RD cost value of the previously encoded blocks in the

corresponding frame contain both required information and can be used for estimating the optimum THs of the blocks in other frames. In other words, the TH of a block in a D_P frame can be calculated from the RD cost of the corresponding block in the P frame from which the disparity vector is derived. Since there is an inverse relation between the RD cost and the TH, a weighting process needs to be applied to the RD cost value before it can be used as a TH in a D_P frame.

Two different weighting functions named linear and quadratic weighting functions are presented in this chapter. The weighting functions are constructed in such a way that they produce a higher weighting coefficient for small RD cost values and lower weighting coefficients for high RD cost values. During the encoding of the D_P frame, the RD cost of the corresponding block in the P frame is weighted by the coefficient calculated by means of the weighting function. It should be noted that the weighting functions are not limited to the two implementations which are presented in this chapter and different weighting functions can be designed. In this case, appropriate parameters to construct the weighting function should also be investigated.

According to the experimental results, the proposed prediction scheme with the D_P frames shows the same RD performance as the traditional prediction scheme with P frames, and this holds true for both proposed weighting functions. However, the proposed prediction scheme allows to skip the disparity estimation process for some of the blocks, which results in a net complexity gain that is most pronounced in the case of the quadratic weighting function.

The results can further be improved by implementing other algorithms mentioned in Section 4.5 to achieve more complexity reduction for the blocks for which the disparity estimation process is run (blocks whose derived disparity vector is non-convenient) as it is done in a regular P frame.



a b c

Figure 6.7: Decoded image samples of top-right view in the image set with experimentally determined values of C_{min}, W_{max} and QP=32. a) Encoded with MVCS-2 b) Encoded with CR-MVCS by employing linear weighting function c) Encoded with CR-MVCS by employing quadratic weighting function.

References

[1] A. Avci, J. De Cock, P. Lambert, R. Beernaert, J. De Smet, L. Bogaert, Y. Meuret, H. Thienpont, and H. De Smet, "Efficient disparity vector prediction schemes with modified P frame for 2D camera arrays", *Journal of Visual Communication and Image Representation*, vol. 23, no. 2, pp. 287–292, 2012 (cit. on p. 94).

[2] A. Avci, J. De Cock, J. De Smet, Y. Meuret, P. Lambert, and H. De Smet, "A content-adaptive scheme for reduced-complexity, multi-view video coding", in *Three-Dimensional Image Processing (3DIP) and Applications II, Proceedings of the SPIE*, 2012, p. 829 014 (cit. on p. 94).

[3] A. Avci, J. De Cock, J. De Smet, Y. Meuret, P. Lambert, and H. De Smet, "Reduced-complexity multiview prediction scheme with content-adaptive disparity vector estimation", *Journal of Electronic Imaging*, vol. 21, no. 3, p. 033 009, 2012 (cit. on p. 96).

7 Complexity efficient B frame

"Everything that is beautiful and noble is the product of reason and calculation."

— **Charles Baudelaire**

7.1 Introduction

In Section 3.4.1 we first mentioned the B frame. Like the P frame it is an existing frame type in the H.264 video coding standard. The use of B frames in addition to the familiar I and P frames leads to a better RD efficiency. Since MVV suffers from the excessive amount of image data, the prediction schemes constructed with B frames can be a better solution for compressing this huge amount of data more effectively. However, the increased efficiency offered by the B frame comes at the cost of increased complexity. That is to say that, the overall complexity of the prediction scheme with B frames is higher than the prediction schemes with P frames described in the previous chapters.

In this chapter, I will introduce the complexity efficient version of the B frame called D_B frame. It will be explained in detail that the complexity of the B frame in a prediction scheme can considerably be reduced without

affecting the RD performance of the encoder. I will also demonstrate novel prediction schemes constructed with B and D_B frames with which one time instant 5x3 view images can be encoded efficiently. I will then compare the efficiency of these prediction schemes to each other and the prediction schemes that I have shown in the previous chapters.

7.2 D_B frame structure

A B frame provides better RD efficiency than a P frame due to the fact that the B frame allows bi-directional prediction [1]. Because of this, the probability of finding a better match for a candidate block increases since the block will be searched in both reference frames. Moreover, an improved match can be obtained by interpolating between many possible combinations of reference blocks. Besides promising better RD efficiency compared to the other frame types, as a big drawback of the B frame, all these attempts to find a more accurate reference block bring extra load to the encoder and cause a drastic increase in its computational load [2–4]. In case of MVV, the complexity of the encoder is an even more severe problem if the B frames are employed in the prediction scheme like, for example, the one proposed in the literature shown in Figure 4.5.c.

Similar to D_P frame, D_B frame achieves a significant complexity reduction in the encoder by deriving the disparity vectors of the candidate blocks. This derivation is based on the strong geometrical relationship between the views as explained in Section 5.2. Similar to the D_P frame, the convenience of the derived disparity vectors need to be verified. The disparity estimation procedure is skipped for those blocks for which the fidelity of the derived disparity vectors is high enough. The more disparity estimation runs can be skipped, the higher the reduction in the complexity of the encoder.

The implementation of D_B frame is more complex and different from the implementation of D_P frame. The disparity estimation process for a block in a D_B frame runs step by step as follows.

1. The disparity vector and the encoding mode of the candidate block are derived from the previously encoded block in another frame based on the same derivation rules for the D_P frame mentioned in Section 5.6.

2. The RD cost value of the block from which the disparity vector is derived is extracted.

3. A proper TH is calculated by using one of the weighting functions (see Chapter 6) for a given extracted RD cost value in the second step.

4. The RD cost value of the derived disparity vector with the extracted encoding mode is calculated.

5. If the calculated RD cost value is lower than the calculated TH, the disparity estimation process is skipped and the derived disparity vector together with the encoding mode are saved for this candidate block.

6. If not, the regular disparity estimation process as in the B frame is run for the same candidate block.

As can be understood from the above mentioned algorithm steps, skipping the disparity estimation process for a candidate block in a D_B frame is completely dependent on the TH as was the case also in D_P frame. The calculation of the TH for the blocks in the D_B frame will be explained later in this chapter.

7.3 Novel prediction schemes with B and D_B frames

In Figure 7.1.a (ADV-MVCS), I proposed a prediction structure consisting of all available frame types (I, P and B frame types). This prediction structure is an advanced version of the prediction structure that I have already proposed in Figure 5.10. While on the middle row, the B frames are sandwiched in between the I frame and the P frame, they are in between two P frames on top and bottom row of the prediction scheme.

In Figure 7.1.b, I introduced another prediction scheme which is the complexity reduced version of ADV-MVCS called ADV-CR-MVCS. All frame types from the video coding standard and the ones that I proposed in this dissertation (D_P and D_B frame) are employed in this prediction scheme. While there is no difference made in the middle row of the prediction scheme compared to the ADV-MVCS, the D_B frames are located in between the P and D_P frames on the top and bottom row. The ADV-MVCS will be used as the reference scheme for the ADV-CR-MVCS later in this chapter. Again, the arrows in these prediction schemes indicate the reference frames for the regular disparity estimation process and give no information about the geometric derivation of the disparity vectors for the D_P and D_B frames.

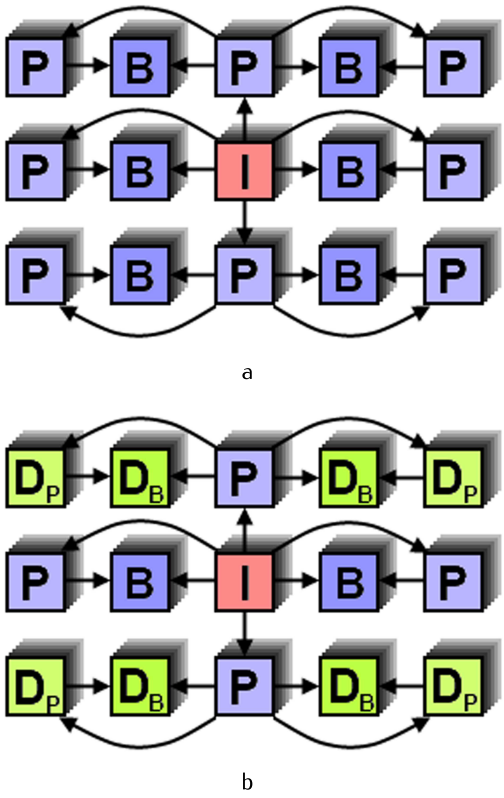


Figure 7.1: Prediction schemes to encode one time instant of multiview videos captured by a 2D camera array. a) The prediction scheme with P and B frames (ADV-MVCS). b) Complexity efficient version prediction scheme of (a) (ADV-CR-MVCS).

The derivation process of the disparity vectors in Figure 7.1 is realized in the same fashion as the prediction scheme in Figure 5.10, which is described in Section 5.6.

7.4 Dynamic TH calculation for the D_B frames

The dynamic TH calculation process is to find the best TH value for a block in the D_B frame. In this process, the QP and the content of the image set is taken into account to obtain a proper TH. This process runs for every block in a D_B frame to increase the efficiency of the derivation process.

Table 7.1: W_{max} parameter values for different QPs.

QP	W_{max}	
	D_P frame	D_B frame
22	0.85	2.20
27	1.15	2.30
32	1.35	2.50
37	1.65	2.60

It has already been proven in the previous chapter that the parabolic weighting function yields better results than the linear weighting function. For this reason, the same parabolic weighting function is employed for calculating the TH of the blocks in a D_B frame dynamically. The W_{max} parameter, which is one of the parameters to construct the weighting function, defines the maximum weighting coefficient corresponding for the minimum extracted RD cost value in the reference frame. Since B frames outperform the P frames mainly due to its structural property of offering bi-directional prediction, the RD cost values of the blocks in the B frame are relatively lower than those in P frames. Therefore, the W_{max} parameter needs to be found and set separately for the D_P and D_B frames in the prediction scheme. The values of the W_{max} parameter for different QPs are listed in Table 7.1 for the blocks in a D_B frame. The values are calculated experimentally by observing the complexity gain and the RD performance of the encoder. Since it is desired to have no deterioration in RD performance of the proposed complexity efficient prediction scheme (ADV-CR-MVCS) in comparison with its reference scheme (CR-MVCS), the W_{max} parameter, thus the weighting function, is selected in such a way that the encoder runs at its minimum complexity while showing the same RD performance. The W_{max} parameter for the blocks in a D_P frame is already given in Table 6.1.

Similar to the dynamic TH calculation process of the D_P frame, the W_{min} parameter which denotes the minimum weighting coefficient is set equal to zero in order to have a continuous function. The C_{max} and C_{min} parameters are dynamically calculated during the encoding. While the C_{max} parameter corresponds to the average of all extracted RD cost values, the C_{min} parameter is chosen equal to the minimum of the extracted RD cost values.

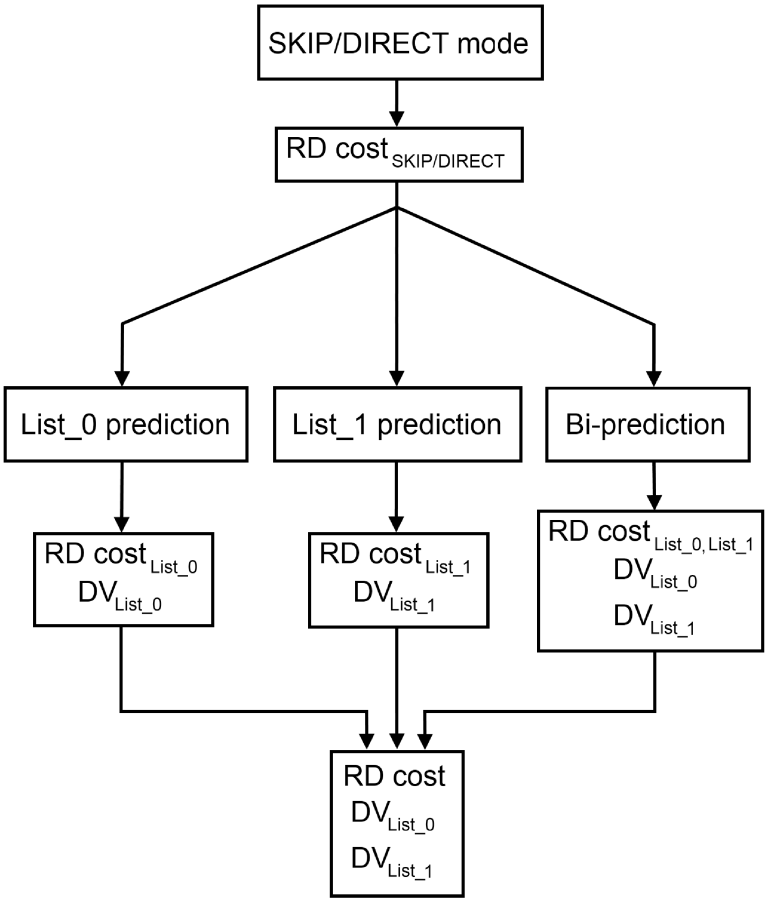


Figure 7.2: Sub-macroblock estimation process.

7.5 Implementation of the D_B frame

I used the JSVM 9.19.7 reference software for the implementation of the D_B frame as it was also the case for the D_P frame. Although a modification of the motion estimation process was enough to implement the D_P frame, many different important functions needed to be revised and customized to implement the D_B frame.

The important part of the modification was mostly realized in the sub-macroblock estimation process where the encoder makes the mode decision for a specific sub-macroblock partition. The size of the sub-macroblock

partition is fixed to 8x8 in this application as mentioned in Table 5.1. In this case, the possible encoding modes for a block are List_0 prediction, List_1 prediction, Bi-prediction and SKIP/DIRECT mode [1]. That means that all the coding options available in the reference software are enabled in this experiment. The List_0 prediction indicates that the prediction is carried out in the reference frame listed in the List_0 index which contains the previously encoded backward frames. In List_1 prediction, the prediction will be done from a previously encoded forward frame listed in the List_1 index. In bi-prediction, an iterative search will be performed on the reference frames from both the List_0 and the List_1 indices to find a more accurate match. In the case of SKIP/DIRECT mode, the block is encoded without transmitting any residual data.

The flow chart of sub-macroblock estimation process is given in Figure 7.2. According to the algorithm, the block is encoded first using the SKIP/DIRECT mode and the RD cost value for this mode is calculated. Then, all the remaining encoding modes for that particular block are tried in search of a better estimation, which is actually the case where the RD cost value of the estimation is lower than that from the SKIP/DIRECT mode. At the end of the process, the encoding mode which gives the minimum RD cost value of all possible modes is selected as the encoding mode for the block and the disparity vector of the best mode is set as the disparity vector of the block.

The process depicted in Figure 7.2 is repeated for every sub-macroblock partition in a B frame. Complexity reduction can be achieved by skipping this burdensome process for some of the blocks, which is the main contribution of the D_B frame. The implementation of the D_B frame is mainly realized in the above mentioned process. Although there are many different ways for implementation, I have chosen to insert the RD cost value, the disparity vector and the encoding mode extraction processes according to the derivation procedure before the sub-macroblock estimation process starts. Then a proper TH is calculated for the extracted RD cost value and this value is compared with the RD cost value of the block which is pointed to the derived disparity vector under the derived encoding mode. The skipping procedure and storing all the relevant information for the candidate block is also done in this code part.

[1] Any frame in a GOP must be encoded first before being a reference for another frame. Such frames are stored in a buffer called Decoded Picture Buffer (DPB) and already encoded frames are listed in this buffer. While the P frames use a single list of reference pictures (List_0), the B frames use two lists (List_0 and List_1). Typically the List_0 contains past pictures and List_1 contains future pictures.

Table 7.2: The complexity and BD performance results of the ADV-MVCS and ADV-CR-MVCS prediction schemes for different image sets.

Image Set	Avg. Searched Blocks (ADV-MVCS) [%]	Avg. Searched Blocks (ADV-CR-MVCS) [%]	BD-Rate [%]	BD-PSNR [dB]
Amethyst	100	71.8	0.0	0.0
S. Bunny	100	79.2	0.0	0.0
Eucalyptus	100	71.7	0.0	0.0
Jelly B.	100	77.9	0.0	0.0
Tre. Chest	100	70.8	0.0	0.0
Truck	100	77.6	0.0	0.0

Apart from the changes realized in the sub-macroblock partition process, some modification in the disparity estimation process was also required in order to calculate the RD cost value of the derived disparity vector.

7.6 Performance comparison

The ADV-MVCS and ADV-CR-MVCS prediction schemes given in Figure 7.1 are used to encode one time instant of view images of the image sets listed in Table 5.2. Since some of the P and B frames in the ADV-MVCS prediction scheme are replaced with their complexity efficient frame types, the ADV-MVCS prediction scheme can be considered as the reference scheme of the ADV-CR-MVCS. In order to make a comparison between the two prediction schemes, the RD and the complexity performances of each view are kept in track for several QPs in the range 22, 27, 32 to 37. In the ADV-CR-MVCS prediction scheme, the parabolic weighting function for both D_P and D_B frames is employed. The construction parameters of the weighting function are experimentally determined and their exact values are already discussed in Section 6.3.1 and Section 7.4.

Experimental results obtained from all the image sets are shown in Table 7.2. The complexity of the encoder is measured by monitoring the

number of blocks in a frame for which the disparity estimation process is run. Since the disparity estimation process has to be performed for all blocks in the P or B frame, the complexity of those frames is depicted as 100%, as shown in the second column in Table 7.2. Note that the numbers related to the complexity performance in Table 7.2 are the averaged numbers for all views in order to avoid cluttering of the table. Note that the complexity of the ADV-MVCS prediction scheme is 100% since it contains only I, P and B frames. It can be seen in the third column of the Table 7.2 that the complexity of the encoder with the ADV-CR-MVCS prediction scheme is reduced significantly. This result also shows the effectiveness of the derivation process based on the existing geometrical correlation between the view images.

As was the case in the previous chapters, the complexity results should be evaluated together with the RD performance of the encoder. The last two columns of the Table 7.2 show the BD differences between the ADV-CR-MVCS and the ADV-MVCS prediction schemes. The BD-PSNR and BD-Rate results are zero up to the first digit following the decimal point for all the image sets, meaning that there is no appreciable RD difference between the ADV-CR-MVCS and ADV-MVCS. As mentioned before, there is a trade-off between the RD performance and the complexity of the encoder. This means that the complexity of the encoder can further be reduced but the RD performance deteriorates. However, the RD performance of the encoder can not be improved further since the ADV-CR-MVCS prediction scheme shows the same RD performance as the ADV-MVCS prediction scheme. This trade-off can completely be controlled by varying the TH, which is possible by changing the construction parameters of the weighting function that is employed.

Apart from the averaged results in Table 7.2, the complexity performances of the D_P and D_B frames in the ADV-CR-MVCS prediction scheme over the P and B frames in the ADV-MVCS prediction scheme respectively are also monitored. This comparison is realized when both prediction schemes are applied to the same image set for the same QPs listed above. As can be seen from the results in Figure 7.3, the complexity performances of the D_P and D_B frames varies over the image set. This is basically due to the content of the image set. If the image set contains more background or less occlusions or illumination effects, the derivation process gives better results and more derived disparity vectors are marked as convenient. In other words, the complexity performance of the proposed complexity efficient frames comes closer to the complexity of the P and B frames if the image set contains less background, more occluded regions or more

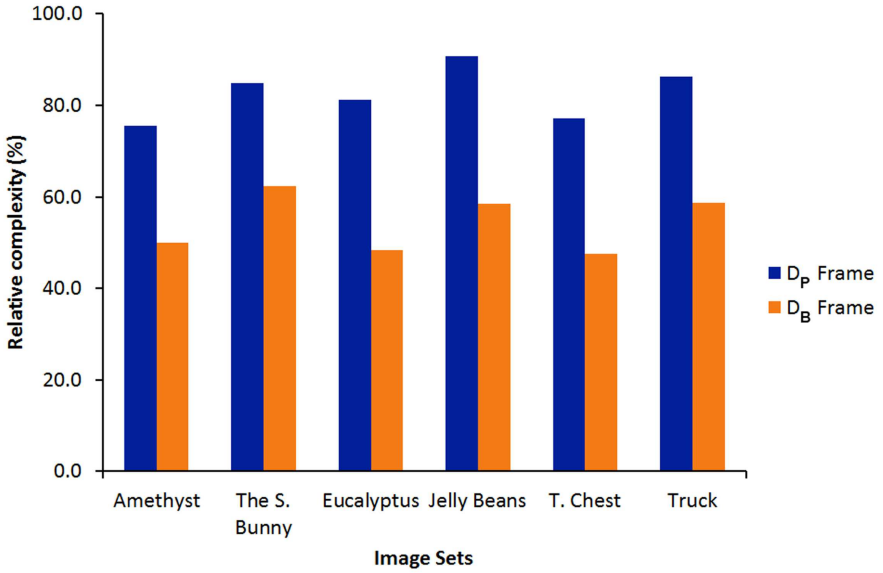


Figure 7.3: Complexity performances of D_P and D_B frames in ADV-CR-MVCS prediction scheme over P and B frames in the ADV-MVCS prediction scheme respectively.

illumination effects. The complexity gain is more visible for the D_B frames since List_1 and Bi-prediction modes are run in addition to the List_0 and SKIP/DIRECT mode, which is the case in the D_P frame.

Average RD results of all the image sets encoded with the MVCS, CR-MVCS, ADV-MVCS and ADV-CR-MVCS prediction schemes are plotted in Figure 7.4. The MVCS prediction scheme is the reference prediction scheme of the complexity efficient prediction scheme called CR-MVCS (in Figure 5.10) in which some of the P frames are replaced with the D_P frame. The prediction scheme in Figure 7.1.a, called ADV-MVCS, is the reference prediction scheme of the novel prediction scheme proposed in this chapter, called ADV-CR-MVCS, which is constructed with D_P and D_B frames in place of some of the P and B frames. Since there is no RD difference between the proposed prediction structures (CR-MVCS and ADV-CR-MVCS) and their reference prediction structures (MVCS and ADV-MVCS), their RD curves are on top of each other in Figure 7.4. The ADV-MVCS and ADV-CR-MVCS prediction structures yield better results than the previously proposed MVCS and CR-MVCS schemes due to the compression efficiency of the B frame and D_B frame which is the complexity efficient

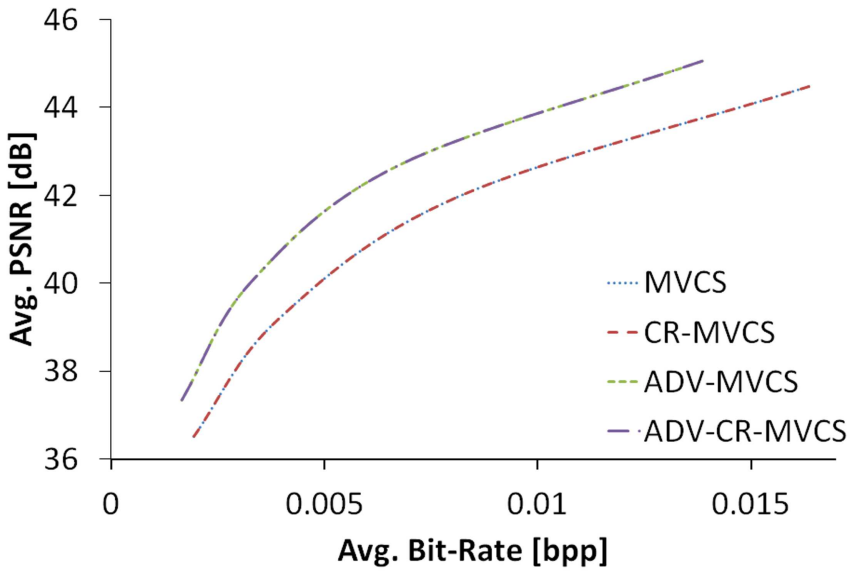


Figure 7.4: Average RD results of all image sets encoded with the MVCS, CR-MVCS, ADV-MVCS and ADV-CR-MVCS prediction schemes. The curves of CR-MVCS and ADV-CR-MVCS are on top of the curve of MVCS and ADV-MVCS respectively.

version of the B frame.

7.7 Conclusions

In the previous chapters, the complexity efficient version of the P frame called D_P frame and the prediction schemes constructed with D_P frames were discussed in detail. In this chapter, the complexity efficient B frame called D_B frame has been introduced and two novel prediction schemes; one containing the well-known frame types from H.264/AVC (ADV-MVCS) and the other one containing the D_P and D_B frames (ADV-CR-MVCS) have been proposed.

In this chapter, the quadratic weighting function has been employed. The parameters to construct the weighting function have been experimentally calculated for the D_B frames while the same construction parameters from the previous chapter are used for the D_P frames.

Experimental results show that the RD performance of the ADV-MVCS and the ADV-CR-MVCS prediction schemes is the same, meaning that

there is no quality and bit-rate difference between the bitstreams generated by means of these prediction schemes. However, from a complexity point of view, the ADV-CR-MVCS prediction scheme has 29.2% (see the experimental results in Table 7.2) lower complexity compared to the ADV-MVCS prediction scheme.

The complexity performance of both D_P and D_B frames in the ADV-CR-MVCS prediction scheme is compared to the corresponding frames in the reference scheme individually. According to the results, considerable complexity gain has been achieved in both proposed complexity efficient frame types. This gain is about 50% without having any deterioration in the RD performance for the D_B frames.

References

[1] F. Pan, X. Lin, S. Rahardja, K. P. Lim, Z. G. Li, D. Wu, and S. Wu, "Fast mode decision algorithm for intraprediction in H.264/AVC video coding", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 7, pp. 813–822, 2005 (cit. on p. 110).

[2] D. Wu, F. Pan, K. P. Lim, S. Wu, Z. G. Li, X. Lin, S. Rahardja, and C. C. Ko, "Fast intermode decision in H.264/AVC video coding", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 7, pp. 953–958, 2005 (cit. on p. 110).

[3] M. Horowitz, A. Joch, F. Kossentini, and A. Hallapuro, "H.264/AVC baseline profile decoder complexity analysis", *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 704–716, 2003 (cit. on p. 110).

[4] M. Jiang and N. Ling, "On enhancing H.264/AVC video rate control by PSNR-based frame complexity estimation", *Consumer Electronics, IEEE Transactions on*, vol. 51, no. 1, pp. 281–286, 2005 (cit. on p. 110).

8
Conclusions

"All's well that ends well."

—**William Shakespeare,** *All's Well That Ends Well*

The number of 3D displays that we come across in our daily life is increasing day by day not only due to the rapid development in the 3D display technology but also due to the advancements in other technology domains such as 3D data acquisition, compression and transmission. All these research domains are interdependent of one another.

For many 3D imaging systems, the multi-view sequences are often obtained by a multiple camera setup which can record the same scene from different perspectives synchronously. As the number of views increases in a multiview video stream, the amount of image data also increases correspondingly, which needs to be compressed (encoded) for more efficient storage and transmission.

The one time instant view images of a multiview video can efficiently be encoded using a prediction scheme with which the inter-view correlations between the views are exploited. However, in this case, the computational load of the prediction scheme remains high and needs to be mitigated.

The periodic arrangement of the cameras results in a strong geometrical relationship among the captured view images. In this dissertation, this spatial redundancy in the view domain is used to our advantage to reduce

the complexity of the multiview video encoder.

The complexity efficient version of the P frame called D_P frame has been presented in Chapter 5. The D_P frame allows us to reduce the complexity of the encoder significantly by skipping the disparity estimation processes for some of its blocks. The skipping process is purely based on the fact that the disparity vector of a block can be derived from a previously encoded block in another frame due to the strong geometrical relationship among views.

Five alternative prediction schemes, three of which were constructed with the D_P frame, have been introduced. The influence of the different locations of I, P and D_P frames on the rate-distortion performance of the encoder has been investigated. Moreover, the success of the disparity vector derivation process and its effects on the coding performance has also been studied.

In the first experiment, a wide range of threshold values were applied globally to all D_P frames in the prediction schemes in order to investigate the effect of the threshold value on the rate-distortion and the complexity performances of the encoder. It has been reported that by employing D_P frames instead of P frames in the prediction schemes, the complexity of the encoder can be reduced drastically. The gain in complexity comes at the expense of only minor losses in quality and/or bit-rate and the trade-off can be tuned by changing the threshold value parameter in the prediction schemes. Since the threshold is an element to adjust the complexity of the encoder, it can be selected depending on the requirements of the user, the exact application and the processing power of the running system.

The experimental results have revealed that the threshold value should be selected depending on the QP and imagery content. In the following experiment, the threshold values were experimentally calculated considering the QP and the content of the image set and applied on an improved version of the scheme 2. It has been reported that both rate-distortion and the complexity results were considerably enhanced. It was also noticed that the complexity of the encoder can be reduced without compromising the quality and the bitrate. It means that the same rate-distortion performance can be obtained with significant complexity reduction.

Since the threshold value is a necessary parameter for a D_P frame, it has to be signalled to the decoder side possibly with the encoded bitstream. In order to avoid this extra burden, a dynamic threshold value calculation method which automatically estimates the optimum threshold values taking the QP and the imagery content into account has been presented. The algorithm estimates the threshold values for every block during

the encoding instead of one fixed threshold value for the whole prediction scheme, which further increases the coding efficiency. In order to calculate the optimum threshold value of a block automatically, the rate-distortion cost value of a previously encoded block from which the disparity vector is derived is utilized. This is due to the fact that the rate-distortion cost value consists of both QP and the imagery content by definition.

Two different weighting functions, named the quadratic and the linear weighting function, underlining the fact that there is an inverse relation between the RD cost values and the weighting coefficients have been proposed in Chapter 6. Different weighting functions than the presented ones can be designed and implemented. In this case, specific construction parameters of the weighting function should be investigated. It has been experimentally demonstrated that dynamic threshold calculation method ensures that the encoder yields maximum complexity gain without deteriorating the rate-distortion performance and without the need of any predefined threshold values. It has also been investigated that the quadratic weighting function outperforms the linear weighting function.

The B frame is another frame type in the video coding standards. It can show better coding performance than the other frame types but brings high computational load to the encoder. In Chapter 7, the complexity efficient version of the B frame called D_B frame and the two more novel prediction schemes have been presented. The quadratic weighting function has been employed in the experiments and its construction parameters are optimized for the D_B frame. According to the experimental results, 29.2% average complexity gain has been achieved without quality and bitrate loss. The complexity performances of the D_P and D_B frames have been evaluated individually and it has been concluded that a drastic complexity gain in a D_B frame can be achieved.

As a drawback, the proposed D_P or D_B frame types require more memory since they utilize information from previously encoded frames during their encoding. This requirement increases with the number of available reference frames that are employed during the derivation process in a D_P or D_B frame. For example, the disparity vector information from four frames, two each at the sides of the I frame, need to be kept in the memory in the case of the prediction scheme presented in Chapter 5.10."

The proposed prediction schemes in this dissertation have been designed for a 5x3 camera matrix but the proposed solutions can easily be extended to larger matrices. In order to avoid longer GOP structures in a prediction scheme, a larger matrix can be divided into sub-camera matrices and each of them can be encoded individually. Moreover, the frames in

the time domain can also be included into the coding process. However, in this case, the complexity of the encoder is expected to be much higher. Nonetheless, it can be an interesting direction of future research.

The resulting bitrate of the proposed prediction schemes can slightly be reduced by not signaling any information for the blocks whose disparity vector is derived. In this case, however, the encoded bitstream becomes non-standard-compliant. That means that a special decoder to regenerate the missing disparity vectors of the blocks is necessary.

The novel prediction schemes that are presented in this work (especially the ones in the Chapter 6 and 7) are compatible for parallelization. In this case, the whole set of multiview images can be encoded by four different processes after encoding the principle frames that will be used in encoding of D_P or D_B frames. Since each process has the same GOP size, their load is fairly balanced.

The ideas in this work can be combined with different methods mentioned in Section 4.5 and more robust and efficient prediction schemes can be formed. That means that the complexity reduction can further be achieved by applying one or more of the proposed algorithms reviewed in Section 4.5 onto the blocks in a D_P or D_B frame for which the disparity estimation process is run (blocks whose derived disparity vector is non-convenient). For example, the complexity of the encoder can be more reduced by employing the mode correlation-based early termination method mentioned in Section 4.5 for the non-convenient blocks. By this way, the disparity estimation process of such a macroblock can be early terminated before running the exhaustive disparity estimation process."

In this work, the disparity vector of a block in a D_P or D_B frame is derived mostly from a single reference frame. However, the derivation efficiency could be increased by taking different available reference frames into account. Although this increases the dependency of a block, more efficient encoding is expected.

Index