CODAGE DES DONNÉES VISUELLES : EFFICACITÉ, ROBUSTESSE, TRANSMISSION

Marco Cagnazzo

HABILITATION À DIRIGER DES RECHERCHES

UNIVERSITÉ PIERRE ET MARIE CURIE

CODAGE DES DONNÉES VISUELLES : EFFICACITÉ, ROBUSTESSE, TRANSMISSION

VISUAL DATA CODING : EFFICIENCY, ROBUSTNESS, TRANSMISSION

présentée et soutenue par
Marco CAGNAZZO

Jury :

M	Pascal FROSSARD	Professeur, EPFL Lausanne	Rapporteur
Mme	Christine GUILLEMOT	Directeur de Recherche, INRIA Rennes	Rapporteur
Mme	Luce MORIN	Professeur, INSA Rennes	Rapporteur
M	Dominique BÉRÉZIAT	Maître de Conférences, HDR, UPMC Paris	Examinateur
M	Patrick LE CALLET	Professeur, Université de Nantes	Examinateur
M	Adrian MUNTEANU	Professeur, Vrije University, Bruxelles	Examinateur

PARIS, 3 SEPTEMBRE 2013

Contents

Acronyms

BMA Block-Matching Algorithm

CQ Context Quantization

CRA Cafforio-Rocca Algorithm

DCT Discrete Cosine Transform

DDE Dense Disparity Estimation

DP Dynamic Programming

DVC Distributed Video Coding

DWT Discrete Wavelet Transform

EBCOT Embedded Block Coding with Optimized Truncation

EZW Embedded Wavelet Zerotree

GG Generalized Gaussian

GOP Group Of Pictures

KF Key Frame

KLT Karhunen-Loève Transform

LS Lifting Scheme

MANET Mobile Ad-hoc Network

MC Motion compensation

MCTI Motion-Compensated Temporal Interpolation

ME Motion Estimation

MPM Most Probable Mode

MRF Markov Random Field

MSE Mean Square Error

MV Motion Vector

MVF Motion vector field

NC Network Coding

PDF Probability Density Function

PMF Probability Mass Function

PNC Practical Network Coding

PSNR Peak Signal-to-Noise Ratio

QMV Quantized Motion Vector

RLNC Random Linear Network Coding

SA-WT Shape-Adaptive Wavelet Transform

SAD Sum of Absolute Differences

SB Subband

SI Side Information

SSD Sum of Squared Differences

TV Total Variation

VLC Variable Length Coding

WZF Wyner-Ziv Frame

Detailed Curriculum Vitae

General information

Name:	Marco
Surname:	Cagnazzo
Birth:	23rd April 1977, Naples, Italy
Nationality:	Italian
e-mail:	cagnazzo@telecom-paristech.fr
Web:	http://cagnazzo.wp.mines-telecom.fr/

Professional status

Position:	*Maître de conférences*
(since 01.02.08)	Associate professor
Title:	*Enseignant-Chercheur en vidéo numérique*
	Teacher-Researcher in Digital Video
Class:	2nd
Institution:	Institut Mines-Telecom – TELECOM ParisTech
Department:	*Traitement des Signaux et Images*
	Signal and Image Processing
Group:	Multimédia
Address:	37/39 rue Dareau 75014 Paris, France

Previous experiences

11.2006 – 01.2008	Post-doc at *Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis* (I3S) UMR-6070 CNRS-Université de Nice-Sophia Antipolis
03.2005 – 10.2006	Assistant professor at *Università "Federico II", Napoli (Italy)*

Education

03.2005 :	Joint PhD from *Université de Nice-Sophia Antipolis* and "Federico II" University of Naples (Italy). Thesis : *Wavelet transform and three-dimensional data compression*
01.2002 :	MSc in Telecommunication Engineering from "Federico II" University of Naples (Italy).

PhD student supervision

Since my arrival at TELECOM-ParisTech, I have co-supervised the following PhD students.

1. Thomas Maugey (from my arrival in February 2008 to the graduation (October 2010);
2. Claudio Greco (September 2008 - July 2012)
3. Abdalbassir Abou El-Ailah (December 2009 - December 2012)
4. Giovanni Petrazzuoli (November 2009 - January 2013)

I am currently (co-)supervising 5 PhD students:

1. Elie Gabriel Mora (start: February 2011)
2. Giovanni Chierchia (start: October 2011)
3. Aniello Fiengo (start: December 2012)
4. Marco Calemme (start: December 2012)
5. Marwa Meddeb (start: February 2013)

Doctorate of Thomas Maugey

I have co-supervised the PhD thesis of Thomas Maugey with professor Pesquet-Popescu since my arrival at TELECOM-ParisTech to the graduation on October 2010.

Dr Maugey has worked on Distributed Video Coding, with particular interest in the estimation of side information. We have been working together mainly on SI generation. We explored methods based on dense motion and disparity fields [27,28,86] (see also Section 3.2), on DVC systems with hashes [87], and iterative refinement [115]. We have also been working on inter-view and temporal side information fusion for multi-view DVC [85]. Working on SI generation, we have observed that often an improvement in SI quality, measured as PSNR with respect to the original Wyner-Ziv frame, does not always translate into an improvement of the global RD performance of the DVC system. Therefore we proposed, implemented and tested other quality metrics then the PSNR for evaluating the SI effectiveness. The findings have been resumed in a paper submitted to IEEE Transactions on Circuits and Systems for Video Technologies (currently we are working at a new version after receiving a "minor revision" decision).

Dr Maugey graduated in October 2010 with the highest marks ("très honorable") and is currently working as post-doctoral fellow at the *École Polytechnique Fédérale de Lausanne*. At the time of graduation he had 13 published conference papers and one published journal paper.

Doctorate of Claudio Greco

The doctorate of Claudio Greco was supervised by professor Pesquet-Popescu and myself from September 2008 to July 2012.

Dr Greco worked on video streaming over unreliable wired and wireless networks. This topic was relatively new in our team, therefore Dr Greco had to start a large part of his work from scratch. Nevertheless, his doctorate has been successful. In our joint work, we designed, implemented and tested new protocols for video streaming on cooperative wired networks [93]; then, in the context of the ANR project "DITEMOI", we developed a protocol for robust video streaming on wireless networks, the ABCD protocol (see also Section 4.1) [60,62]. This protocol allows the streaming of MDC video; in order to perform tests, a simple yet effective MDC technique was proposed [61,65]. The ABCD protocol was later improved taking into account the Congestion/Distortion trade off [63].

The PhD thesis of Claudio Greco was successfully defended in July 2012 ("très honorable"). At that time, the publication record of Dr Greco comprised 5 conference papers (plus 1 accepted and 1 submitted and then accepted) and 2 published journal papers. In the following, Dr Greco has obtained post-doctoral positions in our team (one of them is a "Futur & Ruptures" grant from *Institut Mines-Telecom*).

Doctorate of Abdelbassir Abou El-Ailah

In June 2009 I obtained a PhD fellowship from the "Futur & Ruptures" program of the *Institut Telecom* (now *Institut Mines-Telecom*) with a proposal on distributed video coding. Thanks to our contacts with the Kaslik University (Lebanon), we were able to enroll Abdelbassir Abou El-Ailah, who was therefore co-supervised by our team (professor Pesquet-Popescu, Dr Dufaux and myself) and professor Farah from Kaslik University.

Dr Abou El-Ailah has been working on side information generation using iterative refinement [7–10] and fusion of local and global motion estimation [3, 5], see also Section 3.4. He graduated with the highest marks ("très honorable") in December 2012. At that time he had published 6 conference papers and 2 journal papers. After achieving the PhD, he has kept working with our team as post-doctoral fellow. In this period, we prepared a new journal paper about segmentation-based SI generation [6].

Doctorate of Giovanni Petrazzuoli

In September 2009 I obtained an "Institut Carnot" PhD fellowship for a PhD thesis on distributed video coding. We were able to recruit Giovanni Petrazzuoli, who had just obtained his MSc in Telecommunication Engineering at "Federico II" University of Naples, and had spent 6 months in our team at TELECOM-ParisTech for his engineering graduation project. His PhD thesis was co-supervised by professors Pesquet-Popescu and Poggi ("Federico II" University of Naples), and by myself.

Giovanni Petrazzuoli has been working on SI generation for DVC. A first contribution was based on the idea of using high order trajectories for motion models [10, 112, 113, 115] (see Section 3.3). This method was also used to propose a novel MDC technique [65]. Dr Petrazzuoli has also explored the potential advantages in using DVC for transmitting MVD content [110], in particular in the context of interactive multi-view video streaming [109, 111]. Finally, the last period of the doctorate was devoted to new techniques for DVC in the case of multi-view video, proposing new methods for occlusion detection and management and for DIBR-aided coding. One journal paper is submitted and a second is in preparation [114, 116].

Dr Petrazzuoli graduated in January 2013 with the highest marks ("très honorable") and is currently a post-doctoral fellow in our team.

Current PhD Thesis

Elie Gabriel Mora is pursuing a "CIFRE" PhD thesis with a joint supervision from TELECOM-ParisTech (prof. Pesquet-Popescu and me) and Orange Labs (Dr Jung). His work, started in February 2011, is centered on the upcoming 3D-VC standard [72], see also Section 1.5. Until now, he has published 2 conference articles (two further are to appear in next ICIP and MMSP), 2 standardization contributions for 3D-VC [98]. He has one approved and one pending patent. Finally, he has two submitted journal papers.

Giovanni Chierchia has started his PhD in October 2011 with a "Futur & Ruptures" scholarship that I obtained in June of the same year. He works in tight collaboration with the team lead by professor Pesquet at Paris-Est university. His main research theme is convex optimization with applications to disparity estimation and super-resolution. He also collaborates with the team of professor Poggi at "Federico II" University of Naples on image forensic. He has 4 conference papers and one submitted journal paper.

Marco Calemme has started his doctorate in December 2012 with an "Institut Carnot" grant. He works on multi-view-plus-depth video compression, in particular on depth map representation by elastic deformation of curves. He is co-directed by professor Pesquet-Popescu and Dr Scarpa ("Federico II" University of Naples). *Aniello Fiengo* has begun the PhD program at the same date. He works on rate allocation for HEVC and 3D video coding, and is co-supervised by professor Pesquet-Popescu. *Marwa Meddeb*, also co-supervised by professor Pesquet-Popescu, is pursuing her doctorate with a "CIFRE" contract with a start-up called AMIRIEL. She works on region-based video coding and on optimal rate allocation on HEVC for teleconference applications.

Teaching

After achieving my PhD degree, I have been in charge of the of the *Multimedia Signal Processing* course at the "Federico II" University of Naples (2005-2006 and 2006-2007) and of the *Information Theory* at the University "Parthenope" of Naples (2004-2005).

After joining TELECOM-ParisTech, I have been in charge of the following courses: Compression techniques (2008-2009 – present); Digital video and Multimedia (2008-2009 – present); I am also in charge of the following continuing education short courses: 3D Television; Television for mobile phones; and Digital Television: Systems and Services. Since 2007-2008, I have taught in the following courses: Collaborative Learning Thematic Project; Tools and applications for signals, images and sound; Compression techniques; Image processing and analysis; Advanced methods for image processing; Computer vision; Web Mining; Introduction to image processing (ATHENS); Multimedia Indexing and Retrieval (ATHENS); Short and long student projects ("projet libres" and "stages"); Image and video compression (Continuing education); Video over IP (Continuing education); Signal and image processing (Continuing education). Globally, I have a cumulated teaching record of about 1000 "equivalent hours" (EH). One hour of taught course accounts for 2 EHs, one hour of laboratory or practical works accounts for 0.5 EHs, and one hour of blackboard exercises ("Travaux dirigés") is counted as 1 EH.

A short description of the courses I have been (or I am) in charge of follows.

Information Theory, "Parthenope" University, Naples, Italy (2004-2005). This course (60h) introduces the foundations of information theory to 4th years students in telecommunication engineering. Using the fundamental concepts of entropy and channel capacity, it is shown in a rigorous way, how it is possible to answer to two basilar questions in telecommunication: what the limit of data compression

is; and what the limit of data rate on a lossy channel is. In addition to the information theory notions, this course also introduces some of its main applications: channel coding for error protection, lossless data coding, lossy coding for images and videos. In particular, image and video compression standards are illustrated in the last part of the course.

Multimedia signal processing, "Federico II" University of Naples (2005-2007). This course (60h) illustrates the main techniques for image and video processing. In particular, the focus is on image enhancement and restoration; color image processing; transform, quantization and prediction for image and video compression. I have introduced some new lessons on splines.

Compression techniques, TELECOM-ParisTech (2008-present). In this course (30h) we describe all the main tools for image, audio and video compression: scalar and vector quantization, lossless coding, transform and subband coding, perceptual audio coding, image coding by DCT and wavelet, JPEG and JPEG-2000.

Digital video and multimedia, TELECOM-ParisTech (2009-present). This course (60h) shows the entire life-cycle of multimedia, with particular emphasis on video. In particular, we present the main concepts related to video coding (theory and standards), channel coding and modulation, audio coding, multimedia indexation, streaming, transport and scene composition. Some research-oriented topics are also shown, such us concepts related to distributed video coding, robust coding, ...

Scientific production

My complete scientific production is reported at pages xiii–xviii and includes:

- 14 published journal papers [5, 11, 13, 23–26, 29, 30, 34, 60, 63, 84, 111];
- 5 submitted papers;
- 3 papers in preparation;
- 56 published conference papers, and 3 to appear;
- one book (co-editor) [50];
- 4 published book chapters and 3 further chapters to appear;
- one standardization contribution [98]

Awards. Our paper about Congestion/Distortion optimization [63] has been selected as "High quality paper" by the IEEE MMTC-R Letter board, and included in the January 2013 issue. Two papers appearing in the 2009 IEEE MMSP Workshop have received the "Top 10 % award" [46, 106].

Bibliometrics. According to the *Google Scholar* web site, my H-index is equal to 13 (on August 26, 2013). In Tab.1 there is the list of the 13 most cited articles.

Research projects

I have participated to the following projects:

- LABNET (2001-2002). Low complexity video coding.
- CNRAED (2004-2005). Hyperspectral image compression.
- CPRE 46 04 06 11 (2006-2007). Lossless coding of motion information.
- Secure Media SIM (2007-2008). Secure video coding on SIM card.
- AIBER (2008). Wavelet-based video coding.
- DIVINE (2007-2009). Robust video coding via multiple descriptions.
- DITEMOI (2007-2010). Video streaming on wireless link.
- PERSEE (2009-2013). 2D and 3D video coding (**Scientific responsible for TELECOM-ParisTech**).
- SWAN (2011-2013). Network coding for video.

A few more projects are currently under examination for funding, and for one of them I am the scientific responsible for TELECOM-ParisTech.

Moreover I have obtained scholarships for 4 PhD thesis (2 "Futur & Ruptures", A. Abou-El Ailah and G. Chierchia; 2 "Institut Carnot", G. Petrazzuoli and M. Calemme) and one grant for a post-doc ("Futur & Ruptures" program, C. Greco).

		Citations	year
1	A model-based motion compensated video coder with JPEG2000 compatibility. *M Cagnazzo, T André, M Antonini, M Barlaud*. IEEE ICIP	39	2004
2	Region-oriented compression of multispectral images by shape-adaptive wavelet transform and SPIHT. *M Cagnazzo, G Poggi, L Verdoliva, A Zinicola.* IEEE ICIP	32	2004
3	Scalable context-based motion vector coding for video compression. *V Valentin, M Cagnazzo, M Antonini, M Barlaud.* Proceedings of Picture Coding Symposium	24	2003
4	High order motion interpolation for side information improvement in DVC. *G Petrazzuoli, M Cagnazzo, B Pesquet-Popescu.* IEEE ICASSP.	23	2010
5	Improved class-based coding of multispectral images with shape-adaptive wavelet transform. *M Cagnazzo, S Parrilli, G Poggi, L Verdoliva.* IEEE Geoscience and Remote Sensing Letters	20	2007
6	A comparison of flat and object-based transform coding techniques for the compression of multispectral images. *M Cagnazzo, G Poggi, L Verdoliva* IEEE ICIP	20	2005
7	Region-based transform coding of multispectral images. *M Cagnazzo, G Poggi, L Verdoliva.* IEEE Transactions on Image Processing	17	2007
8	Fusion schemes for multiview distributed video coding. *T Maugey, W Miled, M Cagnazzo, B Pesquet-Popescu.* EUSIPCO	16	2009
9	Low-complexity compression of multispectral images based on classified transform coding. *M Cagnazzo, L Cicala, G Poggi, L Verdoliva.* Elsevier Signal Processing: Image Communication	16	2006
10	A differential motion estimation method for image interpolation in distributed video coding. *M Cagnazzo, T Maugey, B Pesquet-Popescu.* IEEE ICASSP	15	2009
11	JPEG2000-compatible scalable scheme for wavelet-based video coding. *T André, M Cagnazzo, M Antonini, M Barlaud.* Journal on Image and Video Processing	15	2007
12	Costs and advantages of object-based image coding with shape-adaptive wavelet transform. *M Cagnazzo, S Parrilli, G Poggi, L Verdoliva.* Journal on Image and Video Processing	15	2007
13	Improved side information generation for distributed video coding. *A Abou-Elailah, J Farah, M Cagnazzo, B Pesquet-Popescu, F Dufaux.* Visual Information Processing (EUVIP)	14	2011

Table 1: List of most cited publications on August 26, 2013 according to Google Scholar

Other responsibilities

Editorial responsibilities and reviewing

I am an Area Editor for two international journals *Elsevier Signal Processing* and *Elsevier Signal Processing: Image Communication*. Moreover I serve as a reviewer for several journals, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and IEEE TRANSACTIONS ON SIGNAL PROCESSING. I am also a regular reviewer for the major conferences in my area (IEEE ICIP, IEEE ICASSP, EURASIP EUSIPCO, IEEE ICME, IEEE MMSP, IEEE DSP, . . .).

Conference organization

I have been involved in the organization committees of the 2010 IEEE International Workshop on Multimedia Signal Processing (Electronic Media Chair), of the 2011 European Workshop on Visual Information Processing EUVIP (Local arrangement) and of the 2012 EURASIP EUSIPCO (Publicity Chair). Currently I am in the organizing committee of the the 2014 IEEE International Conference on Image

Processing to be held in Paris (Award Chair).

I have co-organized a number of special sessions in scientific conferences:

- EUSIPCO 2010: Special session on Distributed Source Coding: Theory and Applications (co-organized with M. Kieffer)
- DSP 2011: Special session on Multiview and 3D Video Coding (co-organized with B. Pesquet-Popescu)
- WIAMIS 2013: Content Enhancement for Improved Multimedia Applications (co-organized with B. Pesquet-Popescu and F. Dufaux)
- ASILOMAR 2013: 3D content processing (co-organized with B. Pesquet-Popescu and F. Dufaux)

Additional information

I have been a member of IEEE and IEEE Signal Processing Society since 2005 (IEEE Senior Member since February 2011). I am also a member of EURASIP. I speak Italian (mother tongue), English and French.

Publication list

Books

1. F. Dufaux, B. Pesquet-Popescu, and M. Cagnazzo, eds., *Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering*. John Wiley & Sons, Ltd, 2013.

Chapters

1. M. Kaaniche, B. Pesquet-Popescu, and M. Cagnazzo, *Emerging technologies for 3D video: content creation, coding, transmission and rendering*, ch. 5 : Disparity Estimation. Wiley, 2013.

2. M. Cagnazzo, B. Pesquet-Popescu, and F. Dufaux, *Emerging technologies for 3D video: content creation, coding, transmission and rendering*, ch. 6 : 3D Video Representation and Formats. Wiley, 2013.

3. E. G. Mora, G. Valenzise, J. Jung, M. Cagnazzo, B. Pesquet-Popescu, and F. Dufaux, *Emerging technologies for 3D video: content creation, coding, transmission and rendering*, ch. 7 : Depth Video Coding Technologies. Wiley, 2013.

4. B. Battin, P. Vautrot, M. Cagnazzo, and F. Dufaux, *Vidéo 3D : Capture, traitement et diffusion*, ch. 10 : Codage vidéo multi-vues. Hermes, 2013. To appear.

5. E. G. Mora, J. Jung, B. Pesquet-Popescu, and M. Cagnazzo, *Vidéo 3D : Capture, traitement et diffusion*, ch. 12 : Méthodes de codage de vidéos de profondeur. Hermes, 2013. To appear.

6. B. Pesquet-Popescu, M. Cagnazzo, and F. Dufaux, *Elsevier E-References*, ch. Motion Estimation — A Video Coding Viewpoint. Elsevier, 2013. To appear.

7. C. Greco, I. D. Nemoianu, M. Cagnazzo, J. Le Feuvre, F. Dufaux, and B. Pesquet-Popescu, *Electronic Reference Signal Processing, vol. 4*, ch. Multimedia Streaming, pp. 1–80. Elsevier, 2013. To appear.

Journal Papers

1. M. Cagnazzo, F. Delfino, L. Vollero, and A. Zinicola, "Trading off quality and complexity for a low-cost video codec on portable devices," *Elsevier J. Vis. Comm. and Image Repres.*, vol. 17, pp. 564–572, June 2006.

2. M. Cagnazzo, L. Cicala, G. Poggi, and L. Verdoliva, "Low-complexity compression of multispectral images based on classified transform coding," *Signal Proc.: Image Comm. (Elsevier Science)*, vol. 21, pp. 850–861, Nov. 2006.

3. T. André, M. Cagnazzo, M. Antonini, and M. Barlaud, "JPEG2000-compatible scalable scheme for wavelet-based video coding," *EURASIP J. Image Video Proc.*, vol. 2007, pp. Article ID 30852, 11 pages, 2007. doi:10.1155/2007/30852.

4. M. Cagnazzo, S. Parrilli, G. Poggi, and L. Verdoliva, "Costs and advantages of object-based image coding with shape-adaptive wavelet transform," *EURASIP J. Image Video Proc.*, vol. 2007, pp. Article ID 78323, 13 pages, 2007. doi:10.1155/2007/78323.

5. M. Cagnazzo, F. Castaldo, T. André, M. Antonini, and M. Barlaud, "Optimal motion estimation for wavelet video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, pp. 907–911, July 2007.

6. M. Cagnazzo, S. Parrilli, G. Poggi, and L. Verdoliva, "Improved class-based coding of multispectral images with shape-adaptive wavelet transform," *IEEE Geoscience and Remote Sensing Letters*, vol. 4, pp. 566–570, Oct. 2007.

7. M. Cagnazzo, G. Poggi, and L. Verdoliva, "Region-based transform coding of multispectral images," *IEEE Transactions on Image Processing*, vol. 16, pp. 2916–2926, Dec. 2007.

8. M. Cagnazzo, M. Antonini, and M. Barlaud, "Mutual information-based context quantization," *Signal Proc.: Image Comm. (Elsevier Science)*, vol. 25, pp. 64–74, Jan. 2010.

9. M. Agostini, M. Cagnazzo, M. Antonini, G. Laroche, and J. Jung, "A new coding mode for hybrid video coders based on quantized motion vectors," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, pp. 946–956, July 2011.

10. C. Greco and M. Cagnazzo, "A cross-layer protocol for cooperative content discovery over mobile ad-hoc networks," *International Journal of Communication Networks and Distributed Systems*, vol. 7, pp. 49–63, 2011.

11. C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "Low-latency video streaming with congestion control in mobile ad-hoc networks," *IEEE Transactions on Multimedia*, vol. 13, 2012.

12. A. Abou-El Ailah, F. Dufaux, J. Farah, M. Cagnazzo, and B. Pesquet-Popescu, "Fusion of global and local motion estimation for distributed video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, Jan. 2013.

13. G. Petrazzuoli, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu, "Enabling immersive visual communications through distributed video coding," *IEEE MMTC E-Letters*, pp. 17–18, May 2013.

14. T. Maugey, J. Gauthier, M. Cagnazzo, and B. Pesquet-Popescu, "Evaluation of side information effectiveness in distributed video coding," *IEEE Transactions on Circuits and Systems for Video Technology*. To appear. doi:10.1109/TCSVT.2013.2273623.

Submitted Journal Papers

1. E. G. Mora, J. Jung, M. Cagnazzo, and B. Pesquet-Popescu, "Depth video coding based on intra mode inheritance from texture," *APSIPA Transactions on Signal and Information Processing*, Jan. 2013. Submitted.
2. E. G. Mora, J. Jung, M. Cagnazzo, and B. Pesquet-Popescu, "Initialization, limitation and predictive coding of the depth and texture quadtree in 3D-HEVC video coding," *IEEE Trans. Circuits Syst. Video Technol.*, Feb. 2013. Submitted.
3. A. Abou-Elailah, F. Dufaux, J. Farah, M. Cagnazzo, B. Pesquet-Popescu and A. Srivastava, "Fusion of global and local motion estimation using foreground objects for Distributed Video Coding", *IEEE Transactions on Circuits Syst. Video Technol.*, June 2013. Submitted.
4. C. Greco, I. Nemoianu, M. Cagnazzo, and B. Pesquet-Popescu, "A Rate-Distortion Optimized Distributed Social Caching System for Interactive Multi-View Streaming using Network Coding," *IEEE Transactions on Multimedia*, July 2013. Submitted.
5. G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu, "Novel solutions for side information generation and fusion in multiview distributed video coding," *EURASIP Journal of Advances in Signal Processing*, July 2013. Submitted.

Journal Papers in preparation

1. G. Petrazzuoli, T. Maugey, M. Cagnazzo, and B. Pesquet-Popescu, "A novel DVC framework for multiple video plus depth coding."
2. I. Nemoianu, C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "On a Hashing-Based Enhancement of Source Separation Algorithms over Finite Fields for Network Coding Applications."
3. C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "On Reducing Overhead in Practical Network Coding." 2013.

Conference papers

1. M. Cagnazzo, A. Caputo, G. Poggi, and L. Verdoliva, "Codifica video scalabile a bassa complessità," in *Proc. of Didamatica*, (Napoli, Italy), pp. 289–296, Feb. 2002.

2. M. Cagnazzo, G. Poggi, and L. Verdoliva, "The advantage of segmentation in SAR image compression," in *Proceed. of IEEE Intern. Geosc. Rem. Sens. Symp.*, vol. 6, (Toronto, Canada), pp. 3320–3322, June 2002.
3. M. Cagnazzo, G. Poggi, and L. Verdoliva, "Low-complexity scalable video coding through table lookup vq and index coding," in *Proc. of Joint Intern. Workshop on Interactive Distributed Multimedia Systems / Protocols for Multimedia Systems*, (Coimbra, Portugal), pp. 166–175, Nov. 2002.
4. V. Valéntin, M. Cagnazzo, M. Antonini, and M. Barlaud, "Scalable context-based motion vector coding for video compression," in *Proceed. of Pict. Cod. Symp.*, (Saint-Malo, France), pp. 63–68, Apr. 2003.
5. M. Cagnazzo, V. Valéntin, M. Antonini, and M. Barlaud, "Motion vector estimation and encoding for motion compensated DWT," in *Proc. of Intern. Workshop on Very Low Bitrate Video Coding*, (Madrid, Spain), pp. 233–242, Sept. 2003.
6. T. André, M. Cagnazzo, M. Antonini, M. Barlaud, N. Bozinovic, and J. Konrad, "(N,0) motion-compensated lifting-based wavelet transform," in *Proceed. of IEEE Intern. Conf. Acoust., Speech and Sign. Proc.*, (Montreal, Canada), pp. 121–124, May 2004.
7. M. Cagnazzo, T. André, M. Antonini, and M. Barlaud, "A smoothly scalable and fully JPEG2000-compatible video coder," in *Proceed. of IEEE Worksh. Multim. Sign. Proc.*, (Siena, Italy), pp. 91–94, Sept. 2004.
8. M. Cagnazzo, G. Poggi, G. Scarpa, and L. Verdoliva, "Compression of multitemporal remote sensing images through Bayesian segmentation," in *Proceed. of IEEE Intern. Geosc. Rem. Sens. Symp.*, vol. 1, (Anchorage, AL), pp. 281–284, Sept. 2004.
9. M. Cagnazzo, T. André, M. Antonini, and M. Barlaud, "A model-based motion compensated video coder with JPEG2000 compatibility," in *Proceed. of IEEE Intern. Conf. Image Proc.*, vol. 4, (Singapore), pp. 2255–2258, Oct. 2004.
10. M. Cagnazzo, G. Poggi, L. Verdoliva, and A. Zinicola, "Region-oriented compression of multispectral images by shape-adaptive wavelet transform and SPIHT," in *Proceed. of IEEE Intern. Conf. Image Proc.*, vol. 4, (Singapore), pp. 2459–2462, Oct. 2004.
11. M. Cagnazzo, L. Cicala, G. Poggi, and L. Verdoliva, "An unsupervised segmentation-based coder for multispectral images," in *Proceed. of Europ. Sign. Proc. Conf.*, (Antalya, Turkey), Sept. 2005.
12. M. Cagnazzo, G. Poggi, and L. Verdoliva, "A comparison of flat and object-based transform coding techniques for the compression of multispectral images," in *Proceed. of IEEE Intern. Conf. Image Proc.*, vol. 1, (Genova, Italy), pp. 657–660, Sept. 2005.
13. M. Cagnazzo, G. Poggi, and L. Verdoliva, "Costs and advantages of shape-adaptive wavelet transform for region-based image coding," in *Proceed. of IEEE Intern. Conf. Image Proc.*, vol. 3, (Genova, Italy), pp. 197–200, Sept. 2005.
14. M. Cagnazzo, R. Gaetano, S. Parrilli, and L. Verdoliva, "Region based compression of multispectral images by classified KLT," in *Proceed. of Europ. Sign. Proc. Conf.*, (Florence, Italy), Sept. 2006.
15. M. Cagnazzo, R. Gaetano, S. Parrilli, and L. Verdoliva, "Adaptive region-based compression of multispectral images," in *Proceed. of IEEE Intern. Conf. Image Proc.*, (Atlanta, GA), pp. 3249–3252, Oct. 2006.
16. S. Parrilli, M. Cagnazzo, and B. Pesquet-Popescu, "Distortion evaluation in transform domain for adaptive lifting schemes," in *IEEE Workshop on Multimedia Signal Processing*, (Cairns, Australia), pp. 200–205, 2008.
17. S. Parrilli, M. Cagnazzo, and B. Pesquet-Popescu, "Distortion evaluation in transform domain for adaptive lifting schemes," in *Visual Signal Processing and its Application*, (Paris, France), 2008.
18. M. Cagnazzo, T. Maugey, and B. Pesquet-Popescu, "A differential motion estimation method for image interpolation in distributed video coding," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, (Taiwan), pp. 1861–1864, Apr. 2009.
19. S. Corrado, M. Agostini, M. Cagnazzo, M. Antonini, G. Laroche, and J. Jung, "Improving H.264 performances by quantization of motion vectors," in *Picture Coding Symposium*, (Chicago, IL), May 2009.
20. T. Maugey, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu, "Fusion schemes for multiview distributed video coding," in *European Signal Processing Conference*, vol. 1, (Glasgow, Scotland), pp. 559–563, Aug. 2009.
21. M. Cagnazzo, M. Agostini, M. Antonini, G. Laroche, and J. Jung, "Motion vector quantization for efficient low-bitrate video coding," in *SPIE Visual Communications and Image Processing Conference*, vol. 7257, (San Jose, California), Aug. 2009.

22. W. Miled, T. Maugey, M. Cagnazzo, and B. Pesquet-Popescu, "Image interpolation with dense disparity estimation in multiview distributed video coding," in *International Conference on Distributed Smart Cameras*, (Como, Italy), 2009.
23. T. Maugey, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu, "Méthodes denses d'interpolation de mouvement pour le codage vidéo distribué monovue et multivue," in *Colloque GRETSI - Traitement du Signal et des Images*, (Dijon (France)), Sept.2009.
24. M. Antonini, M. Cagnazzo, and M. Oger, "The "secure media SIM" bitstream structure for video encryption and fingerprinting," in *Smart Event*, (Sophia Antipolis), Sep. 2009.
25. I. Daribo, M. Kaaniche, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu, "Dense disparity estimation in multiview video coding," in *IEEE Workshop on Multimedia Signal Processing*, (Rio de Janeiro, Brazil), 2009.
26. S. Parrilli, M. Cagnazzo, and B. Pesquet-Popescu, "Estimation of quantization noise for adaptive-prediction lifting schemes," in *IEEE Workshop on Multimedia Signal Processing*, (Rio de Janeiro, Brazil), 2009.
27. N. Tizon, B. Pesquet-Popescu, and M. Cagnazzo, "Adaptive video streaming with long term feedbacks," in *IEEE International Conference on Image Processing*, (Cairo, Egypt), 2009.
28. M. Cagnazzo, W. Miled, T. Maugey, and B. Pesquet-Popescu, "Image interpolation with edge-preserving differential motion refinement," in *IEEE International Conference on Image Processing*, vol. 1, (Cairo, Egypt), pp. 361–364, Nov. 2009.
29. M. Cagnazzo and B. Pesquet-Popescu, "Perceptual impact of transform coefficients quantization for adaptive lifting schemes," in *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, (Scottsdale, AZ), Jan. 2010.
30. G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu, "High order motion interpolation for side information improvement in DVC," in *International Conference on Acoustics, Speech and Signal Processing*, (Dallas, TX), Mar. 2010.
31. M. Abid, M. Cagnazzo, and B. Pesquet-Popescu, "Image denoising by adaptive lifting schemes," in *European Workshop on Visual Information Processing*, vol. 1, (Paris, France), pp. 1–4, July 2010.
32. M. Cagnazzo and B. Pesquet-Popescu, "Introducing differential motion estimation into hybrid video coders," in *SPIE Visual Communications and Image Processing Conference*, vol. 1, (Huang Shan, An Hui, China), July 2010.
33. G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu, "Fast and efficient side information generation in distributed video coding by using dense motion rep," in *European Signal Processing Conference*, (Aalborg, Denmark), Aug. 2010.
34. M. Abid, M. Kieffer, M. Cagnazzo, and B. Pesquet-Popescu, "Robust decoding of a 3D-ESCOT bitstream transmitted over a noisy channel," in *IEEE International Conference on Image Processing*, (Hong Kong), Sept. 2010.
35. G. Petrazzuoli, T. Maugey, M. Cagnazzo, and B. Pesquet-Popescu, "Side information refinement for long duration GOPs in DVC," in *IEEE Workshop on Multimedia Signal Processing*, vol. 1, (Saint-Malo, France), Oct. 2010.
36. T. Maugey, C. Yaacoub, J. Farah, M. Cagnazzo, and B. Pesquet-Popescu, "Side information enhancement using an adaptive hash-based genetic algorithm in a Wyner-Ziv context," in *IEEE Workshop on Multimedia Signal Processing*, vol. 1, (Saint-Malo, France), Oct. 2010.
37. C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "H.264-based multiple description coding using motion compensated temporal interpolation," in *IEEE Workshop on Multimedia Signal Processing*, vol. 1, (Saint-Malo, France), 2010.
38. M. Cagnazzo and B. Pesquet-Popescu, "Depth map coding by dense disparity estimation for MVD compression," in *IEEE Digital Signal Processing*, (Corfu, Greece), July 2011.
39. A. Abou-El Ailah, J. Farah, M. Cagnazzo, B. Pesquet-Popescu, and F. Dufaux, "Improved side information generation for distributed video coding," in *European Workshop on Visual Information Processing*, (Paris), July 2011.
40. A. Abou-El Ailah, J. Farah, M. Cagnazzo, B. Pesquet-Popescu, and F. Dufaux, "Successive refinement of motion compensated interpolation for transform-domain DVC," in *European Signal Processing Conference*, (Barcelona, Spain), Aug. 2011.
41. C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "Abcd : Un protocole cross-layer pour la diffusion vidéo dans des réseaux sans fil ad-hoc," in *Colloque GRETSI - Traitement du Signal et des Images*, (Bordeaux, France), Sept. 2011.
42. A. Abou-El Ailah, J. Farah, M. Cagnazzo, B. Pesquet-Popescu, and F. Dufaux, "Amélioration pro-

gressive de l'information adjacente pour le codage vidéo distribué," in *Colloque GRETSI - Traitement du Signal et des Images*, (Bordeaux, France), Sep. 2011.
43. G. Petrazzuoli, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu, "Using distributed source coding and depth image based rendering to improve interactive multiview video access," in *IEEE International Conference on Image Processing*, vol. 1, (Bruxelles, Belgium), pp. 605–608, 2011.
44. G. Petrazzuoli, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu, "Wyner-Ziv coding for depth maps in multiview video-plus-depth," in *IEEE International Conference on Image Processing*, vol. 1, (Bruxelles, Belgium), pp. 1857–1860, 2011.
45. C. Greco, G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu, "An MDC-based video streaming architecture for mobile networks," in *IEEE Workshop on Multimedia Signal Processing*, vol. 1, (Hangzhou, China), Oct. 2011.
46. I. Nemoianu, C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "A framework for joint multiple description coding and network coding over wireless ad-hoc networks," in *International Conference on Acoustics, Speech and Signal Processing*, (Kyoto, Japan), 2012.
47. A. A.-E. Ailah, F. Dufaux, M. Cagnazzo, B. Pesquet-Popescu, and J. Farah, "Successive refinement of side information using adaptive search area for long duration GOPs in distributed video coding," in *International Conference on Telecommunications*, (Beirut), Apr. 2012.
48. E. Mora, C. Greco, B. Pesquet-Popescu, M. Cagnazzo, and J. Farah, "Cedar: An optimized network-aware solution for P2P video multicast," in *International Conference on Telecommunications*, (Beirut), 2012.
49. E. Mora, J. Jung, M. Cagnazzo, and B. Pesquet-Popescu, "Codage de vidéos de profondeur basé sur l'héritage des modes intra de texture," in *Compression et Représentation des Signaux Audiovisuels*, vol. 1, (Lille, France), pp. 1–4, 2012.
50. G. Valenzise, G. Cheung, R. Galvao, M. Cagnazzo, B. Pesquet-Popescu, and A. Ortega, "Motion prediction of depth video for depth-image-based rendering using don't care regions," in *Picture Coding Symposium*, vol. 1, (Krakow, Poland), pp. 1–4, 2012.
51. A. Abou-El Ailah, F. Dufaux, M. Cagnazzo, and J. Farah, "Fusion of global and local side information using support vector machine in transform-domain DVC," in *EUSIPCO*, vol. 1, (Bucharest, Romania.), Aug. 2012.
52. C. Greco, I. Nemoianu, M. Cagnazzo, and B. Pesquet-Popescu, "A network coding scheduling for multiple description video streaming over wireless networks," in *European Signal Processing Conference*, vol. 1, (Bucharest, Romania), pp. 1915–1919, Aug. 2012.
53. A. Abou-Elailah, G. Petrazzuoli, F. Dufaux, J. Farah, M. Cagnazzo, and B. Pesquet-Popescu, "Side information improvement in transform-domain distributed video coding," in *SPIE Application of Digital Image Processing XXXV*, (San Diego, California), Aug. 2012.
54. I. Nemoianu, C. Greco, M. Cagnazzo, and B. Pesquet-Popescu, "Multi-view video streaming over wireless networks with rd-optimized scheduling of network coded packets," in *SPIE Visual Communications and Image Processing Conference*, (San Diego, CA (USA)), Nov. 2012.
55. I. Nemoianu, C. Greco, M. Castella, B. Pesquet-Popescu, and M. Cagnazzo, "On a practical approach to source separation over finite fields for network coding applications," in *International Conference on Acoustics, Speech and Signal Processing*, (Vancouver, Canada), May 2013.
56. G. Petrazzuoli, C. Macovei, I.-E. Nicolae, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu, "Versatile layered depth video coding based on distributed video coding," in *International Workshop on Image and Audio Analysis for Multimedia Interactive Services*, July 2013
57. E. G. Mora, J. Jung, M. Cagnazzo, and B. Pesquet-Popescu, "Modification of the merge candidate list for dependent views in 3D-HEVC," in *Proceed. of IEEE Intern. Conf. Image Proc.*, (Melbourne, Australia), Sept. 2013.
58. E. G. Mora, J. Jung, M. Cagnazzo, and B. Pesquet-Popescu, "Modification of the disparity vector derivation process in 3D-HEVC," in *Proceed. of IEEE Worksh. on Multimedia Sign. Proc.*, (Cagliari, Italy), Sept. 2013.
59. G. Petrazzuoli, T. Maugey, M. Cagnazzo, and B. Pesquet-Popescu, "A novel DVC framework for multiple video plus depth coding," in *Proceed of ASILOMAR*, Nov. 2013.

Standardization

1. E. Mora, B. Pesquet-Popescu, M. Cagnazzo, and J. Jung, *Modification of the Merge Candidate List for Dependant Views in 3DV-HTM*. ISO/IEC - ITU, Shangai, PRC, October 2012. Document JCT3V-B0069 for Shanghai meeting (MPEG number m26793).

Thesis

1. M. Cagnazzo, *Wavelet Transform and Three-dimensional Data Compression*. PhD thesis, Università Federico II di Napoli (Italy) and Université de Nice-Sophia Antipolis (France), Mar. 2005.

Introduction

This manuscript resumes my research activity since the achievement of the PhD degree (March 2005). During these years, I have been working in different teams (CNIT National Laboratory on Communications, Naples; "Federico II" University of Naples; I3S Laboratory, Nice; Multimedia team, TELECOM-ParisTech) and on different projects; however I have been almost always working on compression of visual data (image and video), the only relevant exception being the theme of video streaming.

One of the main target of image and video coding is to obtain good rate-distortion (RD) performances. Therefore, it is not surprising that much effort has been devoted to this classical problem, whose results are shown in Chapters 1 and 2. One of the driving ideas behind the proposed techniques is to inject into the encoding algorithm as much as possible *a priori* information about the signal. More precisely, as far as the video is concerned, we explored several *motion models*, since motion is one of the main sources of information for this kind of signals. For images, we consider an object-based model, which drove us to study adaptive image compression techniques.

However, RD performance is not the only important feature of a compression system; as soon as multimedia fruition becomes more and more common, new functionalities are required. One of them, related to the growing diffusion of mobile, wireless devices is the robustness with respect to losses. A second important feature is the one of low-complexity encoding. The second part of this manuscript addresses these two topics.

In the following, the content of this manuscript is detailed. **The first Chapter** is related to "classical" (*i.e.*, RD-related) video compression topics. Much attention has been devoted to the problem of an efficient representation of *motion*. A first approach (see Section 1.1) is based on a reduced-precision representation of motion vectors, allowing a finer regulation of the rate allocation between motion data and transform coefficient. In opposition to this approach, we have tried to improve the quality of the motion representation, by using a dense motion vector field (Section 1.2). A different method is explored in the following (Section 1.3): the lossless representation of motion vector fields. The motion representation is also a key aspect in wavelet-based video compression. We have introduced a motion estimation criterion that is optimized for this framework in Section 1.4. The final part of Chapter 1 is devoted to the recent work on 3D video compression.

Chapter 2 addresses image compression. The main idea explored therein is that images are composed of homogeneous objects, separated by regular contours. We try to implement compression techniques that are aware of this characteristic. A first approach, described in Section 2.1 is based on the segmentation of the signal into objects, followed by a shape-adaptive transform and a shape-adaptive coding algorithm. This approach has good RD performances, but only on some specific class of images. A complementary solution consists in keeping the same signal support, and modifying the transform instead: in particular we analyze the adaptive wavelet transform implemented via lifting schemes (Section 2.2).

Another important theme has been the one of distributed video coding (DVC), explored in **Chapter 3**. Here, the main problem is the generation of the side information. Several methods, based on an effective representation of motion, have been proposed, implemented and tested. We consider methods based on dense motion vector fields (Section 3.2), on high order trajectory interpolation (Section 3.3), on the fusion of local and global motion representations (Section 3.4). We have also considered the application of DVC to other problems, such as the one of multiple description coding (Section 3.5) and the one of interactive multiview streaming (Section 3.6).

In Chapter 4 we report the research results devoted to robust video streaming over networks. We developed a protocol (Section 4.1) for the generation and the management of an *overlay network* for video streaming over MANETs. The overlay network is composed by multicast distribution trees, and allow to diffuse the video content to all the network nodes. This protocol has then been improved by the addition of a congestion-distortion optimization: this feature allows a remarkable reduction of the delay in video display (Section 4.2). The last part of the Chapter is devoted to our contributions based on the network coding (NC) paradigm (Section 4.3): we propose methods to use it jointly with multiple description coding and with multi-view video coding, and finally we propose a method for reducing the NC per-packet overhead using an approach inspired by the blind source separation framework.

Finally, **the last Chapter** of the manuscripts provides conclusions and some perspectives on future works in video compression and transmission, namely based on the concepts of immersive, interactive and proactive communication.

The bibliography and some selected publications (given as annexes) complete this document.

1 Optimization of video coding

In this Chapter we present several methods proposed to improve the rate-distortion performance of video coding schemes, mainly related on the representation of motion. A first contribution (in Section 1.1) is about the quantization of motion vectors in a hybrid video encoder like H.264/MPEG-4 AVC: the basic idea is that a finer control on motion vector rate increases the number of possible rate allocation choices, potentially providing more efficient solutions. The idea of considering new and more flexible representations of motion is further explored in Section 1.2, where we consider a pel-recursive motion estimation technique that can be used at the decoder side to improve the motion description. The potential advantage of a better motion representation is counterbalanced by the additional signaling cost and the drift between the prediction and the prediction error. The study of an effective motion representation is also at the origin of the third contribution, where the need of a lossless coder for region-based motion vectors led to the improvement of the context quantization step, which is a crucial element in the design of context-based entropy coders (Section 1.3). Another contribution is about motion estimation in the context of wavelet-based video coding. The basic idea is that MSE-related criteria are effective for predictive coding, while in the wavelet-based case the temporal redundancy is exploited via a linear transform. Therefore the criterion for motion estimation should rather take into account the coding gain of the transform. This approach (explored in Section 1.4) was firstly introduced in my PhD thesis, and completed shortly after. Finally, the last part of this Chapter (Section 1.5) is devoted to recent contributions on 3D video compression.

1.1 Quantized motion vectors for low bit-rate video coding

This work was achieved during my postdoc at the I3S laboratory, together with prof. Marc Antonini, two PhD student and dr J. Jung from Orange Labs. The results are published in [11, 22, 44].

An effective representation of motion information has the potential for reducing the coding rate of video. For this reason, many studies have been performed on novel ways to describe and use motion in video [54, 82]. In particular, we consider here the rate allocation trade-off between motion vectors (MVs) and transform coefficients, since this problem has a major impact on compression effectiveness. We introduce a new coding mode that improves the management of this resource allocation. The proposed technique can be used within any hybrid video encoder and introduces a new coding mode, called quantized motion vector (QMV) mode.

The key tool of the new mode is the lossy coding of MVs, obtained via quantization: while the transformed motion-compensated residual is computed with a high-precision MV, the latter is quantized before being sent to the decoder. The MVs are quantized in a rate/distortion optimized way. Several problems have to be faced with in order to get an efficient implementation of the coding mode, especially the coding and prediction of the quantized MVs, and the selection and encoding of the quantization steps. This new coding mode allows to improve the rate-distortion (RD) performances of the hybrid video encoder, as confirmed by tests over several sequences.

1.1.1 A new coding mode

The new coding mode is summarized in Figure 1.1, where we describe the encoder and the decoder operation. The decoder perform a subset of the operations performed by the encoder, and it is highlighted by the blue dashed box. The quantities computed at the encoder and sent to the decoder are highlighted

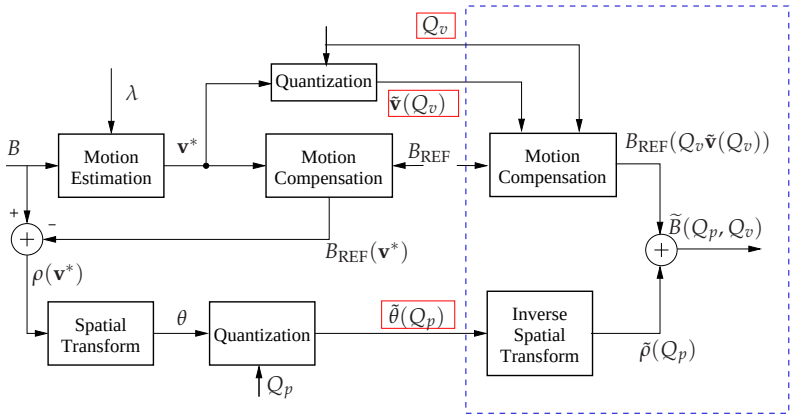


Figure 1.1: The new coding mode operation. In red: the information sent to the decoder, represented by the dashed box. The reference information B_{REF} is known both by the encoder and the decoder.

by a red dotted box. The new coding mode works as follows: first, an accurate (*i.e.* non-quantized) MV $\mathbf{v}^*$ is computed by classical motion estimation, and is used in order to compute the motion-compensated residual $\rho(\mathbf{v}^*)$, for each MB B. This residual is then transformed, quantized with step-size Q_p, and sent to the decoder (after lossless coding), like in all hybrid coders. The difference between this mode and the standard INTER mode is that the MV $\mathbf{v}^*$ is quantized by a simple scalar uniform quantization of its components with a step Q_v. We differ to the Section 1.1.3 the problem of efficiently selecting and encoding Q_v. Knowing the quantization index $\tilde{\mathbf{v}}$ and the quantization step Q_v, the encoder can use them to perform motion compensation on the reference image, which is the same available at the decoder. Adding the residual to the motion-compensated prediction, the decoded block can be computed, exactly as the decoder would do. The encoder can therefore compute the associated distortion, and then the rate, by losslessly encoding Q_v, $\tilde{\mathbf{v}}$ and $\tilde{\theta}$. Now the encoder can take an RD-optimized decision about using the new mode or not. Further details about the RD cost computation are given in the next section.

Figure 1.1 allows to promptly catch the difference between the standard INTER mode and the new QMV: in the former, the same MV $\mathbf{v}^*$ is used to compute the motion-compensated prediction $B_{\text{REF}}(\mathbf{v}^*)$ and the residual $\tilde{\rho}(Q_p)$. In the latter, the MV is quantized before being sent, but the encoder and the decoder still use the same prediction (computed with the quantized vector) and the same residual (computed with the original one $\mathbf{v}^*$). Finally, when the new mode is actually used to encode the current macroblock, the reference frame is updated accordingly. Therefore, there is no drift between the encoder and the decoder. In this sense, the proposed scheme performs a closed-loop prediction.

1.1.2 Computing the cost function for the new mode

As shown in Fig. 1.1, the encoding operation for the new mode begins with a rate-constrained motion estimation: $\mathbf{v}^* = \arg\min_{\mathbf{v}} D_{\text{DFD}}(\mathbf{v}) + \lambda_{\text{ME}} R(\mathbf{v})$. The only differences with respect to the INTER mode are that the Lagrangian parameter is not necessarily the same and that the search grid can be finer. We use the estimated vector $\mathbf{v}^*$ to compute the motion compensated residual $\rho(\mathbf{v}^*) = B - B_{\text{REF}}(\mathbf{v}^*)$. Then, $\tilde{\rho}(Q_p) = T^{-1}\left[\mathcal{Q}^*(\tilde{\theta}, Q_p)\right]$ is the reconstructed residual at the decoder, where T is the direct transform, $\mathcal{Q}^*(\cdot, Q)$ is the de-quantized value for a quantization step Q, and $\tilde{\theta} = \mathcal{Q}\left(T\left[\rho\right], Q_p\right)$ is the quantized transformed residual. The MV $\mathbf{v}^*$ is uniformly quantized with step Q_v, resulting in $\tilde{\mathbf{v}}$ that is sent to the decoder along with Q_v. At the decoder, the MB $\widetilde{B}$ is reconstructed by adding $\tilde{\rho}(Q_p)$, a motion-compensated prediction computed with the quantized vector $\tilde{\mathbf{v}}$, to the residual: $\widetilde{B}(Q_p, Q_v) = B_{\text{REF}}(Q_v\tilde{\mathbf{v}}(Q_v)) + \tilde{\rho}(Q_p)$. Finally, the distortion associated to the QMV mode is:

$$D(Q_p, Q_v) = \left\| B - \widetilde{B} \right\|_\ell = \left\| \rho(Q_v\tilde{\mathbf{v}}(Q_v)) - \tilde{\rho}(Q_p) \right\|_\ell, \tag{1.1}$$

where $\rho(Q_v\tilde{\mathbf{v}}(Q_v)) = B - B_{\text{REF}}(Q_v\tilde{\mathbf{v}}(Q_v))$ is the residual of the motion compensation with the quantized vector; the rate is given by:

$$R(Q_p, Q_v) = R_{\text{mode}} + R[\tilde{\theta}(Q_p)] + R[\tilde{\mathbf{v}}(Q_v)] + R(Q_v). \tag{1.2}$$

Strategy		"Oracle"				"Minsum"			
Rate range		Low		Medium		Low		Medium	
Sequence	Reference	*%R*	*ΔPSNR*	*%R*	*ΔPSNR*	*%R*	*ΔPSNR*	*%R*	*ΔPSNR*
"container"	H264	-5.82	0.26	-11.83	0.41	-6.51	0.31	-9.93	0.34
"container"	H264 $\frac{1}{8}$pel	-5.22	0.24	-4.23	0.15	-5.94	0.29	-2.33	0.08
"foreman"	H264	-3.79	0.19	-3.78	0.18	-3.03	0.15	-0.07	0.00
"foreman"	H264 $\frac{1}{8}$pel	-13.50	0.72	-12.8	0.64	-12.80	0.67	-9.50	0.44
"mobile"	H264	-7.14	0.33	-8.49	0.40	-4.22	0.19	-3.38	0.15
"mobile"	H264 $\frac{1}{8}$pel	-8.03	0.38	-6.00	0.30	-5.15	0.24	-0.87	0.04
"tempete"	H264	-6.74	0.28	-6.62	0.29	-3.87	0.16	-4.20	0.18
"tempete"	H264 $\frac{1}{8}$pel	-8.21	0.36	-6.26	0.28	-5.44	0.24	-3.83	0.17
"city"	H264	-5.04	0.28	-10.04	0.40	-4.73	0.26	-5.60	0.21
"city"	H264 $\frac{1}{8}$pel	-11.39	0.65	-8.55	0.36	-10.98	0.65	-4.08	0.15
"soccer"	H264	-3.92	0.15	-2.99	0.12	-1.74	0.06	-0.01	0.00
"soccer"	H264 $\frac{1}{8}$pel	-12.29	0.47	-7.83	0.33	-10.09	0.36	-4.84	0.20

Table 1.1: Per cent rate savings and differences between PSNR values given by the Bjontegaard metric for the QMV mode at different rates (Q_p between 36 to 39 for medium rates, and Q_p between 39 to 42 for low ones), compared with H.264 and H.264 with 1/8-pel precision (H264 $\frac{1}{8}$pel).

The signaling rate R_{mode} and the coefficient rate $R[\tilde{\theta}(Q_p)]$ are computed as usual via the entropy encoder; the MV rate $R[\tilde{\mathbf{v}}(Q_v)]$ accounts for the quantized vector rate; moreover we observe that in principle a different Q_v could be used in each MB, so we account for its coding cost in Equation 1.2 with the term $R(Q_v)$. In conclusion, the resulting cost function for the QMV mode is: $J_{\text{QMV}}(Q_p, Q_v, \lambda_{\text{mode}}) = D(Q_p, Q_v) + \lambda_{\text{mode}} R(Q_p, Q_v)$. For some assigned Q_v, Q_p and λ_{mode}, we find the cost function for the new mode. This value should be compared with the cost function of the other modes, and if it is the smallest, the QMV mode is selected.

1.1.3 Quantization step selection

Let us indicate with S_Q the set of allowed values for Q_v. Its cardinality, *i.e.* the number of possible quantization steps largely affect the coding performance of the new mode: if too many value of this step are allowed, $R(Q_v)$ could grow too large, and could cancel out any coding gain. To solve this problem, we resort to a *double-pass* coding strategy. We start with a rather dense set S_Q. In a first scanning of the current slice, we gather the estimation of the cost function in $J(Q_v, k)$, where $k \in \{1, 2, \ldots, K\}$ is the MB index. Then we try to represent in an efficient way the whole vector $\mathbf{Q}_v^* = \{Q_v^*(1), Q_v^*(2), \ldots Q_v^*(k), \ldots\}$, where $Q_v^*(k)$ is the best step for the k-th MB. We consider a couple of possible cases, one ideal, the other very simple, but effective:

"Oracle" strategy: The encoder uses the optimal vector $\mathbf{Q}_v^*$, but no bit is accounted for its coding cost. This corresponds to the case of an ideal coding of $\mathbf{Q}_v^*$ (or, to the case of an "oracle" decoder, capable to know the Q_v used for each MB). In other words, in this case we should have $R(Q_v) \approx \log_2 |S_Q|$, but we set $R(Q_v) = 0$.

"Minsum" strategy: We use a single value of Q_v for the whole slice, namely the one minimizing $\sum_k J(Q_v, k)$. In this way the coding cost of Q_v (coded by CABAC or CAVCL) is practically negligible, since it is shared among all the MBs of the slice: $R(Q_v) \approx \log_2 |S_Q|/K$.

1.1.4 Experimental Setup and Results

The QMV mode has been integrated into the H.264/AVC JM-KTA software (v.11.0 KTA 1.4 [74]) with 1/8-pel motion estimation enabled (profile FREXT High); we used full search ME with unrestricted search range, no rate control and a GOP structure of IPP...P. Complete results are reported in Tab. 1.1, where we show the PSNR improvements and the the percent rate savings of the two QMV coders ("Oracle" and "Minsum") with respect to the two "classic" coders H.264 and H.264 with $\frac{1}{8}$-pel ME. We use the Bjontegaard metric [16] to compute the average bit-rate and PSNR variations, considering low rates and medium rates intervals. The proposed mode improves the RD performance in all the tests. The largest improvement are recorded w.r.t H.264 1/8-pel: this was expected since we are observing low-to-medium rates, where the bit-budget needed for high precision MVs is not affordable. When the QMV mode is enabled, the MVs rate and the MVs precision are adapted, and thanks to the quantization step

this precision becomes variable and is optimized according to the RD cost function. On the contrary, with the classical implementation of H.264, the precision of the MVs is fixed. Secondly, we notice that the "Oracle" encoder has in general better performance than the "Minsum" one, but often the difference is small, and sometimes the "Minsum" encoder has better performance. The reason is that the "Oracle" encoder is based on an estimation of the cost function value than can be slightly different from the actual one. As a consequence, the choice indicated by the "Oracle" may be sub-optimal.

1.2 Introducing differential motion estimation in hybrid video coders

This work has been performed at TELECOM-ParisTech, and has been published in [31] (invited paper).

Commonly, motion estimation for video compression is performed using block matching algorithms (BMA). However, it is well known that other ME strategies exist, such as gradient and pel-recursive techniques. These methods produce a dense motion vector field (MVF), which does not fit into the classical video coding paradigm, since it would demand an extremely high coding rate.[1] However, the Caffario-Rocca algorithm (CRA) [19–21] allows to estimate the refined MVF as a function of the prediction error. We have therefore developed a variant of this algorithm that can work within the H.264/MPEG-4 AVC encoder, in order to provide an alternative coding mode for INTER macroblocks The new mode is coded exactly as a classical INTER-mode MB, but the decoder is able to use a modified version of the CRA and then to compute a motion vector per each pixel of the MB. The motion-compensated prediction is then more accurate, and there is room for RD-performance improvement.

1.2.1 The original Caffario-Rocca algorithm

When applying the original CRA, we suppose that we have the current image, indicated as I_k, and a reference one, indicated as I_h. Once a proper scanning order has been defined, the CRA consists in applying for each pixel $\mathbf{p} = (n, m)$ three steps, producing the output vector $\widehat{\mathbf{v}}(\mathbf{p})$.

A priori estimation. The motion vector is initialized with a function of the vectors which have been computed for the previous pixels. For example, one can use the previous pixel's vector, or an average of neighboring vectors.The result of this step is referred to as as $\mathbf{v}^{(0)}$.

Validation. The *a priori* vector is compared to the null vector, computing $A = \left| I_k(\mathbf{p}) - I_h(\mathbf{p} + \mathbf{v}^{(0)}) \right|$ and $B = |I_k(\mathbf{p}) - I_h(\mathbf{p})| + \gamma$. If the prediction error for the current pixel is less than the one for the null vector (possibly incremented by a a positive quantity γ), – that is, if $A < B$ – the a priori is retained as validated vector: $\mathbf{v}^{(1)} = \mathbf{v}^{(0)}$; otherwise, the null vector is retained, that is $\mathbf{v}^{(1)} = \mathbf{0}$.

Refinement. The vector retained from the validation step is refined by adding to it a correction $\delta\mathbf{v}$ obtained by minimizing the energy of first-order approximate prediction error, under a constraint on the norm of the correction. A few calculations show that this correction is given by:

$$\delta\mathbf{v}(\mathbf{p}) = \frac{-e(\mathbf{p})}{\lambda + \|\phi(\mathbf{p})\|^2}\phi(\mathbf{p}) \tag{1.3}$$

where λ is the Lagrangian parameter of the constrained problem; $e(\mathbf{p})$ is the prediction error associated to the MV $\mathbf{v}^{(1)}$, and ϕ is the spatial gradient of the reference image motion-compensated with $\mathbf{v}^{(1)}$. In conclusion, for each pixel $\mathbf{p}$, the output vector is $\widehat{\mathbf{v}}(\mathbf{p}) = \mathbf{v}^{(1)}(\mathbf{p}) + \delta\mathbf{v}(\mathbf{p})$.

1.2.2 Introducing the CRA into H.264

The basic idea is to use the CRA to refine the MV produced by the classical BMA into an H.264 coder. This should be done by using only data available at the decoder as well, so that no new information has to be sent, apart from some signaling bits to indicate that this new coding mode is used.

The operation of the new mode is the following. At the encoder side, first a classical INTER coding for the given partition (e.g. 16×16) is performed and the corresponding cost function $J_{\text{INTER}} = D_{\text{INTER}} + \lambda_{\text{INTER}} R$ is evaluated. Then, *the same encoded information* (namely the same residual) is decoded using

[1] On the contrary, it is quite well suited to the distributed video coding paradigm, as we show in Section 3.2, since in DVC the motion vector are only computed at the decoder.

the CR mode. The RD cost function J_{CR} associated to the mode is computed and the encoder chooses the mode minimizing this cost.

A MB encoded with the CR mode is decoded as follows. The encoded content is identical to the case of an INTER MB (except for a flag indicating the coding mode), therefore the motion vector and the residual are decoded, and we can compute a (quantized) version of the current MB. Moreover, in the codec frame buffer, we have a quantized version of the reference image. We use this information (the INTER MV, the decoded current MB and the decoded reference image) to perform the modified CRA:

A priori estimation. If the current pixel is the first one in the scan order of the MB, we use the INTER motion vector, $\mathbf{v}^{(0)}(\mathbf{p}) = \mathbf{v}_{\text{INTER}}(\mathbf{p})$. Otherwise we can initialize the vector using a function of the already computed neighboring vectors, *e.g.* the previous pixel's vector $\mathbf{v}^{(0)}(\mathbf{p}) = \tilde{\mathbf{v}}(\mathbf{p}_{\text{prev}})$.

Validation. We compare three prediction errors and we choose the vector associated to the least error. First, we have the quantized version of the motion compensated error obtained by first step. Second, we compute the error associated to a prediction with the null vector. This prediction can be computed since we have the (quantized) reference frame, and of the (quantized) current block, decoded using the INTER mode. This quantity is possibly incremented by a positive quantity γ, in order to avoid unnecessary reset of the motion vector. Finally, we can use again the INTER vector. In conclusion, the validated vector is one among $\mathbf{v}^{(0)}(\mathbf{p})$, $\mathbf{0}$ and $\mathbf{v}_{\text{INTER}}(\mathbf{p})$; we keep as validated vector the one associated to the least error.

Refinement. The refinement formula in Eq. (1.3) is modified as follows: $e(\mathbf{p})$ becomes the quantized MCed error; the gradient ϕ is computed on the motion-compensated decoded reference image.

We observe that we only know the quantized version of the motion-compensation error, not the actual one. This affects both the refinement and the validation steps. Moreover, we can compute the gradient only on the decoded reference image. This affects the refinement step. These remarks suggest that the CRA should be used carefully when the quantization is heavy. Finally, the residual decoded in the CR mode is the one computed for the INTER mode, *i.e.* using a different motion field. However, the improvement in vector accurateness leaves room for possible performance gain, as shown in the experiments. Moreover, this is taken into account when computing the coding cost of the mode, which then will be chosen only if it is globally more effective than other modes.

1.2.3 Experimental Results

In a first set of experiments we compared the CR MVs with a MVF obtained by classical block-matching algorithms. For each pair of current-reference images (taken from many sequences), we computed the prediction error energy with respect to the full-search BMA MVF, and the prediction error energy in the case of CR ME. In this case we use the original (*i.e.* not quantized images), and the computation has been repeated for several values of the Lagrangian parameter λ. Excepted the case of very small values of λ, the CR vectors guarantee a better prediction with respect of the BMA. For increasing values of λ the MSE decreases quickly, reaches a minimum and then increases very slowly towards a limit value, corresponding to $\lambda = \infty$. The latter case corresponds to a null refinement.

In order to assess the effectiveness of the proposed method, we have implemented it within the JM H.264 codec. The proposed method seems not to be too affected by the value of the threshold γ provided that it is not too small (usually $\gamma > 10$ works well). Likewise, we found that setting $\lambda = 10^4$ works fairly well in all the test sequences. We considered 8 CIF resolution test sequences, and we computed the Bjontegaard delta rate with respect to the original H.264/MPEG-4 AVC codec. We observed an average rate reduction of 0.3% with gains up to almost 1% for the "flower and garden" sequence. The small improvements are mainly ascribed to the fact the the mode is rarely selected (10% of the blocks in the average).

1.3 Context quantization for lossless coding

This work has been achieved during my post-doc at the I3S Laboratory, in collaboration with professors Antonini and Barlaud. The results have been published in [23].

The initial target of this work was to find out an efficient scheme for lossless coding of complex motion information, such as the MVF resulting from region-based motion estimation [17]. In this frame-

work we considered context based lossless coding in order to take into account high-order and non-linear dependencies between MVs. This approach leads quickly to the *context dilution* problem: having too many contexts makes it difficult or practically impossible to estimate and update the conditional probabilities of symbols during the encoding process. Dealing with the solutions of this problem, we found that the most popular method can be improved since it contains a suboptimal step.

In order to avoid context dilution, all the state-of-the-art algorithms for lossless image compression like CALIC [136], the arithmetic encoder in EBCOT [125], and other popular algorithms proposed in the scientific literature, resort to *context quantization* (CQ). CQ consists in grouping contexts into a relatively small number of conditioning states $c_1, c_2, \ldots, c_F$. Each state, or context class, c_i is made up of one or more contexts x, according to the quantization function which associates a context to a class label. The current input symbol Y is encoded using the probability mass function conditioned to the cluster c_i rather than to the context x. Since the clusters are less numerous than the contexts, the dilution can be kept under control. Of course, the counterpart is that the CQ reduces the mutual information (MI) between the current symbol and the conditioning state. It has been shown that this MI loss affects directly the coding performances, as it is translated into an equal coding rate increase: we remark therefore the importance of designing the best possible context quantizer.

Our contribution to the context quantization problem are the following: first, we propose a more efficient algorithm for the search of local optimal values of the MI loss: MINIMA (Mutual INformation IMprovement Algorithm); second, we generalize the conditions allowing to find the global minimum via dynamic programming (the so-called model-based approach). These two main results are possible thanks to a novel formulation of the CQ problem, which is our third contribution. A brief survey of the state of the art and the experimental results validating the proposed approach complete this section. For more details (in particular for all the proofs) the reader is referred to our paper [23].

1.3.1 Background and Prior Work

We refer the reader to our paper [23] for an exhaustive analysis of the prior work in context quantization. Here we only resume the most influential papers, namely the paper [134] and its extended version [135] by Wu *et al.* who were the first to conveniently analyze the optimal CQ with a given number of classes F and to propose a generic analytical solution. The optimal quantization problem is recognized as a special case of vector quantization, with the Kullback-Leibler divergence to be used as distortion measure. As a consequence, a GLA-like algorithm is proposed to find the optimal CQ, which must minimize the conditional entropy between the current symbol and the quantized context. This approach is then called minimum conditional entropy context quantization (MCECQ). We perform an analysis of the mathematical background of MCECQ, which allows to discover a suboptimal step and then to propose an improved algorithm. As it was already pointed out [66], the paper [134] recognizes that if Y has a binary distribution, then the optimal CQ problem is brought back to a 1D minimization problem, which can be *exactly* solved by means of dynamic programming (DP) techniques. We show here how to extend this approach to M-valued symbols (with $M > 2$) for some specific distributions. We observe that all relevant subsequent papers adopt approaches similar to those of [66] or [134]. This is the case for example of [37] by Chen that develops an analysis of the CQ problem for the lossless coding of WT coefficients and ends up with an algorithm equivalent to MCECQ.

1.3.2 The context quantization problem

In this section we review the problem of CQ and the derivation of the optimal criterion. Let Y be the current symbol of a random process having a discrete alphabet $\mathcal{Y}$. We assume without loosing generality $\mathcal{Y} = \{1, 2, \ldots, M\}$. X is the context, formed by by N already encoded symbols. The alphabet of X is $\mathcal{X} = \mathcal{Y}^N$, thus the number of possible contexts is $|\mathcal{X}| = M^N$. This number can grow up very large, and we risk to incur in the *context dilution* problem: the estimation of $p(Y|x)$, necessary for the encoder to achieve the minimum coding rate, is unreliable. So we consider a classification function $f : x \in \mathcal{X} \rightarrow \{1, 2, \ldots, F\}$ that maps each context in a class label i inducing a partition $\mathcal{C}$ of $\mathcal{X}$ into F classes: $\mathcal{C} = \{c_1, c_2, \ldots, c_F\}$. We use these classes as conditioning information to encode Y, because with a suitable choice of F the estimation of the pmf's $p(Y|c)$ is reliable enough. In the following we consider the problem of optimizing the CQ for a given F. This approach is simpler than a joint optimization of the number of classes and of their composition, but it is popular [37, 66, 79, 134, 135] and sometimes

necessary, as in the case of implementation constraints on the memory or the complexity. On the other hand, a joint optimization has the potential of achieving the best performance.

Given the classification function, we can attain a new entropy bound on the coding cost $H(Y|f(X))$. The loss due to the CQ function f can be therefore measured as $\mathcal{L}(f) = H(Y|f(X)) - H(Y|X) = I(Y;X|f(X))$. The latter identity is true since Y and $f(X)$ are conditionally independent given X. In other words, the residual information between X and Y that is not conveyed by $f(X)$, measures the performance loss ascribable to the CQ function f. Therefore, this function should be chosen so that the loss $\mathcal{L}$ is minimized for a given number of context classes. Introducing the following shorthands, we can write a useful formulation of $\mathcal{L}(f)$. Let us use $p(y,i)$, $p(y|i)$ for $\Pr(Y{=}y, f(X){=}i)$, $\Pr(Y{=}y|f(X){=}i)$, and $p(x,y)$, $p(x|y)$ for $\Pr(Y{=}y, X{=}x)$, $\Pr(Y{=}y|X{=}x)$. After some calculation (see [23]), we find:

$$\mathcal{L}(f) = \sum_{x\in\mathcal{X}} p(x) D\left(p(Y|x) \,||\, p(Y|f(x)) \right). \tag{1.4}$$

We used $D(\cdot\,||\,\cdot)$ to indicate the relative entropy [45]. Defining $d(c_1,c_2) = D\left(p(Y|c_1) \,||\, p(Y|c_2) \right)$, we observe that $d(c_1,c_2) \geq 0$, and we can rewrite (1.4) as

$$\mathcal{L}(f) = \sum_{x\in\mathcal{X}} p(x) d\left(x, c_{f(x)}\right), \tag{1.5}$$

where with a little abuse of notation, in the expression of $d(\cdot,\cdot)$ we indicate with x the class $\{x\}$, and $c_{f(x)}$ is the class containing the context x. We can read (1.5) in this way: the global loss of mutual information due to CQ f, is the average loss that we incur in by substituting the actual context x with its quantized version $c_{f(x)}$.

In order to establish a link with prior works, we observe that our cost function, the mutual information loss (1.5), is equivalent to the loss function definition in [66], and to equation (1) in [134], equation (2) in [135] or equation (6) in [37]. Moreover, it is slightly more general than equation (7) in [79], which only allows to group together contexts with consecutive indexes.

1.3.3 Context quantization properties and the MINIMA algorithm

In the following, we will often refer to the relative entropy between the conditional probabilities functions $p(Y|c_1)$ and $p(Y|c_2)$, referred to as $d(c_1,c_2)$. If $c_1 = \{x\}$ contains only a single context, we will prefer the simpler notation $d(x,c)$ to the more correct one $d(\{x\},c)$, and we will refer to this quantity as *distance* between x and c. We start by computing the effect of the insertion of a context x into a class c. First we observe that inserting a context into a class reduces the context distance from the class.

Proposition 1 *If $x \notin c$ and $c' = c \cup \{x\}$, then*

$$d(x,c') = d(x,c) - \frac{p(c')}{p(x)} d(c',c) - \frac{p(c)}{p(x)} d(c,c'). \tag{1.6}$$

PROOF. See [23]. □

The next proposition allows us to understand the effect on $\mathcal{L}(\mathcal{C})$ of moving a context among classes.

Proposition 2 *When a context x is moved from a class c_1 to c_2, the total mutual information loss varies of:*

$$\Delta = p(c_2) d(c_2,c_2') - p(c_1') d(c_1',c_1) + p(x)[d(x,c_2') - d(x,c_1)], \tag{1.7}$$

where $c_1 = \{x_1, x_2, \ldots, x_N, x\}$, $c_2 = \{z_1, z_2, \ldots, z_M\}$, $c_1' = c_1 - \{x\}$ and $c_2' = c_2 \cup \{x\}$

PROOF. See [23]. □

Of course the Δ resulting from a context displacement can be either positive or negative, *i.e.*, the displacement can worsen or improve the CQ. This is because the effects on the classes c_1 and c_2 have opposite signs, since removing a context from a class allows its pmf to be closer to the pmf's of remaining contexts so that the loss of mutual information becomes smaller; for the symmetrical reason, adding a context to a class increases its contribution to $\mathcal{L}$. Thus, in order to tell whether a context displacement

is suitable or not, we should consider both contributions, and verify that the gains surpass the losses. We remark that the results of propositions 1 and 2 are new; prior works, on the contrary, used directly (1.5) to derive an iterative algorithm for mutual information loss minimization, the MCECQ algorithm (proposed in [134,135]. MCECQ is based on a property for which, with the next proposition, we provide a novel, more explicit demonstration.

Proposition 3 *Let*

$$\begin{aligned} c_1 &= \{x_1, x_2, \ldots, x_N, x\} & c_2 &= \{z_1, z_2, \ldots, z_M\} \\ c_1' &= c_1 - \{x\} & c_2' &= c_2 \cup \{x\} \\ \mathcal{C} &= \{c_1, c_2, c_3, \ldots c_F\} & \mathcal{C}' &= \{c_1', c_2', c_3, \ldots c_F\}. \end{aligned}$$

If $d(x, c_1) \geq d(x, c_2)$, the new context partition obtained by switching x from c_1 to c_2 has a smaller MI loss.

PROOF. See [23]. □

The MCECQ algorithm consists in an iterative scanning of all contexts. For the current context x, its distance from its class $d(x, c_i)$ and from all other classes $d(x, c_j)$ are computed. If for one or more index $d(x, c_j) < d(x, c_i)$, the context x is moved to the "nearest" class. Proposition 3 assures that this algorithm always improves the mutual information loss. However, Eq. (1.7) shows clearly that it is not necessarily the best choice, as it does not take into account the true variation of $\mathcal{L}$, but only a quantity which has the same sign of it.

The proposition 2 suggests as well *how* the MCECQ algorithm should be modified in order to assure the maximal cost reduction at each step: let $c_{f(x)}$ be the class which the context x belongs to according to the current classification function f, let c_k be a generic class, $c'_{f(x)} = c_{f(x)} - \{x\}$ and $c'_k = c_k \cup \{x\}$. Then we define the function δ as:

$$\delta(x, k) = p(c'_{f(x)}) d(c'_{f(x)}, c_{f(x)}) - p(c_k) d(c_k, c'_k) + p(x)[d(x, c_{f(x)}) - d(x, c'_k)] \qquad (1.8)$$

This function gives the mutual information variation associated to the displacement of the context x from its current class to the class c_k. The algorithm that we propose keeps moving the contexts according to the function δ until the relative variation of MI loss is smaller than a given threshold. We call this algorithm MINIMA. We observe that within the inner loop over the class index k, we compare all the F partitions that we would obtain by moving x to each of the classes (by comparing the displacement to the current class with the best displacement so far), and so at the end of loop the best move is kept.

So we have an algorithm that is optimal with respect to the displacement of a given context, but at the same time is "greedy" since it moves the context x before evaluating the effect over the displacement of the others. It is interesting to observe that MINIMA is brought back to the MCECQ if we use the following definition: $\delta(x, k) = d(x, c_{f(x)}) - d(x, c_k)$. We remark that in every single experiment, starting from the same initial conditions MINIMA always achieved better cost function values than MCECQ.

1.3.4 Model-based classification algorithm

As well as its predecessors [37, 79, 134], the proposed algorithm has the problem that it can only find local minima of $\mathcal{L}(f)$. However, if the input alphabet is binary, a DP algorithm exactly solves the minimization problem, *i.e.* it is able to determine the global minimum of $\mathcal{L}(f)$ [134]. In our paper we have found an alternative proof of this property.

In order to verify whether it can be extended to M-ary distributions, we tried to understand in which conditions one can use the DP technique. In our paper we prove that some *sufficient conditions* exist for a set of M-ary conditional probabilities so that the globally optimal classification function could be determined by a 1-D search. These are summarized in the following proposition.

Proposition 4 *If*

1. *the conditional pmf's depend from the context x only by a scalar parameter λ_x: $p(y|x) = q(y, \lambda_x)$;*
2. *for any class c there exists a value λ_c such that the conditional pmf can be expressed as: $p(y|c) = q(y, \lambda_c)$;*

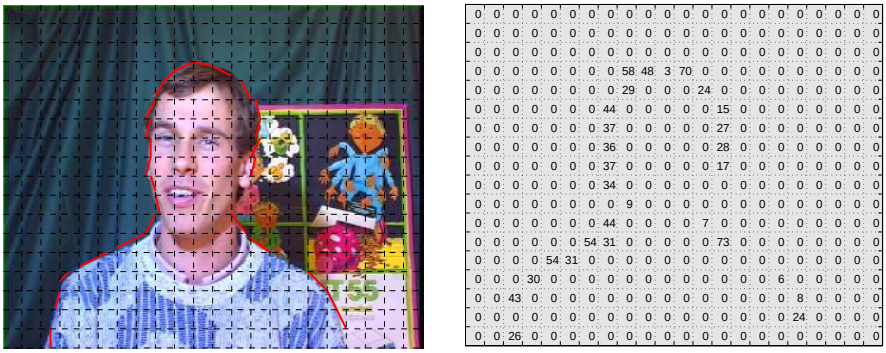


Figure 1.2: A sample segmentation map and its representation as bi-dimensional data set.

3. *the conditional pmf's $q(\cdot,\cdot)$ have relative entropies such that*

$$\lambda_1 < \lambda_2 < \lambda_3 \Rightarrow \begin{cases} D\left(q(y,\lambda_1) \,||\, q(y,\lambda_2)\right) < D\left(q(y,\lambda_1) \,||\, q(y,\lambda_3)\right) \\ D\left(q(y,\lambda_2) \,||\, q(y,\lambda_3)\right) < D\left(q(y,\lambda_1) \,||\, q(y,\lambda_3)\right) \end{cases} \tag{1.9}$$

then, the optimal classification can be found by a DP algorithm, as in the binary case.

PROOF. See [23] □

In our paper we give in an example where this result is applicable.

1.3.5 Experimental Results

Input data. We validate MINIMA for the encoding of the motion segmentation maps produced by a region-based algorithm [17]. An example of segmentation map and the associated label array are in Fig. 1.2. The segmentation map consists in one symbol per block, telling whether the block is split or not. If the block is split, the symbol encodes a straight line contour between the two regions into which the block is divided. In our case, 80 possible splits are possible, which, along with the no-split symbol produce a 81-symbols dictionary. We observe that in this case the statistical dependence among data is strong, but not linear, and even considering a small context (2 or 4 symbols) the number of contexts is quite high: in conclusion we need CQ in order to have an efficient lossless coding.

Coding results We use MINIMA and MCECQ to produce a CQ for the segmentation maps from several test sequences ("eric", "flower and garden", "foreman", "mother and daughter", "paris" and "silent"). The algorithms start from the same initial classification function, chosen randomly, and then the iterative optimization is performed. The probability functions needed to run the algorithms are estimated on a training set obtained from the test sequences[2]. In order to reduce the dependence on the starting configuration, the entire process (random initialization and iterative optimization) is repeated 20 times and the best resulting classification function is retained for both algorithms. We note that in each iteration, MINIMA has reached a lower cost function value than MCECQ. The numerical results show that using MINIMA allows to reduce the final coding rate from up to 3.6 %, with an average reduction of 1.3 %. Further tests show that the proposed algorithm is better than the reference independently from the number of classes.

The model-based algorithm In a second set of experiments, we test the model-based algorithm based on Proposition 4. However, we are aware that this model only loosely matches the characteristics of segmentation maps, and as a consequence we do not expect a large improvement of performances using the model-based algorithm on the segmentation maps. We have validated the model-based algorithm with synthetic data generated so that their statistics exactly match the proposed model. This is of course a favorable case for the model-based algorithm, but anyway it provides some useful insights about the potentialities of the method. In particular, we used an M-ary bi-dimensional random process. The conditional probability distributions have the form $p(y|x) = \lambda_x$ if $y = 0$ and $p(y|x) = \frac{1-\lambda_x}{M-1}$ otherwise. We have found that in this case the model-based algorithm gives a further improvement w.r.t. MINIMA, up to 5 %, and more than 2.5 % in the average.

[2]The data used as training set are not used again to evaluate the compression performance

1.4 Optimal motion estimation for wavelet-based video coding

This work was started during my PhD thesis and completed shortly after, in collaboration with M. Antonini, M. Barlaud, T. André and F. Castaldo. It has been published in [24]

A number of tools employed for wavelet-based video coding [40,108] have been originally conceived for hybrid block-based transform coding. This is the case of motion estimation, which generally aims to minimize the energy or the absolute sum of the prediction error. However, as wavelet video coders do not employ predictive coding, this is no longer an optimal approach, which should instead take into account the coding gain. This may be complex in sight of the recursive nature of the wavelet transform, but we have found a simple solution for a useful class of temporal filters. Experiments confirm that the optimally estimated vectors increase the coding gain as well as the performance of a complete video coder, but at the cost of an increase in complexity.

1.4.1 Coding Gain for Biorthogonal WT

For a given transform, coding gain is defined [57] as the ratio between D_{PCM}, the distortion resulting from quantization of the input signal, and D_{TC}, the distortion achievable by transform coding. For orthogonal subband coding of Gaussian data, in the high resolution hypothesis this turns out to be the ratio between arithmetic and geometric mean of subband variances. We want a suitable expression for this quantity in the case of generic spatiotemporal wavelet decomposition as well. We introduce the following notation: let N be the number of pixels (equal to the number of transform coefficients), L the number of decomposition levels and M the number of resulting subbands. For each $i \in \{1,\dots,M\}$, the i-th subband has N_i coefficients, with a variance σ_i^2. Finally, let us call $a_i = N_i/N$ the fraction of coefficients belonging to this subband, and w_i norm of (any of) the columns of the polyphase matrix corresponding the the i-th subband [128]. It has been shown that, under the hypothesis of high resolution and jointly Gaussian data, the coding gain for this transform can be expressed as [51,76]:

$$\mathrm{CG} = \frac{\sum_{i=1}^{M} a_i w_i \sigma_i^2}{\prod_{i=1}^{M} \left(w_i \sigma_i^2\right)^{a_i}}, \tag{1.10}$$

In the case of orthogonal transform with the same number of coefficients per subband (as for example the DCT), the coding gain is the ratio of the arithmetic and geometric means of subband variances. Equation (1.10) extends this result to the more general case of arbitrary wavelet decomposition with non-orthogonal filters. Thus, we are interested in deriving a criterion which allows to find the MVs that maximize the CG. It can be shown that the numerator of Eq. (1.10) does not depend on the MVs. Therefore they only affect the denominator, and in particular the variances of transformed subbands. In conclusion, our problem is equivalent to the minimization of $\rho^2 = \prod_i \left[(\sigma_i^2)^{a_i}\right]$. When considering the minimum transform coding distortion D_{TC}^*, we should refer to three-dimensional (*i.e.* spatiotemporal) SBs. Actually we will consider only temporal SBs since some earlier studies on this problem showed that considering spatiotemporal instead of temporal SBs gives little or no gain [40].

Unfortunately, in the general case, the minimization of ρ^2 is not an easy task because the MVs computed for a generic decomposition level affect all subsequent levels of WT transform. More precisely, if we consider the i-th level of temporal decomposition, the optimization of the set of motion vectors associated to this level (let it be $V^{(i)}$) must take into account the influence of this MVs set on *all* subsequent temporal subbands: so we should jointly optimize all level MVs in order to minimize ρ^2.

1.4.2 Developing the Criterion for a special case

The minimization of ρ^2 is a difficult problem to approach analytically and extremely demanding in terms of computational complexity. However, it can be remarkably simplified with a suitable choice of temporal filters such as the class of $(N,0)$ lifting schemes (LS) [14]. These filters are quite effective and have good scalability properties. Moreover, when using them, the low-pass branch of the WT filterbank is just the temporally subsampled input sequence, which does not depend on motion vectors. As a consequence, the i-th high frequency subband can be computed directly from the input sequence, independently from other SBs. This means also that all the subband variances are actually independent

from one another and that they can be minimized separately. In other words, each $V^{(i)}$ can be optimized separately providing that it minimizes the i-th high frequency SB variance. For this reason, we will refer from now on to the first level, and will drop the level dependency notation.

Further analytical developments are possible if we refer to a specific $(N,0)$ LS, such as the $(2,0)$. Let us recall the equation for this special case. Let $x_k(\mathbf{p})$ be a pixel from the k-th input image, and let l and h be the low- and high-pass subband respectively. We refer to the motion vector from i to j as $\mathbf{v}_{i\to j}(\mathbf{p})$, and we introduce the backward and forward MVs, $B_k = \mathbf{v}_{k\to k-1}$ and $F_k = \mathbf{v}_{k\to k+1}$. Now we can write the motion-compensated LS equations:

$$h_k(\mathbf{p}) = x_{2k+1}(\mathbf{p}) - \frac{1}{2}\left[x_{2k}(\mathbf{p}+B_{2k+1}(\mathbf{p})) + x_{2k+2}(\mathbf{p}+F_{2k+1}(\mathbf{p}))\right]$$
$$l_k(\mathbf{p}) = x_{2k}(\mathbf{p})$$

Therefore for this LS the vector set to optimize is: $V = \{B_{2k+1}, F_{2k+1}\}_{k\in\mathbb{N}}$. As h_k depends on both B_{2k+1} and F_{2k+1}, and only on them, these vectors have to be jointly minimized, but this can be done independently from other motion vector fields B_j, F_j with $j \neq 2k+1$. Without losing generality, we can refer from now on to the optimization of a vector couple, instead of the whole set V. The optimal couple B^*_{2k+1}, F^*_{2k+1} is the one minimizing the variance of h_k. Since it has zero mean, this is equivalent to minimizing the energy of h_k, indicated with $\mathcal{E}(h_k)$.

$$\left(B^*_{2k+1}, F^*_{2k+1}\right) = \arg\min_{B_{2k+1},F_{2k+1}} \mathcal{E}\left\{h_k(\mathbf{p})\right\}$$

After some calculations, it can be found that the optimal MVFs are given by:

$$\left(B^*_{2k+1}, F^*_{2k+1}\right) = \arg\min_{B_{2k+1},F_{2k+1}} \left[\mathcal{E}(\epsilon_B) + \mathcal{E}(\epsilon_F) + 2\langle\epsilon_B, \epsilon_F\rangle\right] \qquad (1.11)$$

where ϵ_F [ϵ_B] is the forward [backward] motion-compensated prediction error and $\langle\cdot,\cdot\rangle$ denotes the inner product (correlation) between two images.

Equation (1.11) defines the proposed ME criterion. We can compare it to the usual criteria. When for example the SSD is used, one independently minimizes $\mathcal{E}(\epsilon_B)$ and $\mathcal{E}(\epsilon_F)$, so one probably attains a low value of the criterion. This explains why MSE-based ME criteria often perform well with WT video coders. However, they do not necessarily achieve the minimum of criterion (1.11) as the mixed term is not taken into account. This term grows larger when the two error images are more similar, meaning that the optimal backward and forward vectors are not independent: they should produce error images as different as possible, being not enough to minimize error images energies. In other words, regions affected by a positive backward error should have a negative forward error and vice versa. The criterion introduced in (1.11) can easily be modified in order to take into account the motion vector coding cost. At this end it suffices to add a term of the form $\lambda R\left(B_{2k+1}, F_{2k+1}\right)$ to the cost function. However, for simplicity the following experimental results have been obtained neglecting this term. Better global RD-performance can be expected if the vectors rate is taken into account.

1.4.3 Experimental results

In a first experiment we use the ME criterion (1.11) in order to find MV in two test sequences, "foreman" and "akiyo", and we compare the corresponding coding gain. As we expected, we observe a consistent improvement of the coding gain (in average 0.8 dB with respect to SAD), independently from the sequence and the precision (full, half and quarter pixel). We also measured the Bjontegaard Delta Rate [16], and we found an average rate reduction of 8 %. More results can be found in [24] .

As far as the complexity is concerned, the proposed method is more complex than the classical ones, since it needs the joint estimation of backward and forward vectors. Faster, suboptimal search strategies such as those proposed in [107] can be envisaged in order to reduce its complexity. Further analysis could include the impact of taking into account the MV rate into the ME criterion.

1.5 Three-dimensional video coding

The contribution illustrated in this section are the result of the collaboration with team members (B. Pesquet-Popescu, G. Valenzise, I. Daribo, M. Kaaniche, W. Miled, E. Mora), and with external teams (J. Jung from Orange Labs – Issy-Les-Moulineaux, G. Cheung from

Figure 1.3: Left: block-based disparity field, used as initialization of our algorithm. Right: the resulting dense field.

IIT – Tokyo, A. Ortega from USC – Los Angeles). The results have been published in [33, 46, 94, 97, 130] and in two chapters of [50]. The latest results are the object of two submitted journal papers.

The compression of 3D video signals is currently one of my main research domains. The main challenge in this framework is to efficiently exploit all the redundancy present in stereo, multi-view, or multi-view-plus-depth signals. In particular, one should be able to remove inter-view correlation and inter-component correlation (*i.e.* the redundancy between a single view and its associated depth map). In this section we show some contributions to multi-view video coding (Section 1.5.1), to multiple-video-plus-depth compression (Section 1.5.2) and to a new approach based on the concept of *Don't Care Regions*.

1.5.1 Depth map estimation and coding

In this section we illustrate some of our work about multi-view video coding by using inter-view redundancy. We were motivated by the need to efficiently exploit the disparity information between views. We consider the case of multi-view video (without explicit depth information). Using inter-view redundancy is at the basis of the multi-view extension of H.264/MPEG-4 AVC, referred to as H.264/MVC [38, 131]. In this standard, pictures from other views are inserted in the reference picture list, so that they can provide possibly a more efficient prediction than the temporal one. As a consequence, the disparity between two views is implicitly represented via a block-based disparity field. Our contribution here consists in improving the disparity prediction unit in H.264/MVC by using the dense disparity estimation (DDE) method described in [90] followed by a block-based RD segmentation. Based on a set theoretic framework, this DDE approach incorporates various convex constraints corresponding to *a priori* information and yields disparity vectors with ideally infinite precision. We consider a regularization constraint based on total variation (TV), in order to produce a smooth disparity field that preserves discontinuities. As a result, the disparity estimation problem can be written as follows: Find the disparity field d belonging to the intersection of all the constraints ($d \in S = \bigcap_{m=1}^{M} S_m$) and minimizing the cost function $J(d)$ written as

$$J(d) = J_D(d) + \alpha J_S(d)$$

where J_D is a *data term* that measure the distance between corresponding pixels (*i.e.* the disparity-compensated error energy), and J_S is a *regularization term*, also called *smoothing prior*, that enforces the regularity of the solution. In our work, the data term is approximated to the first order Taylor expansion around an initialization disparity field, and the regularization term penalizes solutions too different from the initialization field. As far as the constraints are concerned, we consider a TV constraint with a maximum value τ for the TV of the disparity field. Other constraints limit the dynamics of the disparity field. In Fig. 1.3 we show the initialization disparity field obtained by overlapped block matching and the resulting dense field obtained with our method. The improvement of the estimation is evident.

Once obtained the dense disparity field (*i.e.* one vector per pixel), we reduce its coding cost using a rate-distortion driven segmentation: the dense field in a macroblock, a block or a sub-block is replaced

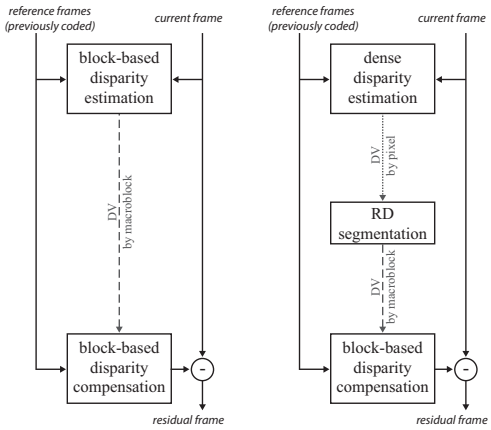


Figure 1.4: Disparity prediction: (left) block-based estimation, (right) enhanced by a dense estimation.

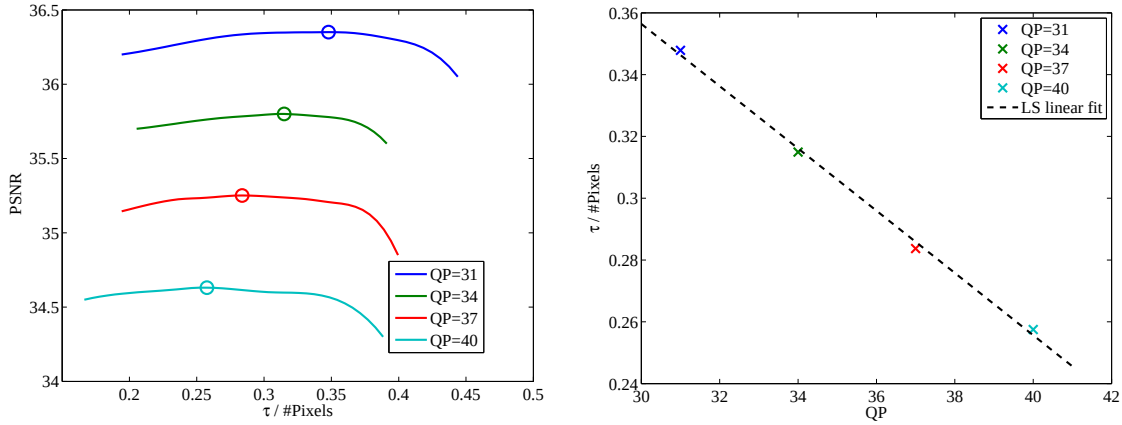


Figure 1.5: Left: Effect of τ (normalized over the number of image pixels) on the disparity quality, for several values of the quantization step. Right: Best normalized τ values in function of QP. The best fitting first order polynomial is shown as well.

with a constant field (represented by just one vector) according to the RD performance of this choice. The resulting coding scheme is shown in Fig. 1.4.

The proposed method remarkably improves the effectiveness of candidate predictors for encoding the current macroblock. As a result, we observe some remarkable coding gain when applying it to H.264/MVC. We compute the bit-rate reductions using the Bjontegaard metric, and observe savings up to 45 % on the test sequence "outdoor". Gains on other sequence range from 3 % to 10 %.

This method was later extended [33] in order to automatically find out the best value of the TV constraint τ. We have found that the optimal value of τ (normalized with respect to the number of pixels per image) decreases linearly with the quantization parameter QP, as shown in Fig. 1.5. We compute the sample correlation coefficient obtaining:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{QP_i - \overline{QP}}{\sigma_{QP}}\right)\left(\frac{\tau_i^* - \overline{\tau^*}}{\sigma_{\tau^*}}\right) = -0.9987$$

where the bar represents the (sample) mean and σ represents the (sample) standard deviation. Finally we computed the least square linear fitting of QP and τ^*. We found the following regression equation:

$$\tau^* = -0.0101QP + 0.6587 \qquad (1.12)$$

Using this equation allows to automatically find effective values for the total variation constraint.

1.5.2 Depth coding for 3D-VC

This and next subsection deal with the multi-views-plus-depth (MVD) video format. When using this format, only a limited number of views (typically two or three) is sent to the decoder, along with the associated depth map that accounts for the distance of each pixel from the camera. This geometrical information allows the decoder synthesizing an arbitrary number of intermediary views (synthetic or virtual views) between those actually transmitted. This paradigm appears to effectively solve a major drawback of the multi-view framework, where the coding rate is proportional to the number of decodable views.

As a consequence, there is a huge research effort about MVD [18,122], culminating with the standardization process of this format [72]. In this section we illustrate our contributions related to the standard, while in the next we describe a different approach.

A first contribution, described in [94], is about INTRA mode inheritance in depth map coding. The framework is the 3D Video Coding (3D-VC) standardization process [72], which is based on the H.264/MPEG-4 AVC or the HEVC standards. In this context we consider the depth map coding problem, and we try to exploit the inter-component dependency to reduce the rate. We consider the case where the depth is encoded after the texture, and therefore the decoded texture is available both at the encoder and the decoder. A preliminary analysis shows that the largest part of the coding rate allocated to depth is usually consumed for INTRA images, and that in turns, the signaling of the coding mode demands as much as the 30 % of the total depth rate. The reason is that on one hand, HEVC allows to use up to 35 different modes to generate a predictor in INTRA slices, so the mode belongs to a large alphabet; on the other, with such a large variety of available predictors, it is probable that the resulting residual can be encoded with a small rate. As a consequence, reducing the coding rate of the INTRA modes may have a relevant impact on depth coding.

Moreover, we have observed that, if we separately encode depth and texture with rate-distortion optimization, often a depth block is encoded with the same mode of the texture. We also observed that this is particularly common when the image presents strong contours, *i.e.* a dominant direction in the texture geometry. Therefore our idea is the following. For a given block of depth to be encoded:

1. Use a contour detector algorithm on the co-located decoded texture block;
2. On the resulting data, compute a criterion accounting for the directionality of the contours;
3. If the criterion is larger than a threshold, the texture mode is "inherited" for depth coding.

We implement these steps as follows. In step 1, we use a Sobel filter to estimate the image gradient on the colocated texture block. In step 2, gradient statistics are used to decide whether the block has a "strong" or dominant direction: the reason to do this is that we observed that coding mode is effectively inherited near object contours. For example, the criterion could consist in evaluating the maximum absolute value of the gradient, or in detecting the presence of a dominant direction, by considering the statistics of the gradient image. Finally, in step 3 we try to take advantage from the inherited mode. At this end, we insert the inherited mode in the most probable mode (MPM) list of HEVC.

The MPM list is a coding tool used to reduce the mode coding cost. It consist in a short list of modes, created by using information available both to the encoder and the decoder. Signaling a mode in this short list is much less expensive than signaling one out of 35 possible INTRA modes. Our algorithm consists in modifying the content of the short list, using the information of the texture: hopefully, the inherited mode is effective and actually selected in the following RD-optimized mode selection. In this case the proposed method would give a coding gain, since the mode is encoded from the MPM and not from the long list of 35 modes.

This algorithm has been tested in the reference software of the HEVC standard, namely in HM-3.3. We considered the MPEG test sequences used for the 3DV normalization process, and we applied the common test conditions [72]. We observed average rate reductions (using the Bjontegaard metric [16]) around 1 %. We are recently working on an improved version of the algorithm, where the decision about inheritance is made after a more careful analysis of the texture block, and the first results show higher gains. This new algorithm is described in a submitted journal paper [95].

Other contributions to the field of the 3DV normalization are about MV representation [97] and the quad-tree representation of depth in 3D-HEVC video coding. The former tool was accepted in the standardization process [98]. For the latter, we propose an algorithm for predicting and limiting the partition of depth coding units, thus allowing for a remarkable complexity reduction with a very small rate-distortion loss. This algorithm is the subject of a submitted paper [96].

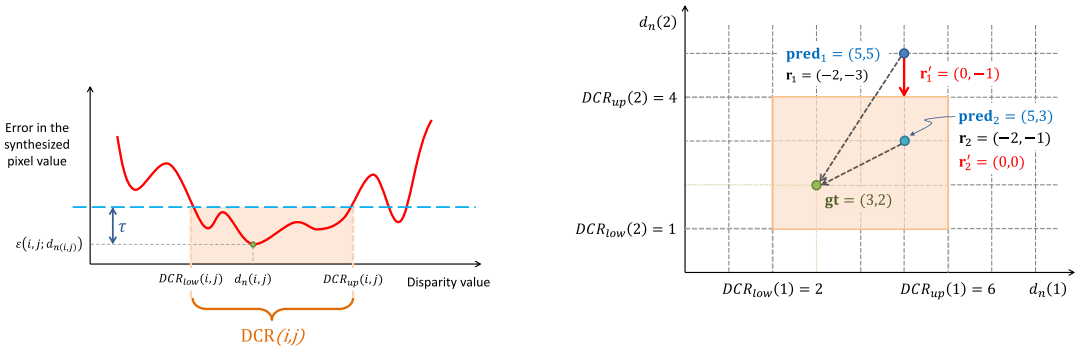


Figure 1.6: Left: Definition of DCR for a given threshold τ. Right: Coding the residuals using DCR with a toy example with just two pixels ($d_n(1)$ and $d_n(2)$). In conventional coding, given predictor (**pred**), one aims to reconstruct the original ground truth (**gt**). However, considering DCR, it is sufficient to encode a generally smaller residual, i.e. one that enables to reconstruct a value inside or on the border of the DCR (shaded area in the picture).

1.5.3 MVD video compression via *Don't Care Regions*

As previously observed, depth maps are not directly viewed, but are only used to provide geometric information of the captured scene for view synthesis at decoder. Thus, as long as the resulting geometric error does not lead to unacceptable quality for the synthesized view, each depth pixel only needs to be reconstructed at the decoder coarsely within a tolerable range. We first formalize the notion of tolerable range per depth pixel as *Don't Care Region* (DCR), by studying the synthesized view distortion sensitivity to the pixel value—a sensitive depth pixel will have a narrow DCR, and vice versa. Given per-pixel DCRs, we then modify inter-prediction modes during motion prediction to search for a predictor block matching per-pixel DCRs in a target block (rather than the fixed ground truth depth signal in a target block), in order to lower the energy of the prediction residual for the block. This introduces a potential rate reduction for the same reconstructed quality.

Definition of Don't Care Regions (DCR). We now define per-pixel DCRs for depth map $\mathbf{D}_n$, assuming target synthesized view is n. In the following, we will rather refer to the *disparity* field $\mathbf{d}_n$, which can be obtained from the depth once the camera parameters are known. A pixel $v_n(i,j)$ in texture map $\mathbf{v}_n$, with associated disparity value $d_n(i,j)$, can be mapped to a corresponding pixel in view $n+1$ through a view synthesis function $s(i,j;d_n(i,j))$. In the simplest case where the views are captured by purely horizontally shifted cameras, $s(i,j;d_n(i,j))$ corresponds to a pixel in texture map $\mathbf{v}_{n+1}$ of view $n+1$ displaced in the x-direction by an amount proportional to $d_n(i,j)$. The view synthesis error, $\varepsilon(i,j;d)$, can thus be defined as the absolute error between reconstructed and original pixel value, given disparity d for pixel (i,j); i.e., $\varepsilon(i,j;d) = |s(i,j;d) - v_n(i,j)|$. If $\mathbf{d}_n$ is compressed, the reconstructed disparity value $\tilde{d}_n(i,j)$ employed for view synthesis may differ from $d_n(i,j)$ by an amount $e(i,j) = \tilde{d}_n(i,j) - d_n(i,j)$, resulting in a (generally larger) view synthesis error $\varepsilon(i,j;d_n(i,j)+e(i,j)) > \varepsilon(i,j;d_n(i,j))$. We define the *Don't Care Region* $\mathrm{DCR}(i,j) = [\mathrm{DCR}_{low}(i,j), \mathrm{DCR}_{up}(i,j)]$ as the *largest* contiguous interval of disparity values containing the ground-truth disparity $d_n(i,j)$, such that the view synthesis error for any point of the interval is smaller than $\varepsilon(i,j;d_n(i,j)) + \tau$, for a given threshold $\tau > 0$. The definition of DCR is illustrated in Figure 1.6 (left). Note that DCR intervals are defined *per pixel*, thus giving precise information about how much error can be tolerated in the disparity maps.

The DCRs give us a new degree of freedom in the encoding of disparity maps, where we are only required to reconstruct each depth pixel at the decoder to within its defined range of precision (as opposed to the original depth pixel), thus potentially resulting in further compression gain. Specifically, we change three aspects of the encoder in order to exploit DCRs: i) motion estimation, ii) residual coding, and iii) skip mode.

Motion estimation. During motion estimation for depth map encoding, the encoder searches, for each target block $\mathcal{B}$, a corresponding predictor block $\mathcal{P}$ in a reference frame which minimizes the La-

grangian cost function

$$\mathcal{P}^* = \arg\min_{\mathcal{P}} D_{\mathrm{MV}}(\mathcal{B},\mathcal{P}) + \lambda_{\mathrm{MV}} R_{\mathrm{MV}}(\mathcal{B},\mathcal{P}), \tag{1.13}$$

where $R_{\mathrm{MV}}(\mathcal{B},\mathcal{P})$ is the bit overhead required to code the motion vector from position of $\mathcal{P}$ to $\mathcal{B}$, and λ_{MV} is a Lagrange multiplier. For a given predictor block $\mathcal{P}$, we can reduce the energy of the prediction residuals using defined per-pixel DCRs as follows. We first define a per-block *DCR space* for a target block $\mathcal{B}$ as the feasible space containing depth signals with each pixel falling inside its per-pixel DCR. As an example, Figure 1.6(right) illustrates the DCR space for a two-pixel block with per-pixel DCR $[2,6]$ and $[1,4]$. For a given predictor block, to minimize the energy of the prediction residuals, we identify a signal in DCR space closest to the predictor signal in Euclidean distance. In Figure 1.6(right), if the predictor is $(5,5)$, we identify $(5,4)$ in DCR space as the closest signal in DCR space, with resulting residuals $(0,-1)$. If the predictor is $(5,3)$, we identify $(5,3)$ in DCR space as the closest signal with residuals $(0,0)$.

In mathematical terms, we compute a prediction residual $r'(i,j)$ for each pixel (i,j) given predictor pixel value $\mathcal{P}(i,j)$ and DCR $[\mathrm{DCR}_{low}(i,j), \mathrm{DCR}_{up}(i,j)]$ according to a soft-thresholding function:

$$r'(i,j) = \begin{cases} \mathcal{P}(i,j) - \mathrm{DCR}_{up}(i,j) & \text{if } \mathcal{P}(i,j) > \mathrm{DCR}_{up}(i,j), \\ \mathcal{P}(i,j) - \mathrm{DCR}_{low}(i,j) & \text{if } \mathcal{P}(i,j) < \mathrm{DCR}_{low}(i,j), \\ 0 & \text{otherwise.} \end{cases} \tag{1.14}$$

We then use the residuals $r'(i,j)$ with respect to DCR to calculate D_{MV} in (1.13). If SAD is used as distortion metric, we simply get: $D'_{\mathrm{MV}} = \sum_{(i,j)\in\mathcal{B}} |r'(i,j))|$.

Since the distortion D_{MV} is now zero for any motion vector which points to a predictor inside DCR, the encoder can select from a potentially larger set of zero-distortion candidate predictors. Among them, the one with the smallest rate term R_{MV} will be selected.

Coding of prediction residuals. Once the optimal predictor $\mathcal{P}^*$ for a given target block has been found, we encode r' with respect to the per-block DCR, in place of the residuals r computed with respect to ground truth depth signal. Notice that this applies also to INTRA coding modes as well. In general, since both rate and distortion terms are computed using minimum-energy residuals r' for inter and intra modes, the actual selected mode for a given target block will be different from the one selected when coding residuals with respect to the ground truth signal.

Skip mode. In H.264/MPEG-4 AVC, when the SKIP mode is used, the prediction residuals are not encoded. Thus, the reconstructed pixels could be potentially far away from DCR. This could be harmful since, by construction, there is no upper bound to the distortion in the synthesized view when a depth pixel is reconstructed outside DCR. This requires SKIP mode to be handled differently from INTER and INTRA: in particular, we prevent the SKIP mode to be selected from the encoder if *any* reconstructed pixel of that macroblock violates DCR. Although this could be conservative in terms of rate optimization, it guarantees that the distortion in the synthesized view for SKIP macroblocks will be bounded by τ.

Experimental results. We have modified an H.264/MPEG-4 AVC encoder (JM reference software v. 18.0) in order to include DCR in the motion prediction and coding of residuals. Our test material includes 100 frames of two multiview video sequences, *Kendo* and *Balloons* with spatial resolution of 1024×768 pixels and frame rate equal to 30 Hz. For both sequences we coded the disparity maps of views 3 and 5 (with IPP...GOP structure), using either the original H.264/AVC encoder or the modified one. In the latter case, we computed per-pixel DCRs with three values of τ, namely $\tau \in \{3,5,7\}$. Given the reconstructed disparities in both cases (with/without DCR), we synthesize view $\mathbf{v}_4$ using the uncompressed views $\mathbf{v}_3$ and $\mathbf{v}_5$ and the compressed depths $\tilde{\mathbf{d}}_3$ and $\tilde{\mathbf{d}}_5$. Finally, we evaluate the quality of the reconstructed view $\hat{\mathbf{v}}_4$ w.r.t. ground-truth center view $\mathbf{v}_4$.

For the *Kendo* sequence, using $\tau = 5$ we obtain an average gain in PSNR of 0.34 dB and an average rate saving of about 28.5 %, measured through the Bjontegaard metric. From the encoder statistics, we notice that most of the rate savings are obtained through a more efficient use of SKIP mode (which increases by over 18% in this case), and by a more efficient prediction of motion and coding of residuals.

2
Adaptive image compression

In this chapter we report the research work on image compression using adaptive methods. The main problem here is how to deal with discontinuities in images (*i.e.* object contours). We consider two approaches: in the first, the signal is segmented and thus the discontinuity is removed from it, allowing for a better energy concentration. However, the difficulties related to the segmentation, to the contour information coding and to the management of non-rectangular (*i.e.*, shape-adaptive) transforms, limit the effectiveness of this method to some particular class of problems. This approach is illustrated in Section 2.1. The second approach consists in considering an adaptive system in which the linear operators used for the transform computation are modified according to the signal characteristics. In this case we are obliged to give up the transform isometry; as a consequence, the resource allocation becomes more challenging. Therefore, the second part of the chapter (Section 2.2) is devoted to the evaluation of quantization noise in the transform domain for this transform. The results can be applied to coding (perceptual and non-perceptual) and denoising.

2.1 Object-based image coding

This work is the continuation of some topics started during my PhD thesis. It was carried out at "Federico II" University of Naples in collaboration with prof. Poggi, Dr. Verdoliva, and Dr. Parrilli, and resulted in two journal publications [29, 34].

Object-based image coding has gathered attention from the research community in the late 90's and in the 2000's, since it has some very interesting characteristics. The major conceptual reason for object-based coding is that images are *naturally* composed by objects, and the usual pixel-level description is only due to the lack of a suitable language to efficiently represent them. Once objects have been identified and described, they can be treated individually for the most diverse needs. This was the driving idea of the MPEG-4 standard [73].

Here we report our main results about object based image coding. In a first part (Section 2.1.1), we consider a general approach to the problem, with the target of pointing out strength and weaknesses of the object-based coding paradigm. The second part (Section 2.1.2) deals with a more specific problem, the one of multi-spectral image compression. Also thanks to the results of the first part, we have developed an efficient coding framework for this kind of images.

2.1.1 Costs and advantages of object-based image compression

In terms of coding efficiency, the object-based description of an image presents some peculiar costs which do not appear in conventional coding. First of all, objects' shape and position must be described by means of some segmentation map, sent in advance as side information. In addition, most coding techniques become less efficient when dealing with regions of arbitrary size and shape. Finally, each object needs its own set of coding parameters, which adds to the side information cost. On the positive side, an accurate segmentation carries with it information on the graphical part of the image, the edges, and hence contributes to the coding efficiency and perceived quality. Moreover, component regions turn out to be more homogeneous, and their individual encoding can lead to actual rate-distortion gains. In any case, to limit the additional costs, or even obtain some performance improvement, it is necessary to select appropriate coding tools, and to know in advance their behavior under different circumstances.

Here we focus on a wavelet-based shape-adaptive coding algorithm. We implemented an object-based coding scheme with the following elementary steps (see Fig. 2.1)

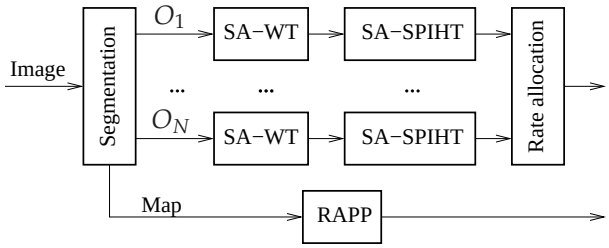


Figure 2.1: Object-based coding scheme

1. image segmentation;
2. lossless coding of the segmentation map (object shapes);
3. shape-adaptive wavelet transform of each object;
4. shape-adaptive SPIHT coding of each object;
5. optimal post-coding rate allocation among objects.

As for image segmentation, we consider both ideal and practical segmentation. The first case is useful to assess the effect of more or less complex contours; the second to validate the result with an actual segmentation method, based on Bayesian techniques. The segmentation maps are losslessly encoded by means of a modified version of the RAPP algorithm [118], which proves very efficient for this task. As for the transform, we use the shape-adaptive wavelet transform (SA-WT) proposed by Li and Li [78]; the coding technique is a shape-adaptive version of SPIHT (SA-SPIHT) [35] (similar to that formerly proposed in [77] and further refined in [91]) which extends to objects of arbitrary shape the well-known image coder proposed by Said and Pearlman [120]. The attention on wavelet-based coding is justified by the enormous success of this approach in conventional image coding [129] leading to the wavelet-based standard JPEG-2000 [123]. After coding, the rate-distortion (RD) curves of all objects are analyzed so as to optimally allocate bits among them for any desired encoding rate, like in the post-compression rate allocation algorithm of JPEG-2000. This coding scheme is compared with its "flat" version, *i.e.* a version using ordinary WT and SPIHT. By the way, this coder does not need segmentation, map coding nor rate allocation.

The main focus of this work is to analyze the general mechanisms that influence the efficiency of wavelet-based shape-adaptive coding and to assess the difference in performance with respect to conventional wavelet-based coding.

In more detail, we can identify three causes for the additional costs of object-based coding: 1) the reduced energy compaction of the WT and 2) the reduced coding efficiency of SPIHT that arise in the presence of regions with arbitrary shape and size, and 3) the cost of side information (segmentation map, object coding parameters). As for the possible gains, they mirror the losses, since they arise for the increased energy compaction of the WT, when dominant edges are removed, and for the increased coding efficiency of SPIHT when homogeneous regions have to be coded.

The two coding schemes (flat and SA) were compared on several images, both synthetic and natural. We refer the reader to our paper [29] for the complete results. For the sake of brevity, here we report only the main conclusions. Our aim is to assess the rate-distortion performance of our object-based coder by means of numerical experiments in typical situations of interest, and single out, to the extent possible, the individual phenomena that contribute to the overall losses and gains. Since the usual coding gain does not make sense for SA-WT, we measure its compaction ability by analyzing the RD performance of a virtual oracle coder which spends bits only for quantization. SA-WT losses turn out to be quite significant, especially at low rates. Although the quantization cost is by itself only a small fraction of the total cost, the reduced compaction ability of SA-WT has a deep effect also on the subsequent coding phase, the sorting pass of SPIHT. In fact, our experiments reveal this to be the main cause of SPIHT losses, while the presence of incomplete trees plays only a minor role. This is also confirmed by the fact that SA-SPIHT performs about as well as more sophisticated coding algorithms, and suggests that algorithms that code significance maps equally well perform equivalently at shape-adaptive coding regardless of how carefully their coding strategies have been tailored to accommodate object boundaries, and hence improving boundary handling is largely a wasted effort. As for the gains, our analysis showed that they can be significant when the image presents sharp edges between relatively homogeneous regions but also that this is rarely the case with real-world images where the presence of smooth contours, and

the inaccuracies of segmentation (for a few objects) or its large cost (for many objects) represent serious hurdles towards potential performance gains.

The experimental evidence (the bulk of which was not presented here) allows us to provide some simple guidelines for the use of object-based coding, by dividing operative conditions in 3 major cases.

1. **A few large (say, over ten thousand pixels) objects with smooth contours.** RD losses and gains are both negligible, hence performance is very close to that of flat wavelet-based coding, the best currently known. In this case, which is the most explored in the literature, resorting to wavelet-based coding, with SA-WT and SA-SPIHT or BISK [55], is probably the best solution.
2. **Many small objects (or a few large objects with very complex boundaries) at low rates.** There are significant RD losses, both because of the reduced compaction ability of SA-WT and because the coding cost of the segmentation map is not irrelevant. This is the only case, in our opinion, where the wavelet-based approach leaves space for further improvements, as the introduction of new tools explicitly thought to encode objects rather than signals with arbitrary support.
3. **Many small objects (or a few large objects with very active boundaries) at high rates.** Here, the losses due to the SA-WT and the side information become almost negligible, and the performance comes again very close to that of flat coding, making wavelet-based coding very competitive again.

This list accounts mainly for the losses, as performance gains are currently achievable only for some specific source, like multispectral (MS) images. In this case, SA approach have two potential sources of improvement: first, the segmentation map cost is shared by several images (the components of an MS image); second, adaptive spectral transforms can attain better compaction performances when carried off on homogeneous data, *i.e.* on objects. This approach is illustrated in the following section.

2.1.2 Object-based multispectral image compression

In this section we show our work about an efficient, region-based algorithm for the compression of multispectral (MS) remote-sensing images. An example of false color MS image is given in Fig. 2.2(a)-(b). As shown in the previous Section, this approach, with the multiple pieces of information required, may be inherently inefficient. The goal of this research is to show that, in the case of multispectral images and by carefully selecting the appropriate segmentation and coding tools, region-based compression can be also effective in a rate-distortion sense. To this end, we resort to the coding scheme shown in Fig. 2.1, using Bayesian image segmentation, class-adaptive Karhunen-Loève spectral transform and shape-adaptive wavelet spatial transform, which outperforms state-of-the-art and carefully tuned conventional techniques, such as JPEG-2000 multicomponent or SPIHT-based coders.

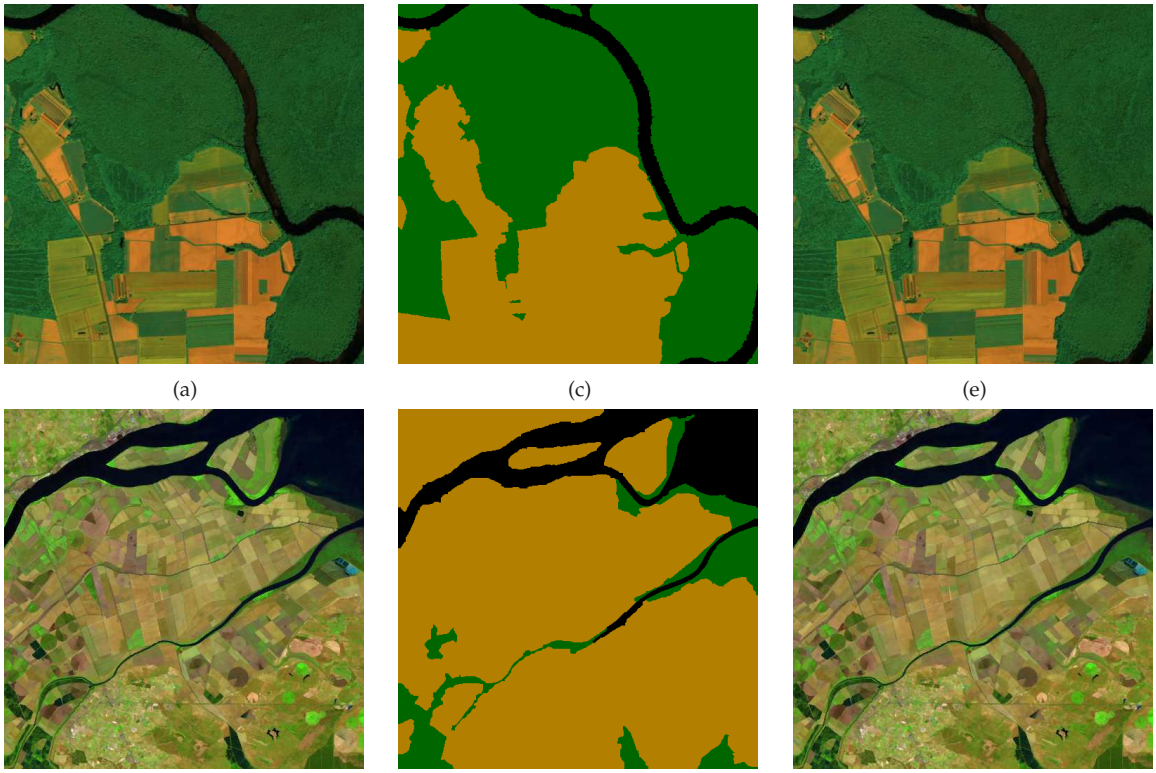
The details of the coding tools are given in the following.

Segmentation. In our application we have two requirements: on one hand, we want each region to be formed by pixels of the same type, so as to exhibit homogeneous statistics and increase the efficiency of subsequent encoding. On the other hand, we would like to segment the image in a small number of large regions, in order to have a simple map to encode, and to use shape-adaptive transforms on nice regular shapes. We resort to the algorithm developed in [48], based on a tree-structured Markov random field (TSMRF) model, which provides the segmentation maps with few, compact regions with regular boundaries. Possible residual isolated points and small fragments are eliminated by merging them with the dominant neighboring region. An example of resulting segmentation map (along with a false color representation of the associated MS image) is given in Fig. 2.2(c)-(d).

Map coding. Just as in the previous section, we resort to the efficient RAPP algorithm [118].

Transforms. In accordance with the different nature of spectral and spatial dependencies, we use a 1-d spectral transform first, and then a further 2-d (shape-adaptive) transform in the spatial domain. As for the spatial transform, we turn to the shape-adaptive wavelet transform (SA-WT) algorithm proposed by Li and Li [78]. For the spectral transform, several alternatives are possible. A promising candidate is the Karhunen-Loève transform (KLT) which, being data dependent, allows one to adapt the transform to the data to be encoded, but requires significant computation and calls for the transmission of some side information. However, KLT-based solutions fit much better the region-based approach and provide consistently superior RD performance.

There are at least three possible ways to use the KLT in this context: 1) use a single transform for the whole image; 2) use a different transform for each region; 3) use a different transform for each class. In the first case, KLT is used almost like a fixed transform, but for the fact that it is adapted to



(a) (c) (e)

(b) (d) (f)

Figure 2.2: Compression of multispectral images. Left: original MS images in false colors; middle: segmentation maps obtained by TSMRF; right: decoded images (0.3 and 0.6 bits per sample)

the image statistics. On the contrary, using a different KLT for each region is fully in the spirit of the region-based approach, since the transform is now adapted to the *local* statistics of the region, which might be markedly different from those of other regions. As a consequence one obtains a better energy compaction, but also a heavier computational burden and increased side information. The third option consists in using a single KLT matrix for all regions that belong to the same class (*e.g.*, trees, water *etc.*), with an obvious advantage in terms of computation and side information, but also with a good adaptation to the local statistics. For the sake of completeness, we also considered WT as a possible spectral transform.

Region coding. As for the previous Section, we will consider here the shape-adaptive version of SPIHT proposed in [35], with the modifications needed to manage 3D trees of wavelet coefficients.

Rate allocation. We implement a post-compression resource allocation algorithm: we first compute the rate-distortion (RD) curve of each object by encoding it at high rate, and then deallocate resources progressively, acting each time on the object with the lowest RD slope so as to keep an approximate equi-slope condition while reaching the desired overall coding rate. An "object" may be a different set of transform coefficients, since in the spatial domain we can work with the whole image, or the elementary regions, and in the spectral domain we can work with the whole set of bands or with each single band individually.

Experimental results. We consider two test MS images: a Landsat TM image (6 bands, 512×512 pixels, 8 bits per sample) and an AVIRIS image (32 bands, 512×512 pixels, 16 bits per sample). A false-color version of them is shown in Fig. 2.2(a)-(b). For complete results we refer the reader to our paper [34]. Here we report rate-distortion performances of several algorithms (in Fig. 2.3). We observe that the best region-based algorithm, using class-based KLT, outperforms even the best flat algorithm (JPEG-2000 multicomponent), excluding very low rates because of the rate penalty coming of side information. The advantage of the proposed technique is even more evident when it comes to post-processing or visual quality – see Fig. 2.2(e)-(f). As shown in our paper, post-processing tasks such as classification have better performances on shape-adaptive compressed data than on flat-compressed data.

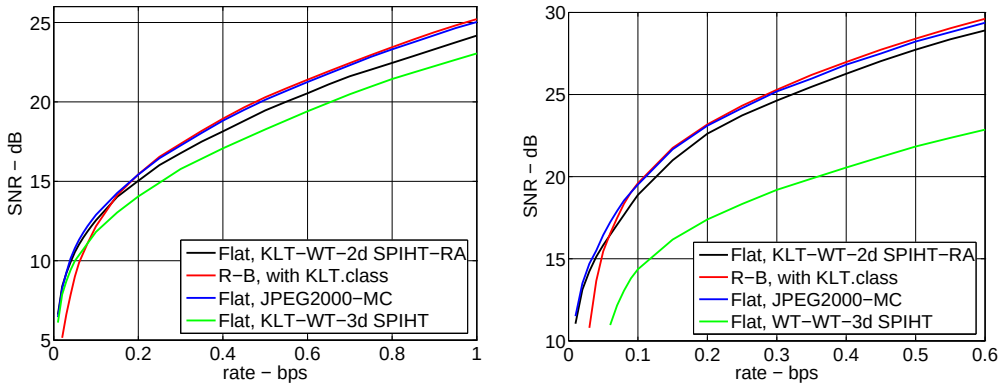
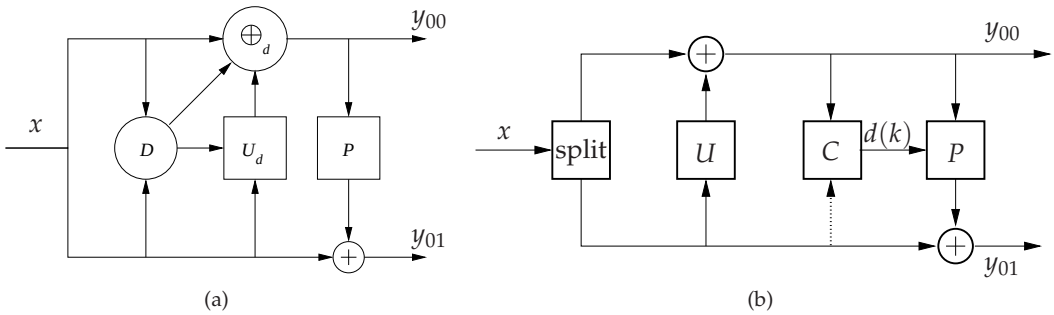


Figure 2.3: Rate-distortion performance of shape adaptive coding algorithms



(a) (b)

Figure 2.4: Adaptive lifting schemes. (a): adaptive update. (b): adaptive prediction

2.2 Adaptive wavelet basis for image compression

This work was carried out at TELECOM-ParisTech in collaboration with prof. Pesquet-Popescu and dr. Parrilli. It has been published in [2, 32, 104–106].

As previously observed, standard wavelet transforms do not completely fit to image representation, because they are very effective in representing smooth signals with pointwise discontinuities, but fail in representing discontinuities along regular curves, as image contours [49]. In the previous section, we illustrated in which condition shape-adaptive transforms may be used to effectively deal with this problem. Here we consider an alternative approach where the support of the signal is not changed (no segmentation is needed). On the contrary, different transform filters are used according to the local characteristics of the signal. This results into an adaptive lifting schemes (ALS) architecture.

One of the problem to be solved when using ALS for image compression is rate allocation among transform coefficients. This is not trivial, since ALS results into a non-isometric, non-separable, and even non-linear transform. In this section we show how this problem can be solved, *i.e.* how to estimate the quantization noise in the transform domain when ALS are used. Our approach consists in showing that ALS are equivalent to time-varying, linear filters, for which a normalization is possible, such that the transform-domain energy is equal to the signal energy at least for uncorrelated signals. The proposed method finds applications in image compression (allowing to reduce both MSE and perceptual distortion) and denoising.

Lifting schemes [47] can be used to build content-adaptive wavelet decompositions [42, 56, 88]. In particular, Heijmans, Piella and Pesquet-Popescu [70, 117] have proposed an adaptive update lifting scheme (AULS) using seminorms of local features of images in order to build a decision map $d(k)$ that determines the lifting update step: for example, when the decision map highlights important features like contours or singularities, a weaker filter (or no filtering at all) can be used. On the other hand, the prediction step is fixed (see Fig. 2.4(a)). One of the most interesting features of this adaptive transform is that it does not require the transmission of side information, since the decision on the update step can be made with the information available at the synthesis stage.

We also consider the case of adaptive prediction lifting schemes (APLS) [42], whose scheme is shown

(a) (b) (c) (d)

Figure 2.5: Lena details decoded at 0.25 bpp. APLS (a,c) shows less artifacts than non-adaptive 9/7 wavelet (b,d): in the first row the mirror border is more regular; in the second less spurious structures appear in the hairs and near the nose.

in Fig. 2.4(b). We decided to analyze this class of filters because of its popularity and good performance. In order to guarantee perfect reconstruction (in absence of quantization) at the synthesis stage it is important that the decoder can reproduce all the encoder decisions. To obtain this goal without sending the side information $d(k)$, the update stage is applied first and the decision is based on the approximation signal only, that is, the dotted line in Fig. 2.4(b) does not operate.

We note that, according to the value of the decision map at position k, we use one out of a certain number D of linear update or prediction filters. Since the decision map depends at its turn on the input signal, the whole system is inherently non-linear. However, if we forget about the dependence of $d(k)$ on x and just look at its dependency on index k, we can see it as a linear, time-varying system. For such a system we can find a polyphase representation, and then, as shown in [128], the distortion D in the original domain is related to the distortion D_{ij} in the wavelet subband y_{ij} by the relation: $D = \sum_{ij} w_{ij} D_{ij}$. Note that the subband y_{ij} is the j-channel of the i-th decomposition level. The terms w_{ij} are computed based on the reconstruction polyphase matrix of subband i, j. In our work [104] we show how to compute the polyphase matrix for the general case of N-levels, 2D AULS transform. As for the APLS case, the problem is quite similar, and just a bit more complex; the solution is provided in [106]. For both cases, we found a surprisingly simple result, at least for a single decomposition level: the term w_{ij} to be used for each subband is a weighted average of the norms of the D adaptive filters, and the weight of the n-th filter is the relative frequency of that filter in the decision map. For more decomposition levels we provide an algorithm for weight computation, but there is no simple interpretation.

Experimental results. A first set of experiments is performed in order to evaluate the ability of our algorithm to estimate the energy of an uncorrelated signal (as the quantization noise is commonly modeled to be) in the transform domain. Using the weights, the relative error in energy estimation is in the range 0.17 % $\sim$ 1.15 % for the AULS and 0.14 % $\sim$ 0.83 % for the APLS, while if one is not using the weights (which is the only way to go without our technique), the error ranges rather from 40 % to 94 %.

As a consequence, using the proposed weights in a rate allocation strategy allows to improve the RD performance of an ALS-based coding scheme of up to 3.2 dB for the AULS and of up to 1.0 dB for the APLS. In particular, the latter is improved to the point of having practically the same PSNR performance of the most-celebrated Daubechies 9/7 filters. Moreover, since the APLS is better suited to object contour representation, images encoded with it have higher values of perceptual metrics such as the weighted PSNR [92]. This is also visible in a decoded image, see Fig. 2.5

Other applications. Our algorithm for weights computation can improve performances whenever one want to use ALS (with respect to the case where constant weights are applied). For example, we applied these weights to compute saliency masks in the transform domain [32] achieving an improvement of the objective perceptual quality measured by the SSIM [133]. Similarly, we implemented a simple denoising algorithm via soft-thresholding on APLS coefficients [2]: using weights allow to obtain PSNR gains up to 1 dB and SSIM gain up to 3 % with respect to the case where no weights were employed.

3
Distributed video coding

Distributed Video Coding (DVC) [67] is a new promising paradigm in video communication. It refers to the compression of multiple outputs of correlated sources which do not communicate with each other. The targeted applications are numerous, such as video compression on mobile devices, multi-sensor surveillance systems, and so on. Even though well known theoretical results [124, 137] indicate that distributed source coding has the same performance bound than joint coding, practical DVC scheme are still quite far from the performance of the most recent video standard.

However, recently there has been a huge effort to improve DVC. Most of the attention has been devoted to the problem of side information (SI) generation. In the most popular DVC architecture, introduced by Aaron *et al.* [1], the so-called key frames (KF) are compressed independently from all the other frames, using an "INTRA" coding technique (*e.g.*, H.264 in INTRA mode). They are used at the decoder to generate an estimation of the other frames, called Wyner-Ziv frames (WZF). This estimation, referred to as side information (SI), is corrected by the parity bits from a suitable channel coder, produced by the encoder upon the decoder's request.

We have been working on the problem of SI generation, since DVC performances strongly depend on the SI quality. The proposed solutions are illustrated by comparing them to one of the most popular DVC scheme, the one of the DISCOVER project [15], see Section 3.1. We considered approaches based on dense motion field generated with differential ME algorithms (Section 3.2), on high order trajectory interpolation (Section 3.3), on fusion of local and global motion information (Section 3.4). Applications of DVC to robust video coding and to interactive streaming are presented in Section 3.5 and 3.6. We conclude the Chapter by describing on-going work on DVC (Section 3.7).

3.1 Reference image interpolation scheme

In DISCOVER [15], the decoder produces an estimation of the current WZF, let it be I_k, by using the adjacent KFs, let them be I_{k-1} and I_{k+1}, as shown in Fig. 3.1. After a spatial smoothing of the two KFs, a block-matching ME is performed between them. The resulting motion vector field $\mathbf{v}(\cdot)$ is split into a couple of fields $\mathbf{v}_B(\cdot)$ and $\mathbf{v}_F(\cdot)$ (pointing from k respectively to $k-1$ and $k+1$). This step is called bidirectional motion estimation (or sometimes vector splitting). In order to illustrate it, let us consider a block of the WZF centered in pixel $\mathbf{p}$. The split is performed by looking for the motion trajectory passing closest to $\mathbf{p}$. If the trajectories are modeled as linear, a block centered in $\mathbf{q}$ at time $k+1$, is centered in $\mathbf{q}+\frac{1}{2}\mathbf{v}(\mathbf{q})$ at time k. Then, we select $\mathbf{q}^*$ such that its position at time k is the nearest to $\mathbf{p}$, *i.e.*:

$$\mathbf{q}^*(\mathbf{p}) = \arg \min_{\mathbf{q}\in\mathcal{N}(\mathbf{p})} \left\| \mathbf{q} + \frac{1}{2}\mathbf{v}(\mathbf{q}) - \mathbf{p} \right\|^2 \tag{3.1}$$

where $\mathcal{N}(\mathbf{p})$ is the set of the block center positions near $\mathbf{p}$. The vectors $\mathbf{v}_B$ and $\mathbf{v}_F$ are defined as $\mathbf{v}_B(\mathbf{p}) = \frac{1}{2}\mathbf{v}(\mathbf{q}^*(\mathbf{p}))$ and $\mathbf{v}_F(\mathbf{p}) = -\frac{1}{2}\mathbf{v}(\mathbf{q}^*(\mathbf{p}))$. Next, the bidirectional motion refinement (BMR) is applied: let $B_\ell^{\mathbf{p}}$ be the block centered in $\mathbf{p}$, and belonging to the ℓ-th frame; we look for the vector correction $\mathbf{e} \in W = \{-1, 0, 1\}^2$ that gives the best matching between the blocks $B_{t-1}^{\mathbf{p}+\mathbf{v}_B(\mathbf{p})+\mathbf{e}}$ and $B_{t+1}^{\mathbf{p}+\mathbf{v}_F(\mathbf{p})-\mathbf{e}}$. BMR is applied a second time, reducing the block size and adapting the search area. Finally, the two motion vector fields are smoothed using a weighted median filter, and then used to compensate the previous and the next key frame. The resulting images are averaged to produce the SI. We observe that, even

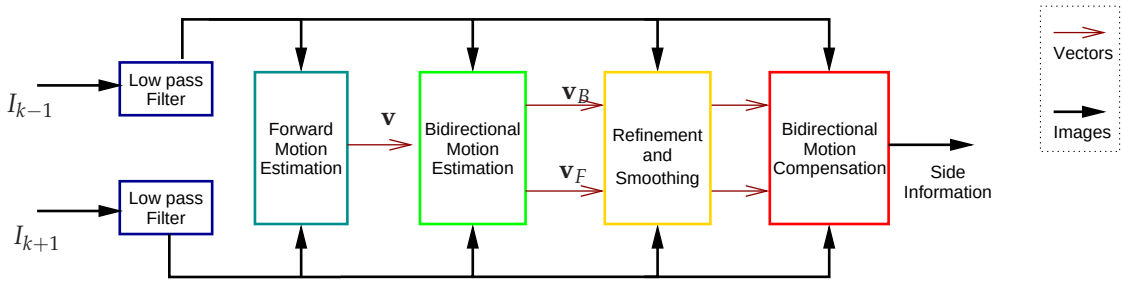


Figure 3.1: Discover side information generation scheme.

if DISCOVER uses a simple motion model (linear trajectories), yet it gives very good results in several cases.

3.2 Dense motion vector fields for DVC

This work was performed at TELECOM-ParisTech, in collaboration with prof. Pesquet-Popescu, dr. Miled and dr. Maugey. The results were published in [27, 28, 86, 89].

In this section we show how to modify the motion estimation steps in the DISCOVER side information generation process using the Cafforio-Rocca algorithm (CRA, see also 1.2.1).

We propose two variants: in a first one, the CRA is used to modify the bidirectional motion estimation step, using as initialization the backward and forward MVFs produced within DISCOVER. In the second we apply a modified version of the CRA to the forward ME, taking into account the image edge information. Both techniques resulted in an improved side information, and in a reduced coding rate for the same quality. We also considered the case where the CRA is jointly used with total variation-based estimation algorithms. For the sake of brevity, we do not report here the results, and refer the reader to [86, 89] for more details.

3.2.1 Using the CR on bidirectional MVFs

This new version of the CRA is used to refine the MVF $\mathbf{v}_B$ and $\mathbf{v}_F$ produced by the DISCOVER splitting step. There are two main difference with respect to the original CRA: first, the current image I_k is not available, so we must use the error between I_{k+1} and I_{k-1}; the second is that we perform a joint refinement of the two field. Our algorithm still consists in the initialization, validation and refinement steps, but they are modified to fit the new context; we also use a different scanning order: the DISCOVER blocks are scanned line-by-line, and within each block the pels are scanned in raster order.

Initialization. If $\mathbf{p}$ is the first pel in the block the initialization vectors $\mathbf{v}_B^{(0)}(\mathbf{p})$ and $\mathbf{v}_F^{(0)}(\mathbf{p})$ are those estimated for the current block by DISCOVER $\mathbf{v}_B^{\text{DISC}}$ and $\mathbf{v}_F^{\text{DISC}}$; otherwise, we use a weighted average of the left, up, and up-right neighboring vectors, with different weights if the neighbors are in the same block or not.

Validation. The validation step amounts to compute the motion compensated errors associated to $\mathbf{v}_{B,F}^{(0)}$, to the null vector and to $\mathbf{v}_{B,F}^{\text{DISC}}$, and to choose those with the least error. More precisely we compute the following quantities:

$$A = \left| I_{k+1}(\mathbf{p} + \mathbf{v}_B^{(0)}) - I_{k-1}(\mathbf{p} + \mathbf{v}_F^{(0)}) \right|, \qquad B = |I_{k+1}(\mathbf{p}) - I_{k-1}(\mathbf{p})| + \gamma,$$
$$C = \left| I_{k+1}(\mathbf{p} + \mathbf{v}_B^{\text{DISC}}) - I_{k-1}(\mathbf{p} + \mathbf{v}_F^{\text{DISC}}) \right|$$

If A [resp. B, C] is the least quantity, we use $\mathbf{v}_{B,F}^{(0)}$ [resp. the null vector, $\mathbf{v}_{B,F}^{\text{DISC}}$] as validated vectors. Note that, like for the original CR algorithm, a threshold γ is used to penalize the reset of the estimated vector. A high threshold causes less vector resets, producing more regular but maybe less accurate MVFs.

Refinement. In the last step, we refine the validated MVs $\mathbf{v}_B^{(1)}$ and $\mathbf{v}_F^{(1)}$ by adding a correction ($\delta\mathbf{v}_B$ and $\delta\mathbf{v}_F$). The cost function J, depending on both refinements, is defined as the squared bidirectional

motion-compensated error, with penalization on correction norms:

$$J(\delta\mathbf{v}_B, \delta\mathbf{v}_F) = [I_{k+1}(\mathbf{p} + \mathbf{v}_B^{(1)} + \delta\mathbf{v}_B) - I_{k-1}(\mathbf{p} + \mathbf{v}_F^{(1)} + \delta\mathbf{v}_F)]^2 + \lambda_B\|\delta\mathbf{v}_B\|^2 + \lambda_F\|\delta\mathbf{v}_F\|^2$$

We have been able to find an analytical solution to the problem of minimizing J. We introduce the following short-hands for the motion-compensated error and gradients:

$$\epsilon = I_{k+1}(\mathbf{p} + \mathbf{v}_B^{(1)}) - I_{k-1}(\mathbf{p} + \mathbf{v}_F^{(1)}) \qquad \phi_B = \nabla I_{k+1}(\mathbf{p} + \mathbf{v}_B^{(1)}) \qquad \phi_F = \nabla I_{k+1}(\mathbf{p} + \mathbf{v}_F^{(1)})$$

Then, we can show that the optimal refinements (found by setting to zero the partial derivatives of the first-order approximation of J) are:

$$\delta\mathbf{v}_B^* = \frac{-\epsilon\phi_B}{\lambda_B + \|\phi_B\|^2 + \frac{\lambda_B}{\lambda_F}\|\phi_F\|^2} \qquad \delta\mathbf{v}_F^* = \frac{\epsilon\phi_F}{\lambda_F + \|\phi_F\|^2 + \frac{\lambda_F}{\lambda_B}\|\phi_B\|^2}. \tag{3.2}$$

Previous equations further simplify since usually $\lambda_B = \lambda_F$, as discussed in the following. The resulting final formulas are formally very similar to those of the original algorithm, see Eq. (1.3).

Experimental results. We have performed preliminary experiments to determine the best value for parameters γ, λ_B and λ_F, by maximizing the quality of the SI. Complete results can be found in [27]; here we report the main conclusions. We have found that the threshold γ should be greater or equal than 50, so we use this value for the following. Moreover the parameters λ_F and λ_B should have very close values ($|\lambda_F - \lambda_B| < 0.1\lambda_B$). As a consequence, in the following we take $\lambda_F = \lambda_B$ and we drop the subscript. Finally, we look for the best value of λ. We have computed the SI PSNR over the test sequences for several values of the parameter between 1000 and 15000. The average PSNR performance are quite consistent for $\lambda \geq 5000$, with a maximum around 7500, which has been used as value for λ in the following.

With these values of parameters, we have compared our algorithm with DISCOVER by running them over the same test sequences and using several QPs for the KF coding. We observe that the proposed method is able to improve the WZF quality, up to over 0.6 dB in the average and to over 2 dB on the single image. The best results have been obtained for the "foreman" sequence, characterized by a complex motion. The gain is still interesting for sequences with more regular motion. Smaller gains are obtained when the movement is irregular or very small. We observe as well that the gain is smaller at very low rates: severely quantized KFs provide a less reliable gradient information, and thus a worse estimation of $\delta\mathbf{v}_B^*$ and $\delta\mathbf{v}_F^*$.

In the last set of experiments, we used the new SI compute the end-to-end RD performance for QP=31, 34, 37, 40. This was compared with the RD performance over the test sequences of the reference DISCOVER coder, and the results are compared using the Bjontegaard [16] metric. The proposed method allows some interesting rate reductions (3.5% for "foreman" and 2.0% on the average).

3.2.2 The CRA at the beginning of the estimation process

In this section we apply the CRA to the forward motion estimation step of DISCOVER, see Fig. 3.1, refining the vectors $\mathbf{v}$. These vectors are used in the **initialization** step: if $\mathbf{p}$ is the first position in the block, the vector $\mathbf{v}^{(0)}(\mathbf{p})$ is initialized as $\mathbf{v}(\mathbf{p})$. Otherwise, we use a weighted average of neighbors.

Validation. We compute three motion-compensated errors: the one associated to $\mathbf{v}^{(0)}(\mathbf{p})$, the non-compensated error, and the error for $\mathbf{v}(\mathbf{p})$, and we choose the vector with the least absolute error. As in the original algorithm, the non-compensated error is increased by a threshold γ in order to reduce the reset frequency.

Refinement We propose the following cost function:

$$J(\delta\mathbf{v}) = [I_{k+1}(\mathbf{p}) - I_{k-1}(\mathbf{p} + \mathbf{v}^{(1)} + \delta\mathbf{v})]^2 + \lambda\delta\mathbf{v}^T\mathbf{D}\delta\mathbf{v} \tag{3.3}$$

Like in the original algorithm, the correction should minimize the prediction error, under the constraint of a regularization condition. We improve the regularization conditions using the Nagel-Enkelmann constraint [99], that allows larger corrections across object boundaries. This is possible since $\mathbf{D}$ is the diffusion matrix:

$$\mathbf{D}(\nabla I) = \frac{1}{|\nabla I|^2 + 2\sigma^2}\left[\begin{pmatrix} \frac{\partial I}{\partial y} \\ -\frac{\partial I}{\partial x} \end{pmatrix}\begin{pmatrix} \frac{\partial I}{\partial y} \\ -\frac{\partial I}{\partial x} \end{pmatrix}^T + \sigma^2\mathbf{I}_2\right]$$

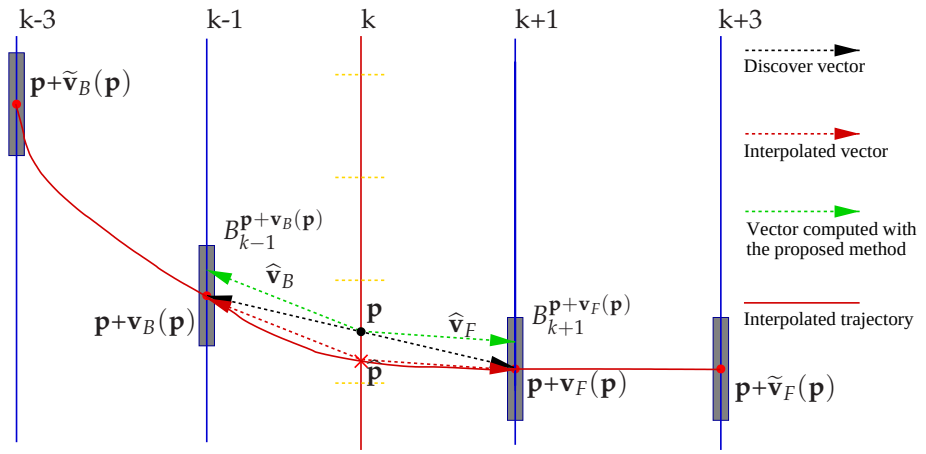


Figure 3.2: Proposed high order motion interpolation (HOMI) method.

In Eq. (3.3) we used the shorthand notation $\mathbf{D} = \mathbf{D}(\nabla I_{k-1})$. We notice that, in the homogeneous regions where $\sigma^2 \gg |\nabla I_{k-1}|^2$, the cost function becomes equivalent to the one used in the original algorithm, see Eq. (1.3). Our contribution here was to find an analytical solution to the problem of minimizing Eq. (3.3). We refer the reader to [28] for details. The result is:

$$\delta\mathbf{v}^* = \frac{-\epsilon\mathbf{D}^{-1}\phi}{\lambda + \phi^T\mathbf{D}^{-1}\phi} \tag{3.4}$$

Experiments. First, we have performed some experiments in order to tune the parameters λ, γ, and σ of the proposed algorithm. We refer the reader to [28] for the complete results, and report only the conclusions: the best values are for the parameter $\lambda = 2000$ $\gamma = 20$ and $\sigma = 50$.

With these values for the parameters, we compute the SI on four test sequences. Its quality, measured as PSNR with respect to original WZF, is better than the one obtained by DISCOVER. We obtain higher gain when there is high, regular motion like in "mobile" and "city" (up to more than 1.1 dB). When the motion is less regular we have a bit smaller but still significant gain (up to about 0.5 dB for "foreman"). Finally, some gains are still obtained for the *head and shoulder* sequence "eric", around 0.2 dB. The gain are slightly reduced for larger GOP sizes. We also compare this method with the one proposed in the previous section: the new method consistently outperforms the previous one (average improvement of 0.33 dB), suggesting that the earlier the CRA is introduced into the processing chain, the better the results are. Similar results are obtained for other GOP sizes.

Finally, we assess the impact of the new SI generation method in the complete DVC system comparing it to DISCOVER by the means of the Bjontegaard [16] metric at four QPs $(31, 34, 37, 40)$. We observe an average rate reduction of more than 5.9 % for the sequences of the test set. These results validate the proposed method, and, along with those shown in the previous Section, confirm the intuition that differential ME methods can be profitably employed in the framework of DVC.

3.3 High order motion interpolation

This work was performed in collaboration with prof. Pesquet-Popescu and Dr. Petrazzuoli. The results were published in [110, 112, 113, 115].

As shown in Section 3.1, DISCOVER performs a linear trajectory interpolation for deducing the object position in the missing WZFs. However, whenever the motion in the video deviates from this model the quality of SI is affected. So we propose to perform a *higher order motion interpolation* (HOMI) of object positions. This SI generation method is described in Fig. 3.2. Let us consider a block of the current WZF centered on the pixel $\mathbf{p}$. The DISCOVER algorithm provides us with the vectors $\mathbf{v}_B(\mathbf{p})$ and $\mathbf{v}_F(\mathbf{p})$ Let $B_{k-1}^{\mathbf{p}+\mathbf{v}_B(\mathbf{p})}$ and $B_{k+1}^{\mathbf{p}+\mathbf{v}_F(\mathbf{p})}$ be the corresponding block of pixels in the reference images. Now we want to refine the movement of this block by taking into account the images I_{k+3} and I_{k-3}. Therefore, we start

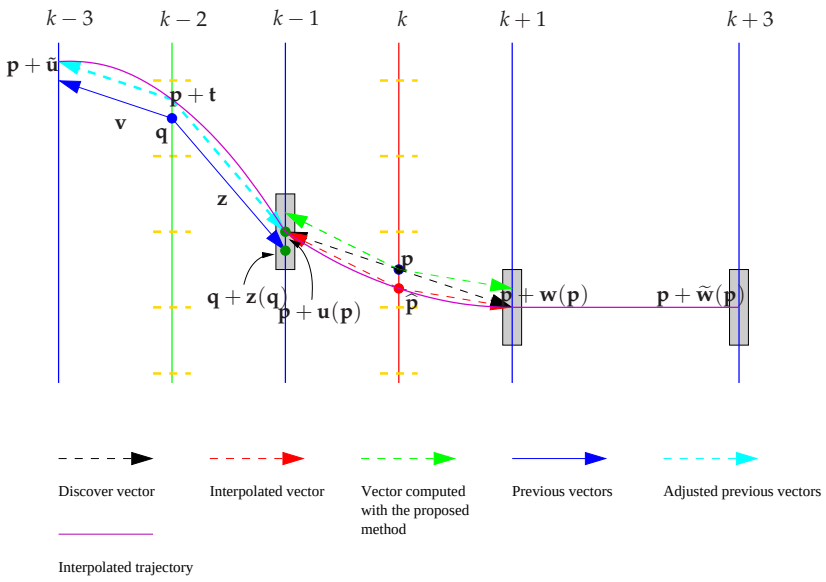


Figure 3.3: Fast HOMI for WZF estimation by exploiting the previous estimated MVF.

by looking for the position that the block $B_{k-1}^{\mathbf{p}+\mathbf{v}_B(\mathbf{p})}$ has in I_{k-3}, by using a regularized block-matching search: we look for the vector $\widetilde{\mathbf{v}}_B$ that minimizes:

$$J(\widetilde{\mathbf{v}}_B) = \sum_{\mathbf{q}} \left| B_{k-1}^{\mathbf{p}+\mathbf{v}_B(\mathbf{p})}(\mathbf{q}) - B_{k-3}^{\mathbf{p}+\widetilde{\mathbf{v}}_B}(\mathbf{q}) \right|^n + \lambda \|\widetilde{\mathbf{v}}_B - 3\mathbf{v}_B\| \qquad (3.5)$$

The regularization term penalizes too large deviations from the linear model. Likewise, we can find the vector $\widetilde{\mathbf{v}}_F$ allowing to match the block $B_{k+1}^{\mathbf{p}+\mathbf{v}_F(\mathbf{p})}$ with another block in I_{k+3}.

The second step of the proposed method consists in interpolating the positions of the current block in the four images. In other words we interpolate a vector function with the values $\mathbf{p}+\widetilde{\mathbf{v}}_B$, $\mathbf{p}+\mathbf{v}_B$, $\mathbf{p}+\mathbf{v}_F$ and $\mathbf{p}+\widetilde{\mathbf{v}}_F$ respectively at instants $k-3$, $k-1$, $k+1$, $k+3$, in order to find its value at instant k, be it $\widehat{\mathbf{p}}$. We use a piecewise cubic Hermite interpolation to find the position. As a consequence, the interpolated motion vectors are $\mathbf{v}_B(\mathbf{p})+\mathbf{p}-\widehat{\mathbf{p}}$ and $\mathbf{v}_F(\mathbf{p})+\mathbf{p}-\widehat{\mathbf{p}}$. These vectors are shown in dark red in Fig. 3.2. The last step consists simply in choosing the interpolated trajectory passing closest to the block center and in assigning the associated vector to the position $\mathbf{p}$, just as in DISCOVER. The difference is that the trajectory is interpolated using four points instead of two. The new vectors, shown in green, are:

$$\widehat{\mathbf{v}}_B(\mathbf{p}) = \mathbf{v}_B(\mathbf{p}) + \mathbf{p} - \widehat{\mathbf{p}}, \qquad \widehat{\mathbf{v}}_F(\mathbf{p}) = \mathbf{v}_F(\mathbf{p}) + \mathbf{p} - \widehat{\mathbf{p}}. \qquad (3.6)$$

Experimental results. We computed the SI with the HOMI method on several test sequences, and for various GOP sizes, see [113]. The proposed method allows good gains in side information quality, up to more than 0.5 dB. The corresponding Bjontegaard rate reductions range between 1.1 % and 3.3 %. We also notice that the highest gains are for a GOP size equal to 4. In fact, when KF are very close (GOP size = 2), linear interpolation is not very bad, so our gains are a little smaller, while when they are too far apart, any interpolation method would have a difficult task.

3.3.1 Variants of the HOMI algorithms

The complexity of the interpolation procedure described in the previous section can be reduced, because at the instant k, we have already estimated the MVF $\mathbf{v}$ from the frame I_{k-2} to the frame I_{k-3} and the MVF $\mathbf{z}$ from I_{k-2} to I_{k-1}. The new procedure, called FastHOMI, consists in the following steps (see Fig. 3.3):

1. **Initialization.** - We estimate $\mathbf{v}_B$ from I_k to I_{k-1} and $\mathbf{v}_F$ from I_k to I_{k+1} using DISCOVER.
2. **Motion estimation from I_{k+1} to I_{k+3}.** We perform a block matching motion estimation from I_{k+1} to I_{k+3} and we find the position $\mathbf{p}+\widetilde{\mathbf{v}}_F$.

3. **Motion estimation from I_{k-1} to I_{k-2} and from I_{k-2} to I_{k-3}.** We search for the vector $\mathbf{z}(\mathbf{q})$ that points in I_{k-1} to the position closest to $\mathbf{p} + \mathbf{v}_B(\mathbf{p})$. Then, we estimate the intersection of the trajectory in the frame I_{k-2} as the point $\mathbf{p} + \mathbf{t}$, with $\mathbf{t} = \mathbf{v}_B(\mathbf{p}) - \mathbf{z}(\mathbf{q})$. For the estimation of the intersection of the trajectory in I_{k-3}, we use the vector $\mathbf{v}(\mathbf{q})$. The intersection point will be $\mathbf{p} + \widetilde{\mathbf{v}}_B$, with $\widetilde{\mathbf{v}}_B(\mathbf{p}) = \mathbf{t} + \mathbf{v}(\mathbf{q})$.
4. **Interpolation.** Finally, we interpolate a vector function with the five values $\mathbf{p} + \widetilde{\mathbf{v}}_B$, $\mathbf{p} + \mathbf{t}$, $\mathbf{p} + \mathbf{v}_B(\mathbf{p})$, $\mathbf{p} + \mathbf{v}_F(\mathbf{p})$ and $\mathbf{p} + \widetilde{\mathbf{v}}_F(\mathbf{p})$ respectively, at the instants $k-3$, $k-2$, $k-1$, $k+1$ and $k+3$, in order to find its value at the instant k, which will be denoted by $\widehat{\mathbf{p}}$.
5. **Vector computation.** Given $\widehat{p}$, it is performed as in the previous section.

We observe that the complexity of FastHOMI is about the half of the original algorithm, but we expect slightly reduced SI quality.

We also designed an increased-complexity version of HOMI that is expected to have better performances. This new method simply doubles the MV density on the rows and on the columns, keeping the same block size for the matching. In other words, we use overlapping block matching. In order to tell apart the different methods, we use the following terminology: HOMI8 is the original version of the algorithm, since we have a vector for each 8×8-pixels block. HOMI4 is the overlapping version: we have four times as much vectors. Likewise for FastHOMI8 and FastHOMI4.

Now we can compare the different methods. We use the test sequences *book arrival*, *ballet*, *jungle* and *breakdancer* at a resolution of 384×512 pixels. We encoded the KFs with H.264/INTRA, using four QP values (31, 34, 37 and 40). For each of the four methods, we compute the SI PSNR difference (averaged along each sequence) with respect to DISCOVER. We refer the reader to [112] for the complete results. However we observe that in almost all cases, the quality of the side information is improved with respect to DISCOVER. The only exception is for GOP size equal to 8, when a good SI estimation is difficult, and all methods are almost equivalent. We observe that denser MVFs improve the SI quality for GOP size equal to 2, while they do not help in the case of long-term estimation. We ascribe this behavior to the difficulty of estimating images that are quite far from the references. Finally we observe that the fast versions of HOMI have fairly good performances, since the quality of the SI is almost unchanged in many cases, while the computational complexity is halved.

The last experiment consists in computing end-to-end performances of the proposed techniques when inserted into a complete DVC coder by using the the Bjontegaard metric [16]. The results (reported in [112]) are not surprising: the proposed methods are in general better than the reference DISCOVER, excepted for GOP size equal to 8, where they are practically equivalent. Moreover, even from the point of view of RD performances, denser MVF are better than sparser ones, and the fast version of HOMI are nearly as effective as the original algorithms. We remark that globally, the best technique is HOMI4, which allows rate reductions up to 8% with respect to the reference.

As a final observation, we note that increasing the quality of the side information does not mean always an increasing of the RD performances. For example, the HOMI8 method has a better side information than DISCOVER for GOP size equal to 8, but worser RD performances. This confirms the intuition that the PSNR with respect to the original WZF is not necessarily an accurate method for evaluating the SI quality, even though for the moment is the most common, since in most cases the RD performances are well correlated to the side information PSNR. This observation is at the basis of a theoretical study, originating a submitted journal paper [84].

Other variants of HOMI were implemented. In [115] we improved the RD performance for long GOPs using SI refinement. In [110] HOMI was adapted to the distributed coding of MVD video (see Section 1.5). In both cases we remarked improved performance with respect to the state of the art.

3.4 SI generation by local and global ME

This work was performed for the PhD thesis of A. Abou-El Ailah, in collaboration with co-supervisors dr. Dufaux, dr. Farah and prof. Pesquet-Popescu. The results were published in [3–5, 7–10]

In this section, we illustrate a method for enhancing the SI in transform-domain DVC. This solution consists in combining global and local SI at the decoder. The global motion parameters are computed at the encoder, while keeping a low complexity. For a given WZF, feature points of the original reference frames and of the original WZF are extracted by carrying out the Scale-Invariant Feature Transform (SIFT) [80] algorithm. Then, a matching between these feature points is applied. Next, we need to

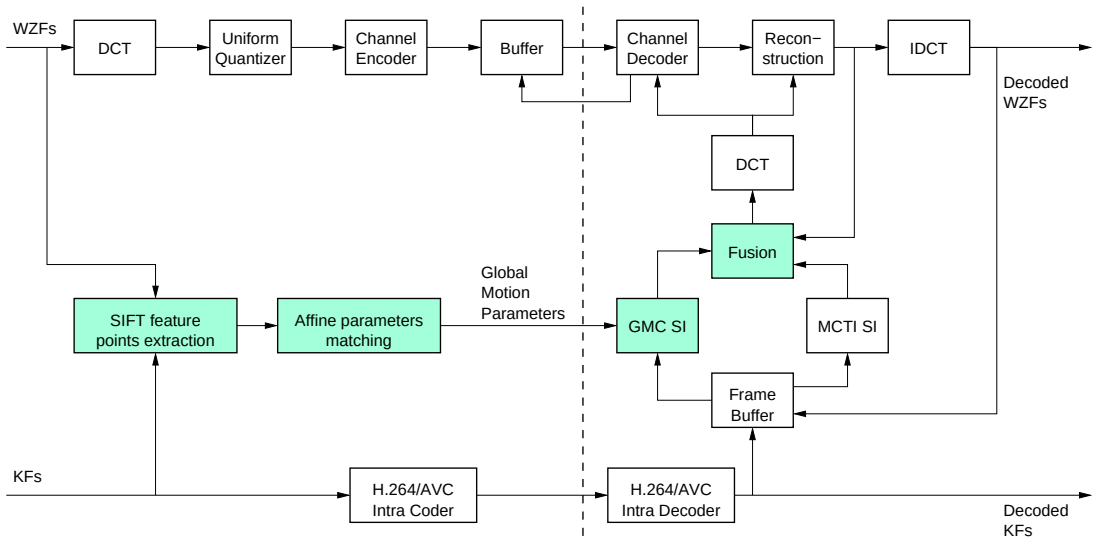


Figure 3.4: Overall structure of the proposed DVC codec based on GMC and MCTI.

find the matches which belong to the global motion in the scene. We propose an efficient algorithm which consists in eliminating iteratively the false matches due to local motion, in order to estimate the parameters of a global motion model between the current WZF and the backward or forward reference frame. The parameters of the global model are sent to the decoder in order to generate a SI based on Global Motion Compensation (GMC), and referred to as GMC SI. On the other hand, another SI is estimated using the same motion-compensated temporal interpolation (MCTI) as in DISCOVER codec [15]. This a "local" method, since it is based on the matching of (small) blocks. Then, a fusion of GMC SI and MCTI SI is performed; it will be referred to as the First Fusion SI (FFSI).

In addition, we also propose to successively improve the fusion of GMC SI and MCTI SI, after the decoding of each DCT band. Starting with the FFSI, the decoder reconstructs a Partially Decoded Wyner-Ziv Frame (PDWZF) by correcting the FFSI with the parity bits of the first DCT band. Two variations are proposed to enhance the fusion. The first one consists in improving the fusion after decoding the first DCT band, using the decoded DC coefficients of the PDWZF. It is important to note here that this method is very efficient in terms of computational load. The second method consists in improving the FFSI using the PDWZF along with the backward and forward reference frames. This method consists in re-estimating the false motion vectors obtained by the MCTI technique, after the decoding of each DCT band. Finally, the fusion between GMC SI and MCTI SI is iterated after each improvement of the PDWZF.

The block diagram of our proposed codec architecture is depicted in Figure 3.4. The shaded (green) blocks correspond to the four new modules introduced here: SIFT feature points extraction, affine parameters matching, computation of GMC SI, and fusion of GMC SI and MCTI SI. At the encoder, global motion parameters are estimated between the current original WZF and the original reference frames. First, SIFT feature points are extracted from the original reference and WZ frames. Second, global motion parameters are derived from matched feature points. This slightly deviates from the DVC paradigm, where WZF are encoded independently from KF. However, only very light computation is required in order to perform these operation, so the proposed system can be seen as a DVC-like low-complexity video coding system. At the decoder, the MCTI SI generation is based on block matching between the decoded reference frames (using DISCOVER) while the GMC SI is estimated by applying the global parameters on the decoded reference frames. Afterwards, the fusion of the two SI is carried out in order to obtain the FFSI. This fusion is performed as follows: for each 4×4 block, we compute the motion-compensated errors associated to GMC and MCTI on the available key frames, and we pick the block corresponding to the smallest error. The resulting image is improved by a first partial decoding, consisting in correcting the DC subband of the SI using the corresponding parity bits. Finally, the partially decoded WZF is used to validate the estimated motion vectors from MCTI.

Experimental results. A first set of experiments was devoted to validate the fusion technique. We refer to reader to our paper [5] for complete results. However we show here some obtained SI images.

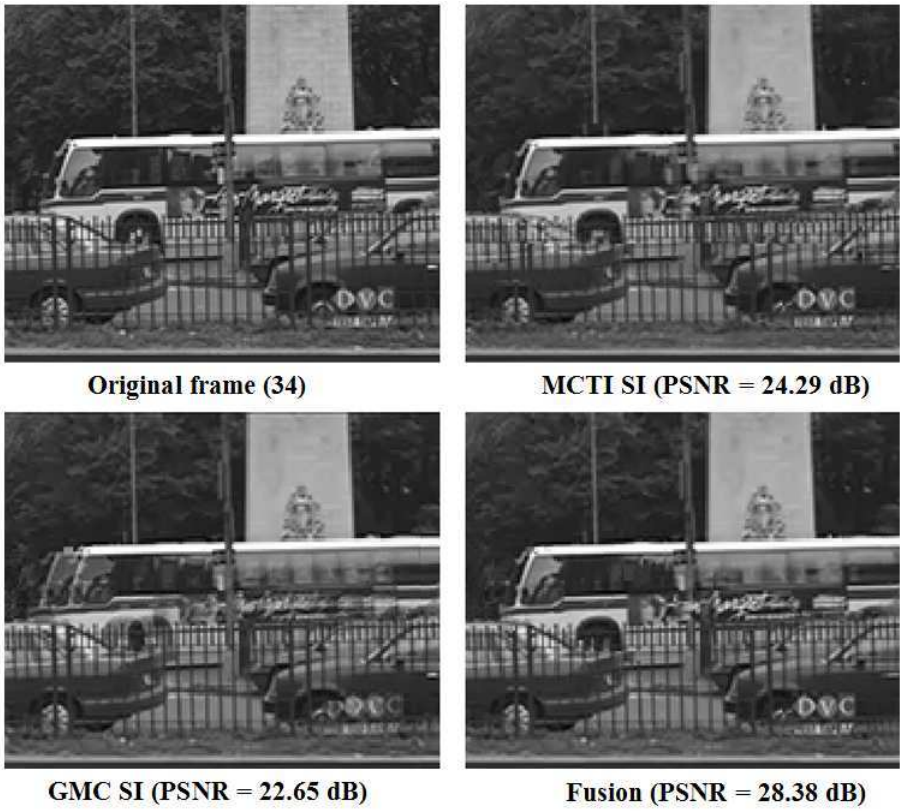


Figure 3.5: MCTI SI (top-right) - GMC SI (bottom-left) - Fusion of MCTI SI and GMC SI (bottom-right) - Frame number 31 of Bus sequence.

For example, in Fig. 3.5, we observe that MCTI (*i.e.*, DISCOVER) provides better SI than GMC. However, combining both allows a further quality gain of more than 4 dB.

Rate-distortion performances measured with the Bjontegaard metric show large rate reductions for most of the test sequences. For example, we obtained 22 % to 48 % rate reduction for "stefan", 14 % to 37 % for "bus", 7 % to 20 % for "foreman", and 1 % to 22 % for "coastguard". Globally, the proposed system has consistently better performances than H.264/INTRA, and almost always even better than H.264/No-motion. Moreover, the performance gap between the proposed DVC scheme and H.264/INTER is significantly reduced

3.5 Multiple description coding using DVC

This work was achieved in collaboration with prof. Pesquet-Popescu, dr. Greco and dr. Petrazzuoli. The results were published in [61, 65].

Multiple Description Coding (MDC) is a framework that allows an improved immunity to losses on error prone channels, when no back channel is available or when retransmission delay is not tolerable [53]. Using MDC, robustness is traded off with coding efficiency in terms of compression ratio for a given quality. Given an input signal – image, audio, video, etc. – an MD coder produces a set of independently decodable, mutually refineable description of equal (or almost equal) rate and importance; each description provides low, yet acceptable, quality; as soon as any further description is received, the quality of the reconstruction increases, independently on which description it is [59]. The decoding system used when all descriptions are received is referred to as *central decoder*; the one used when any subset of the description is received is referred to as *lateral decoder*.

Several ways to achieve MDC have been explored. Here we consider a simple temporal channel splitting approach: the original sequence is split up into even and odd frames; each sub-sequence is then separately encoded with a video coder (*e.g.* H.264/MPEG-4 AVC) to produce the two descriptions. Lateral

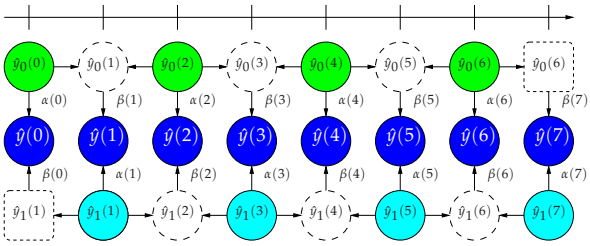


Figure 3.6: Structure of central decoder. Solid circles represent received frames; dashed circles represent interpolated frames. Horizontal arrows represent interpolation. Vertical arrows represent weighted sum. With $\hat{y}_n(k)$ we denote the k-th frame of description n. For each k, $\beta(k) = 1 - \alpha(k)$.

decoding is performed decoding the received description with an H.264 decoder, then reconstructing the missing frames via temporal interpolation; in our scheme, we shall use the DISCOVER [15,67] or the HOMI [113] technique of temporal interpolation.

When both decoded descriptions are available, central decoding is performed as a block-wise convex combination of the sub-sequences. For each frame, the relative weight α of each block in the received frame with respect to the corresponding block in the interpolated frame is computed at the encoder to minimize the distortion between the block in the original frame and the convex combination; then the sequence of α's is sent along with the descriptions as side information. A scheme of the central decoder is shown in Figure 3.6 The idea of reusing information from the lower fidelity version of a frame in central decoding by means of a convex combination has been originally proposed by Zhu et al. [138]; however, in their work, the lower fidelity frame was a transmitted B frame of lower hierarchical level, whereas we propose to use an interpolated frame generated at the decoder side.

Even though in theory the relative weight α is a continuous variable, experimental results show that quantizing α on three bits – i.e., eight levels – introduces a negligible error on the reconstructed sequence. We shall refer to the quantized version of α with $\bar{\alpha}$. In order to reduce the bit-rate needed to transmit the sequence of weights $\bar{\alpha}$, we adopt a context-based coding. We have found that there is some statistical dependence between the values of $\bar{\alpha}$ and the quantity E, defined as the MSE between received blocks and interpolated blocks. This quantity measures the similarity between the two descriptions. When they are very different, usually the received block is a better representation of the original one than the interpolated block. Thus, the mass probability function of $\bar{\alpha}$ is more concentrated around 1. Therefore, the context-based coding has a rate bounded by $H(\bar{\alpha}|E)$, and $H(\bar{\alpha}|E) < H(\bar{\alpha})$ because of the dependency. However, since the number of possible contexts (*i.e.*, of MSE values) is very high, we risk to incur into a *context dilution* problem, see Section 1.3. Given the relatively simple structure of the quantizer, this can be achieved by the means of as simple algorithms as the gradient descent, and we do not need more complex iterative techniques as the popular MCECQ [134] or MINIMA [23].

Results We used version 17.0 of the H.264/MPEG-4 AVC reference software, JM [69], to encode the sequences in our proposed scheme and in the scheme of Zhu et al. [138], which we took as reference. For the complete set of results, we refer the reader to our paper [61]. In the following we report some comparison w.r.t. the reference scheme.

A set of eight QPs has been selected (namely 22, 25, 28, 31, 33, 36, 39, and 42) in order to compare the RD performance of the two methods. According to the metric proposed by Bjontegaard [16], our technique has an average gain of 1.80dB in Y-PSNR corresponding to a reduction of 25.84% rate at low bitrates for central decoding. This improvement can be explained as the DISCOVER interpolation technique provides a reconstruction of the missing frames better than the very coarse version provided by the lowest level B-frames in the reference method. Also, even when such a coarse quantization is used, B-frames still need a certain bit-rate in order transmit motion vector, mode selection and so on.

It should be expected that the rate-distortion performance of a scheme based on temporal interpolation highly depends on the motion content of the video sequence. We have performed tests for several sequences with increasing motion content. It should be noticed that, whereas lateral decoding is severely impaired for sequences with fast movement, central decoding is still more efficient than the reference, since the low fidelity of lateral sequences is compensated with an appropriate value of α.

A performance comparison with the reference method as a function of the packet loss rate is illustrated in Figure 3.7-(a). Packet losses are modeled as independent and identically distributed Bernoulli

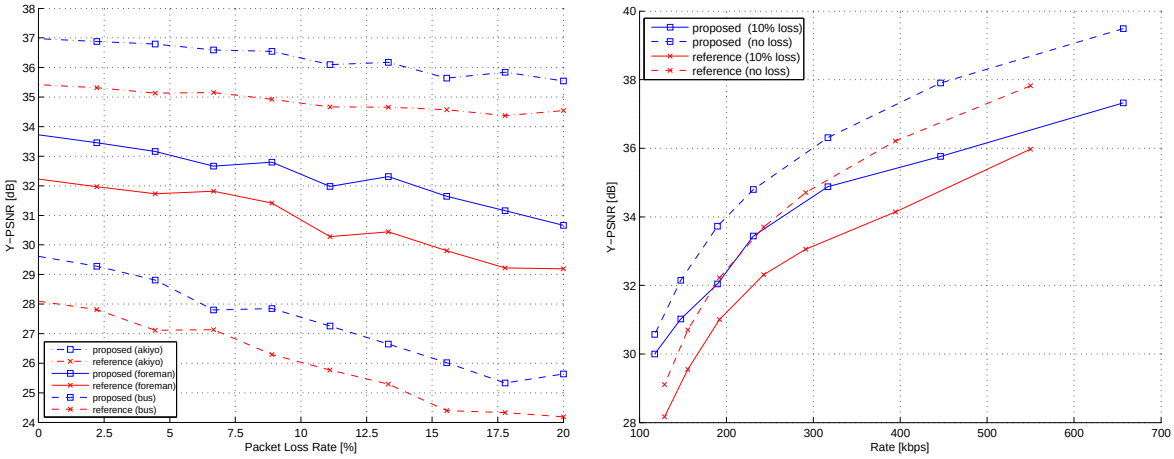


Figure 3.7: (a) Performance versus packet loss rate comparison for fixed bit-rate (200 kbps). (b) Performance versus bit-rate comparison for fixed packet loss rate (10%).

random variables with success probability p equal to the loss rate. As expected, sequences with higher motion content are more affected by packet loss; however, our technique consistently outperforms the reference method by 0.5–1.5 dB. In Fig. 3.7-(b), the two methods are compared over several bit-rates for a fixed packet loss rate of 10%, on sequence "foreman" (CIF, 30 fps). It can be seen how at low bit-rates out method outperforms even the lossless reference method.

Further improvements have been achieved introducing two modifications [65]: a rate-distortion optimized selection of weights α's; and the HOMI trajectory interpolation. This results into an average improvement of SI quality of 0.4 dB, and an average rate reduction of 5.4 % for the side decoders.

3.6 Interactive multiview streaming with DVC

This work has been carried out at TELECOM-ParisTech in collaboration with prof. Pesquet, dr. Dufaux and dr. Petrazzuoli. The results have been published in [109, 111].

As discussed in Section 1.5, multiple-views-plus-depth is emerging as the most relevant format for 3D video representation. Its huge redundancy have to be exploited in order to reduce the storage space on the server and the bandwidth used for transmission. These two requirements are equivalent in the context of non-interactive scenarios (like TV broadcasting), when all the video stored on the server will be sent to the user. For example, all the views are sent when MVD is used on a auto-stereoscopic display. On the contrary, the interactive multiview video streaming (IMVS) [39, 83] is a paradigm that enables the client to select interactively the view that he/she wants to display. Given this constraint, the server will send only the data needed to display the views according to the switch pattern decided by the user. However, the video is first encoded and stored in a server and afterwards it is sent to the clients. We would like to minimize both the storage space demanded by the compressed video and the bandwidth needed to interactively send the requested view to the user. These requirements are conflicting in the case of IMVS which makes the problem challenging.

In particular, let us consider a client switching from the view v_1 to v_2. The images of v_2 cannot be encoded using previous images from the same view, since the decoder will not have them. Therefore, two contrasting approaches emerges: on the one hand, we could reduce the bandwidth requirement if for any view, any image is coded N times, using any other views as reference. In this case the storage requirement is multiplied by a factor N (since at the encoding time we do not know which view the user will choose at any time), but the required bandwidth is minimized. This approach is called **Redundant P-frames**. On the other hand, we could encode each image only once but as an Intra frame, thus reducing the storage space. However in this case the bandwidth requirements are more demanding. This approach is called **I-frames** [39].

An alternative approach exploits principles coming from DVC. When a user switches to a novel view,

Name	Description
m	switching time: user wants view 1 up to $m-1$ and then view 2
k	time of the KF of the GOP affected by the switch on view 2. $k \leq m$
N	GOP size; for simplicity we only consider the case $N = 4$
I_n	decoded frame for the first view at instant n
J_n	an estimated frame of the target view, taken at time n and used as reference for motion interpolation for the remaining WZFs of the GOP
$\widetilde{I}_n$	estimation of the second view at time n obtained by depth-aided DIBR on I_n
$\widehat{I}_n$	$\widetilde{I}_n$ corrected by parity bits

Table 3.1: Notation

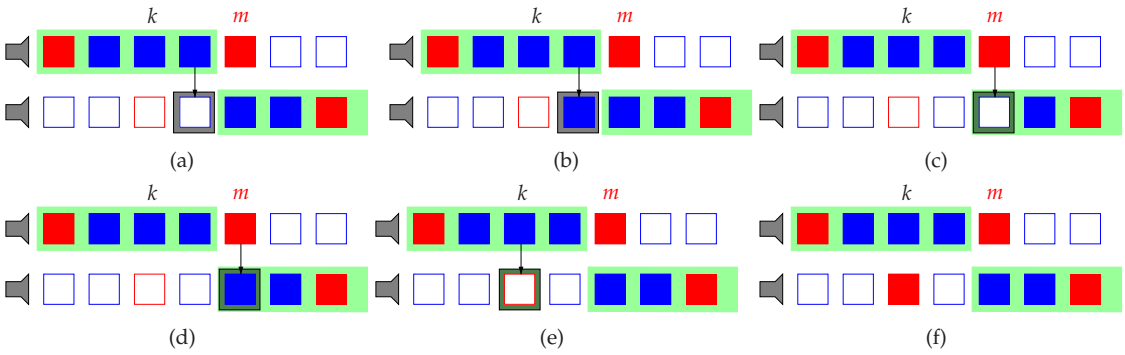


Figure 3.8: Strategies for IMVS with $N = 4$ and $m = k+2$: (a) A-DIBR; (b) A-DIBRc; (c) I-DIBR; (d) I-DIBRc; (e) GOPp; (f) noDIBR. The KFs are in red and the WZFs in blue. The frames with green background are displayed. The filled frames are sent by the video server to the user. The black arrow shows that DIBR is applied.

the next image to be displayed can either be a KF or a WZF. In the first case, it is decoded as usual; in the second one, we do not change the encoded representation of the frame stored on the server, *i.e.* the parity bits of the frame. This is interesting, since on one hand we improve the storage cost (only one version of any view is stored on the server), and on the other hand we do not send I-frames but only WZFs, which in theory could require as few bits as a P-frame, while in practice have an intermediate coding rate between I's and P's [58]. Cheung et *al.* [39] proposed to insert the Wyner-Ziv Frames, used as **M-Frames (Merge Frames)**, in the video stream, but they are interested in the case of multiview video coding *without* the depth information.

To the best of our knowledge, only a few papers deal with IMVS in MVD with the constraint of ensuring that the video playback is not interrupted during switching. In this work, we analyze and propose effective switching strategies for IMVD in MVD.

For the sake of simplicity, we consider the case where we have two Z-cameras, there is a KF out of N images, and the KFs on the two views are shifted by half a GOP. We have to choose what information is send to the decoder when the user makes a switch between two views, and how this is used. We propose six different techniques, exploiting DIBR and DVC. In order to describe the proposed methods, we introduce the following notation. We call m the instant of the switch, and k the instant when the GOP of frame m begins in view 2: in other words the previous KF for the target view happens to be in k. Because of the periodical GOP structure, we have to consider only the cases $m = k, k+1, k+2, \ldots k+N-1$. Moreover, we call J_n the reconstructed image for the second view *used for creating the side information*. This image will not necessary be displayed, but will be used, together with the KF $k+N$ that will always be sent, to produce the estimation for the WZF of the current GOP in the target view. We call this image the *reference image*. Finally we denote by $\widetilde{I}_n$ the estimation of the target view at time n obtained by only using depth-aided DIBR on the first view, and with $\widehat{I}_n$ an improved version of this image, obtained by using the parity bits of the second view image to correct $\widetilde{I}_n$. The introduced notation is summarized in Table 3.1 for ease of reading.

Now we can conceive several methods for IMVS using DIBR. They are shown in Fig. 3.8 for $N = 4$ and $m = k+2$, but this can be easily generalized to any N and m. A first one, that we call **Advance DIBR** (A-DIBR) consists in computing J_{m-1} at the decoder by using DIBR on the last received image from the first view. A variation of this method, that we can use only when $m \neq k+1$ consists in additionally sending parity bits in order to improve the reference image: $J_{m-1} = \widehat{I}_{m-1}$. This second method is

not necessarily better than the first one, since the higher reference quality is traded-off with a higher coding rate. We call this method **Advance DIBR + correction** (A-DIBRc). Another couple of methods is obtained if we perform DIBR at switch time instead of the previous instant. The resulting methods are called respectively **Immediate DIBR** (I-DIBR) and **Immediate DIBR + correction** (I-DIBRc). We can also think about preserving the GOP decoding structure in the second view. If we compute the previous KF on the GOP of the target view by DIBR, then we can use it to perform image interpolation in the GOP. In this case we just need to compute J_k as $\widetilde{I}_k$. We call this method **GOP preserving with DIBR** (GOPp). The last method does not use DIBR, and consists in directly sending the key frame of the current GOP for the target view. With respect to the GOPp method, it demands more rate but provides a better representation of the reference frame. It is called **GOP preserving without DIBR** (noDIBR).

Results. We refer the reader to our paper [109] for complete results. As a summary, we report here some interesting findings. On the average, the best method is A-DIBRc, achieving an average bit-rate reduction up to 13 % w.r.t. a solution that does not employ DIBR. However, according to the position of the switching point within the GOP, the performance of the methods change, and some of them become equivalent, while some other are not available. For example, for $m = k$ the best solution is noDIBR, which was expected: when the switching point coincides with a KF, the best is to send it directly. For $m = k + 1$ the best solution is A-DIBR/GOPp (The two methods coincide). For $m = k + 2$ the best solution is I-DIBRc: in this case A-DIBR is less effective because the reference frame is farther apart. For $m = k + 3$ the most effective method is A-DIBRc because it demands the smallest number of non-displayed frames. We observe that, when parity bits are available, it is always better to send them since the quality of the frame J_n used for creating the side information for the rest of the GOP is improved.

As a last test, we consider an adaptive method that uses the best strategy as a function of the switching point. We can obtain improved performances with respect to a given fixed strategy: from 1 % rate reduction with respect to A-DIBRc up to 6 % with respect to I-DIBRc.

3.7 On-going work on DVC

We are currently finalizing several studies on DVC-related topics. A summary of this on-going activity is given here.

1. We are working in the context of multi-view SI generation, using occlusion estimation and avoidance, adaptive fusion and validation methods. First results show improvements with respect to the state of the art. A journal paper has been submitted to the EURASIP Journal of Advances in Signal Processing [114] and is currently in the second review round.
2. We are developing a system for DVC of MVD video, with a new paradigm based on geometrical projection and rate allocation; a journal paper is in preparation.
3. A journal paper on segmentation-based DVC [6] has been submitted to IEEE Transactions on Circuits and Systems for Video Technology.
4. Finally, a journal paper on SI quality assessment [84] has just been accepted for publication in IEEE Transactions on Circuits and Systems for Video Technology.

4

Robust video distribution using cooperative networking and network coding

In this chapter we consider the problem of robust video delivery over unreliable networks. A first contribution is the ABCD protocol (Section 4.1), which allows the efficient construction of an overlay network over a mobile ad-hoc network (MANET), exploiting the inherent broadcast nature of the medium. The performance of this protocol are improved when we take into account the congestion/distortion trade-off for real-time applications (Section 4.2). In the last part of the Chapter (Section 4.3), we illustrate our contributions based on the network coding (NC) paradigm.

4.1 The ABCD protocol

This work was achieved within the PhD thesis of C. Greco, and published in [60, 62].

Video delivery on MANETs is a difficult problem because of the unreliable and time-varying nature of the underlying network. A popular solution consists in using multiple description coding on multiple paths, hoping that at each instant at least one description is available to the nodes. Building these paths (the so-called overlay network) in an efficient and resilient way is one of the most important challenges to achieve operational video delivery on these networks.

The ABCD protocol [60, 62] was introduced to enable the construction of an overlay network composed of N multicast trees, one for each description. We consider a multitude of cooperative mobile nodes connected by wireless links in a mesh topology, which can connect and disconnect abruptly. The application aims to deliver a video stream to the nodes thank to the building of an overlay that is efficient and robust. Efficient here means that the stream is delivered to all nodes with the minimum use of resources; robust, that the overlay is not severely affected by burst packet losses, due to collisions, node mobility, or abrupt disconnection of nodes. Each node aims to receive as many descriptions as possible, in order to maximize its video quality. Building and maintaining an overlay network inevitably requires that a number of packets is exchanged; in order to reduce this number, nodes have to be able to gather information without making explicit requests and to infer as much information as possible from any packet they receive.

We exploit the fact that the wireless medium is inherently broadcast, and thus each node can intercept any packet sent within its transmission area as long as it does not collide. However, using the wireless channel as a broadcast medium conflicts with the fact that the 802.11 MAC layer was mainly designed for one-to-one communications, while its one-to-many communication is known to be unreliable and inefficient [127]. To overcome this difficulty, we provided ABCD with an application-driven reliable broadcast [81]. In particular, even though packets are sent in broadcast, we do enforce an RTS/CTS/ACK exchange with one neighbor, specified by the protocol accordingly with the application logic, and referred to as *control peer*. The choice of the control peer is implemented as a biased random choice. Even though this technique cannot entirely prevent collisions, experimental evidence suggests that it reduces this phenomenon to the point of being negligible in our simulations. Such a reduction of the collision probability allows the protocol to perform better than the standard 802.11 in the trade-off between rate and diffusion area.

This trade-off is typical of self-limiting multi-hop broadcast in wireless networks [52], where a greater number of retransmissions increments the spread of the content (diffusion area), but reduces its throughput (rate) because of the limited channel capacity. Our improvement comes at the price of an increased

congestion, as the channel reservation has an overhead in terms of the time needed to transmit a packet; it is therefore advisable to use this technique when in the network only a small subset of nodes is transmitting at the same time, even if they send with a high bit-rate.

Once a reliable channel for broadcasting is available, the protocol design is very intuitive. The video source sends an advertisement message, and its neighbors reply with an attachment message for each description. Attachment messages are interpreted as a subscription to the description, so as soon as the stream has at least one subscriber, the source starts broadcasting the video packets for that description. We define a node active on a description when it is transmitting that description. Conversely, we say that a node *deactivates* on a description when it stops broadcasting it, willingly or otherwise. The subscribing nodes – that we define as the source's *children* – send periodical attachment messages to the source in order to keep it active. Each node that is not in the source's neighborhood, but that is in the neighborhood of at least one of its children, becomes aware of the availability of the descriptions since it receives its *peers*' attachment messages (a peer is any node other than the source). It then chooses one of these peers as its parent and sends it an attachment message; the node thus chosen will activate, starting to forward the video packets it receives from the source. The attachment messages sent by the newly subscribed node will now advertise the description within their neighborhood, generating other subscriptions; this process is reiterated, independently on each description, until all nodes have one parent per description. A node can have a different parent for each description; the overlay is thus formed of the superposition of N different trees. In conclusion, a node becomes aware of a path to the video resource as it intercepts an attachment message. Quite often, it actually intercepts attachments from multiple peers, and has to decide which peer provides the best path. Even if the node already has a parent, it could become aware of a better path, created by the connection of a peer or the mobility. Therefore, nodes need a metric for selecting the best path. To this end, we designed a metric that takes into account the above discussed objectives. Each node minimizes, over all candidate parents the following metric:

$$J = \omega_{\mathrm{h}} h + \omega_{\mathrm{a}} a + \omega_{\mathrm{d}} d - \omega_{\mathrm{g}} g - \omega_{\mathrm{q}} q, \tag{4.1}$$

where h is the number of hops to the source, a is the number of active peers in the node's neighborhood, d is the number of descriptions, other than the current one, for which the node is already subscribed to the candidate parent, g is the number of peers subscribed to the same candidate parent, q is the average signal-to-noise ratio of the link to its parent, and the ω values are a set of positive real weights, chosen experimentally so that the average PSNR of the video sequences decoded by the nodes is maximized. Experiments show that these weights need not to be adjusted at run-time, as the optimization is quite robust with respect to their choice.

Hop count minimization should always be preferred over all other parameters in the function, since it assures that the overlay graph is acyclic (*i.e.*, a proper tree). Also, it is beneficial to the minimization of the end-to-end delay. As a result, in an overlay generated by ABCD, a node cannot have, in a steady state, a peer in its neighborhood whose hop-count is smaller than that of its current parent, since in that case it would simply switch parent in order to prevent loops. The number of active nodes per neighborhood is also minimized, for two reasons: reducing the number of packets injected in the network, hence the congestion, and reducing the total amount of resources demanded to the nodes, which pay an energy cost to relay a description. The protocol also aims, by minimizing $\omega_{\mathrm{d}} d$, to ensure path diversity among the descriptions, which is advisable for both fairness and robustness. We also note that the term $-\omega_{\mathrm{g}} g$ implies that the nodes try to maximize the number of peers subscribed to their same parent (*siblings*), in order to concentrate subscriptions on fewer active nodes, making deactivation more frequent. As a result, the overlay trees generated by ABCD tend to be short and wide, and the number of active nodes tends to be small. This allows to mitigate the collision problem in the ABCD protocol.

However we implemented a number of other techniques in order to reduce the collision probability. In order to make it unlikely that two (or more) active siblings rely a video packet at the same time, we used a random assessment delay (RAD, [81]). To reduce the collision probability among video packets belonging to different descriptions, the source relay them as temporally far apart as possible.

We have extensively tested ABCD using a simulation environment (*ns2*) which models the 802.11 MAC/PHY layers under several conditions of node density, number of nodes, and stream bit-rate [60]. Also, two different mobility models have been experimented: Random Way-point and Reference Point Group Mobility [65]. The protocol has proven to be able to ensure that 100 % of the nodes receive almost all frames of all descriptions for a node density up to 20 nodes per neighborhood, which is three times as high as the optimal density (in the sense of the trade-off between the number of hops to reach a

destination and the collisions occurring at each node) [119]. The average delay is kept in the order of the hundreds of milliseconds as the topology is slowly changing, but the maximum delay can have much higher peaks if the topology is changing quickly, *e.g.*, a flash-crowd or a high mobility happens. We refer the reader to our papers [60, 62, 65] for full experimental results.

4.2 Congestion-distortion optimization

This work was achieved within the PhD thesis of C. Greco, and published in [63].

In most scenarios, ABCD performs well in terms of availability, robustness, scalability, and presents a low and stable latency. However, in large dense networks it is not able to abide by a stringent low-delay requirement. This is a common problem in the context of video streaming over MANETs, due to the fact that, even under optimal assignment of transmission ranges and traffic patterns, the throughput of each node in a wireless network diminishes to zero as the number of users is increased [68].

In this Section, we illustrate a congestion control framework for real-time multiple description video multicast over wireless ad-hoc networks, in the hypothesis of cooperative nodes. This framework includes models for congestion and distortion that take into account both the video stream coding structure and the unavoidable redundancy of the overlay network; it also provides the MAC layer with video-coding awareness, and thus makes possible to perform a Lagrangian optimization of congestion and distortion. This framework can be integrated into any tree-based video streaming protocol for MANETs to improve its performance; we show here that, if integrated into the ABCD protocol, it attains a significant reduction of both average (over time and nodes) and maximum end-to-end delay, maintaining a delivery rate close to 100%. The proposed protocol adjusts its parameters on-the-fly, and is based on a distributed estimation of both the network topology, in order to capture the multiple paths that a video packet may follow, and the channel conditions, in order to estimate the effects on end-to-end delay. This information is propagated in an efficient and compact way through the network, leading to significant improvements in terms of both delay and objective video quality, as demonstrated by the simulations.

The Congestion Distortion problem. When node density is very high or a sudden change in topology occurs, ABCD and, to a greater extent, generic-purpose protocols show that the average delay may become so high that some video frames are received beyond their playback deadline; in the following we shall assume that the maximum accepted delay from the video source to the end user is in the order of one hundred milliseconds, and the total bit-rate of the stream is in the order of a few megabits per second.

To reduce the delay in ABCD, we introduce a Congestion/Distortion Optimization (CoDiO) criterion in the per-hop forwarding of the protocol; namely, we adjust the retry limit to be passed to the RTS/CTS mechanism of the MAC, in a Co-Di optimized fashion. CoDiO is an approach already proven viable in the design of cross-layer protocols for video streaming on MANETS [121]. We start from the observation that the congestion/distortion trade-off can be adjusted by tuning the retry limit k in the RTS/CTS mechanism. Small values of k would reduce the congestion, since less requests are sent to try and obtain the channel, but the expected distortion would increase, as it would increase the probability of not obtaining the channel and being unable to send the current packet. On the other hand, higher values of k would lower the expected distortion, since the probability of sending the packet is higher, but would also imply a higher congestion due to the channel occupation. We end up with a multi-objective minimization problem; specifically, for each video packet, we want to find the optimal value k^* for the retry limit, defined as:

$$k^* \triangleq \arg\min_{k\in\mathbb{N}} \{D(k) + \lambda C(k)\}, \tag{4.2}$$

where $D(k)$ is the expected total distortion over a set of frames depending on the current packet (*i.e.*, contained in the packet or predicted upon it), for all the nodes in the sub-tree rooted in the current node, and $C(k)$ the expected congestion of the channel seen by the current node, both resulting from the retry limit k for the current packet. The multiplier λ has been determined experimentally to minimize the distortion for a given delay constraint.

While the congestion model can be computed locally without need to propagate information through the overlay, since it depends on the channel that the nodes can observe directly, the distortion model offers several challenges. A missing packet affects in general several frames of the decoded video sequence; moreover, the effects are in general different for each node, depending on its reception of the

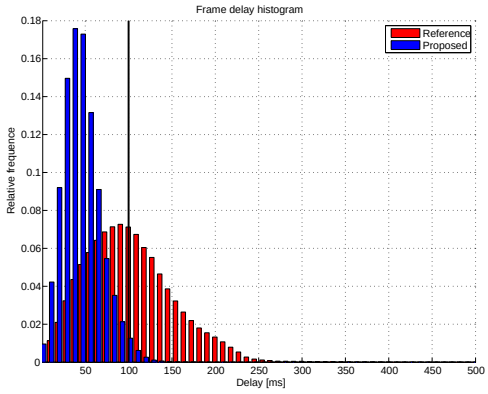


Figure 4.1: Histogram of frame delay. The vertical bar marks the maximum delay for frame decoding in the case of conversational pattern.

same packet from another path and of the complementary description. Finally, the effects of a loss propagate along the multicast tree, while a node can only communicate with its direct neighbors. Our main contribution amounts to a reliable and fast estimation of these quantities, performed in a distributed and adaptive fashion in the network

Distortion model. The distortion resulting from any single decision is modeled by splitting the descendants of a node n in four categories, according to the effect of its decision on their distortion (for simplicity, we consider in the following the case where only 2 description, d_0 and d_1 are used in the MDC system). If node n stops sending description d_0, each node in the sub-tree rooted in n belongs to one of the following sets:

- $\mathbf{S}_c$, nodes able to receive both descriptions even if node n stops sending d_0;
- $\mathbf{S}_0$,nodes able to receive only d_0 if node n stops sending d_0;
- $\mathbf{S}_1$, nodes able to receive only d_1 if node n stops sending on d_0;
- $\mathbf{S}_f$, nodes unable to receive either description if node n stops sending on d_0.

Using these groups, the estimation of the effect of the decision on the total distortion becomes easier. In particular in our paper we show how to estimate the group sizes efficiently and reliably, using the concept of available alternative path. Moreover the proposed technique gathers local information that is refined during its propagation in the network. As a result, errors do not propagate but rather they tend to be corrected during the process. Other contributions of our paper include a dynamic delivery ratio estimation and a locally adaptive congestion estimation.

Results. Here we present the results of the performance tests of the proposed CoDiO extension of the ABCD protocol in comparison with the conventional ABCD implementation (described in Sec. 4.1). A set of nine video sequences (CIF at 30 fps), concatenated and then looped to match the total simulation time of 300 s, has been encoded in multiple descriptions using our MDC technique (see Section 3.5), with a coding rate of about 1.8 Mbps, resulting in an average PSNR of 39.94 dB for central decoding and 35.20 dB for side decoding.

The MANET has been simulated using the *ns2* discrete event simulator [103]. The Lagrangian multiplier in the minimization problem (4.2) has been found experimentally, by maximizing the average quality (in terms of PSNR) of the decoded sequences, to have a value of $\lambda = 1.4$. In the experiments, we compared ABCD and ABCD+CoDiO in a network with 100 nodes and a density of 40 nodes per neighborhood. This is an extremely high density, chosen in order to appreciate the capability of the proposed framework to deal with very harsh conditions of the network. In Fig. 4.2, we compare the histogram of the frame delay for the two versions of the protocol, collected from all the nodes in the simulation. Note that any frame with a delay higher than 100 ms is dropped. We observe that, in the reference version, more than one half of the frames are too late to be decoded (55 %), while in the proposed version, only a light tail of the histogram (2.7 %) crosses the deadline. The video quality is also improved, since the reference technique uses central decoding for 73 % of frames and side decoding for 19 %, while the proposed technique uses central decoding 94 % of frames and side decoding for 5 % (concealment is used for the remaining frames). With the proposed MDC technique, this is translated into an increase of the

average PSNR per node of 1.5 dB. Moreover, with the new protocol all nodes achieve a very high quality: even though some frames are decoded with a relatively small PSNR, this hardly happens repeatedly to the same nodes. In summary, we find a significant gain both in terms of PSNR and in average end-to-end delay, while the delivery rate is kept close to 100 %, making the technique suited for conversational video applications over MANETs. These results have been obtained in experimental conditions of high bit-rate, high density, and large number of nodes, *i.e.*, conditions prone to generate a severe congestion on the channel; also, a stringent constraint on delay has been imposed. Tests have been performed in less harsh scenario as well, but – even though the proposed technique is never out-performed by the reference technique – the gain is less significant in situations where congestion is less relevant (because a longer delay is accepted) or less likely to occur (because the node density and the bit-rates are small); this depends on the fact that this framework is designed for congested networks, and is unnecessary in more tolerant and less crowded networks.

4.3 Network coding

This work was carried out for the PhD of I. Nemoianu, in collaboration with prof. Pesquet-Popescu, prof. Castella and dr. Greco. The results were published in [64, 100–102]. A journal paper is also in preparation.

Network Coding (NC) [12] has recently been investigated as an alternative to classical routing for multicast streaming. Using NC, a multi-hop communication is relayed at intermediate nodes by sending combinations of the received messages, rather than mere copies. An interesting application of NC is to grant partial loss immunity to data streams in unreliable wireless networks [75]. Using Random Linear Network Coding (RLNC) [71], a technique in which nodes send random linear combinations of their received packets, with coefficients taken from a finite field of proper size, the communication can be routed in unreliable networks with dynamically varying connections with no need for node coordination. A practical implementation of RLNC [41], called practical network coding (PNC), can be achieved segmenting the data into groups of packets called generations and combining only packets belonging to the same generation. All packets in a generation are jointly decoded as soon as enough linearly independent combinations have been received, by means of simple linear system solving. Since the coefficients are taken from a finite field, perfect reconstruction is assured regardless of the precision of the implementation. Recently [126], it has also been proposed to apply NC to video content delivery, dividing the video stream into layers of priority and providing unequal error protection for the different layers via PNC. Layered coding requires that all users receive at least the base layer, hence all received packets must be stored in a buffer until a sufficient number of independent combinations are received, which introduces a decoding delay that is often unacceptable in real-time streaming applications. There exist several techniques aimed to reduce the decoding delay, proposed by both the NC and the video coding communities.

From a network coding perspective, a viable solution is to use Expanding Window Network Coding (EWNC) [132]. The key idea of EWNC is to increase the size of the coding window (*i.e.*, the set of packets in the generation that may appear in combination vectors) for each new packet. Using Gaussian elimination at the receiver side, this method provides instant decodability of packets. Thanks to this property EWNC is preferable over PNC in streaming applications. Even though PNC could achieve almost instant decodability using a small generation size, this would be ineffective in a wireless network, where a receiver could be surrounded by a large number of senders, and if the size of the generation is smaller than the number of senders, some combinations will necessarily be linearly dependent. On the other hand, EWNC automatically adapts the coding window size allowing early decodability, and innovation (*i.e.*, linear independence) can be achieved if the senders include the packets in the coding window in a different order. However, these orders should take into account the RD properties of the video stream.

NC and MDC. Another possibility is to employ NC jointly with multiple description coding. In a first contribution, appeared in [100], we formulate the problem of broadcasting a video stream encoded in multiple descriptions on an ad-hoc network in terms of finding an optimal set of combination coefficients. Then, we introduce an objective function that takes into account the effect that decoding a given number of descriptions has on the total distortion. This framework has been integrated with the ABCD protocol that provides both an acyclic overlay network and knowledge of the neighbors' state. We have compared the performance of this technique with the well-known random linear coding technique. We

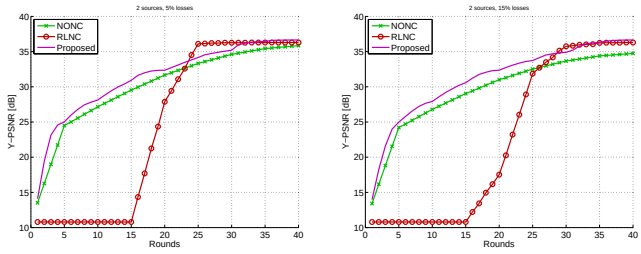


Figure 4.2: Comparison of the average Y-PSNR of the decoded sequences, for $M = 2$ sources and packet loss probability $p = 5\,\%$ (left) and $p = 15\,\%$ (right).

observe that the limitations of the generation size to the number of descriptions, imposed by the delay constraints, severely affect the performance of the reference technique, which as a result is consistently outperformed by the proposed approach. More precisely, the proposed technique performs on average about 2 dB better than RLNC. We refer the reader to [100] for more results.

In a second contribution, appeared in [64], we propose to jointly use EWNC and video MDC, in order to provide a robust video delivery over an unreliable wireless network, without any need for centralized control or feedback channel. To this end, we design a Rate-Distortion Optimized scheduling algorithm that, at each sending opportunity, selects which video packet has to be added to the coding window in such a way as to maximize the expected video quality perceived by the receiver. Since the wireless medium is inherently broadcast, we want to exploit the possibility of the receiver being exposed to multiple senders. In other words, we assure that the senders transmit innovative coding vectors even though they do not coordinate their actions. Introducing this optimized scheduling strategy allows to reach a good quality for the received video with much less transmission rounds than the reference scheduling strategy (see [64]).

NC and multi-view video coding. Similar ideas have been employed for multi-view video coding [101]. We consider again EWNC over wireless networks, but we aim at the robust delivery of multi-view video. Again, we design a Rate-Distortion Optimized, based on the known dependencies among data units in a multi-view video stream. We introduce the concept of clusters of video frames. A cluster of frames is made up of frames that affects similarly the total rate and distortion of the multi-view video. For example, the frames that are the first of a GOP in a non-base view, affect similarly the remainder of the view, so they are typically in the same cluster. The scheduling among clusters is the same for all the node; however, in order to achieve the needed degree of randomness without deviating too much from the optimal scheduling, each node select randomly the order within a cluster.

In our simulations, the proposed approach has proven to be able to deliver an acceptable video quality to the receiver in a shorter number of rounds than the reference techniques. As an example, in Fig. 4.2, we report a comparison with the reference technique. We observe that, thanks to the variety in the scheduling, our technique is able to reduce the number of linearly dependent coding vectors, and is therefore able to provide a better video quality (in terms of Y-PSNR) in fewer rounds.

Using blind source separation to improve NC. In practical Network Coding approaches, the random coefficients for packet combinations must be added to the packet as headers [41], incurring an overhead that can be prohibitive if the maximum packet size is small. One could resort to a blind source separation (BSS) approach to find these coefficients without sending them. BSS [36, 43] consists in recovering a set of source signals $\mathbf{S}$ from a set of mixed signals $\mathbf{X} = f(\mathbf{S})$, also referred to as observations, without knowing the sources themselves nor the mixing process parameters. In this framework, the observations are the received packet, and the source signals are the original packets. Our main idea is to increase the discriminating power of the algorithm by pre-processing the sources with an error detecting code. The code should be such that the probability of a mixture belonging to the code is small. Also, the code cannot be linear, otherwise mixtures would always belong to it; we therefore consider only non-linear codes. A simple example of non-linear code is the odd-parity bit-code, that is obtained by inverting the parity-bit of the packet. This very simple approach, coupled with a minimum-entropy search, allow very good demixing performances: the failure rate (*i.e.* the rate of unsuccessful demixing) is reduced to less than the half with respect to a reference technique without the odd-parity bit. We refer the reader to our work [102] for more complete results. Future works are devoted to even more efficient demixing using more complex codes than the simple odd-parity bit.

5

Conclusions and perspectives

In this manuscript, the main research results obtained since the achievement of the PhD degree have been resumed. Some very general conclusion can be drawn at this point. A first one is about the importance of the signal model when designing a compression algorithm. The more the model fits the underlying signal characteristics, the more the encoding algorithm is potentially effective. Of course, finding a good model is the most difficult part of the problem. Our results showed that one can obtain improvements in video coding if motion is correctly represented. For the emerging framework of 3D video, taking into account inter-view and inter-component redundancy appears to be equally very important. As for image compression, even though intuition tells us that object-based compression has potential advantages over "flat" compression, an extensive analysis of the problem revealed that the losses related to the irregular shape or to the adaptive filters can easily surpass the gains, unless particular class of images are considered.

As far as the robust representation is concerned, once again motion modeling has proven to be one of the key challenges in DVC and in MDC. New streaming protocols as ABCD can be improved if they are designed keeping in mind the target application *i.e.* the transport of video information: here, a model of distortion and congestion propagation becomes of primary importance.

Building on the achieved results, new perspective for upcoming research are opened. In particular, it is envisaged to keep working on compression, considering the new challenges related to 3D video: compression efficiency, new forms of interactivity, robustness with respect to the loss of information unreliable channels. The goal is an interactive, immersive and proactive communication; once again, the key tools are the mathematical models able to capture the true nature of this type of data.

Immersive Communication: Compression of 3D video. This research theme focuses on the compression of multi-view video with depth information. It is an extremely redundant format: in addition to the usual spatial and temporal correlation, there is correlation between the different views, and the between each image and its depth map. New compression techniques are then possible: a first approach consists in providing innovative rate-distortion models; another could be the representation of object contours in depth maps with advanced mathematical models such as the elastic curves. In the medium and long term, new formats of representation will be considered, such as holoscopic or holographic formats.

Interactive Communication: IMVS. One of the most interesting applications of MVD is the interactive multiview streaming (IMVS). The problem of IMVS is very recent, and very few solutions exist for the MVD case. We intend to continue the exploration of the DVC-based approach, and at the same time, consider the joint use of new coding techniques based on complex movement models.

Proactive communication: Social Network Coding. Today the users of multimedia services have a more and more active role: they create, share and disseminate content, usually within a circle of users having the same interests, called "friends": this is the paradigm of social networks. In such a network, users interested in a specific content have the ability to retrieve pieces of information from friend nodes, and at the same time, they can provide information to other nodes. This scenario can greatly benefit from the use of new techniques of network coding (NC), above all when network resources are limited. The most effective solutions will probably take into account the content to share, the network status, along with contextual information and the interaction history between users. To meet this challenge, one promising idea is to design the network in a socio-centric manner, where the design and optimization of NC and transport policies are influenced and changed according to social interaction between users, rather than just to the communication between nodes. For example, the content shared by a so-called "trendsetter" will be released more quickly than the one shared by a user with only a few

"friends". Another case of potentially important use is the one of advanced video services on mobile ad-hoc wireless networks. The scientific objective is to show that it is possible to develop new tools based on the theory of network coding in the framework of the joint optimization of network and video aspects, and that they can successfully be used for video streaming, particularly for the specific aspects related to 3D video.

Bibliography

[1] A. Aaron, R. Zhang, and B. Girod. Wyner-Ziv coding of motion video. In *Asilomar Conference on Signals and Systems*, Pacific Grove, California, Nov. 2002.

[2] M. Abid, M. Cagnazzo, and B. Pesquet-Popescu. Image denoising by adaptive lifting schemes. In *European Workshop on Visual Information Processing*, volume 1, pages 108–113, Paris, France, 2010.

[3] A.-b. Abou-Elailah, F. Dufaux, J. Farah, and M. Cagnazzo. Fusion of global and local side information using support vector machine in transform-domain distributed video coding. In *European Signal Processing Conference (EUSIPCO 2012)*, Bucharest, Romania, Aug. 2012.

[4] A.-b. Abou-Elailah, F. Dufaux, J. Farah, M. Cagnazzo, and B. Pesquet-Popescu. Successive refinement of side information using adaptive search area for long duration gops in distributed video coding. In *International Conference on Telecommunications*, Jounieh, Lebanon, Apr. 2012.

[5] A.-b. Abou-Elailah, F. Dufaux, J. Farah, M. Cagnazzo, and B. Pesquet-Popescu. Fusion of global and local motion estimation for distributed video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(1):158–172, Jan. 2013.

[6] A.-b. Abou-Elailah, F. Dufaux, J. Farah, M. Cagnazzo, A. Srivastava, and B. Pesquet-Popescu. Fusion of global and local motion estimation using foreground objects for distributed video coding. 2013. In preparation.

[7] A.-b. Abou-Elailah, J. Farah, M. Cagnazzo, B. Pesquet-Popescu, and F. Dufaux. Amelioration progressive de l'information adjacente pour le codage video distribue. In *GRETSI 2011*, Bordeaux, France, Sept. 2011.

[8] A.-b. Abou-Elailah, J. Farah, M. Cagnazzo, B. Pesquet-Popescu, and F. Dufaux. Improved side information generation for distributed video coding. In *EUVIP2011*, Paris, France, July 2011.

[9] A.-b. Abou-Elailah, J. Farah, M. Cagnazzo, B. Pesquet-Popescu, and F. Dufaux. Successive refinement of motion compensated interpolation for transform-domain dvc. In *EUSIPCO2011*, Barcelona, Spain, Sept. 2011.

[10] A.-b. Abou-Elailah, G. Petrazzuoli, F. Dufaux, J. Farah, M. Cagnazzo, and B. Pesquet-Popescu. Side information improvement in transform-domain distributed video coding. In *SPIE Application of Digital Image Processing XXXV*, San Diego, California, Aug. 2012.

[11] M.-A. Agostini, M. Cagnazzo, M. Antonini, G. Laroche, and J. Jung. A new coding mode for hybrid video coders based on quantized motion vectors. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(7):946–956, July 2011.

[12] R. Ahlswede, N. Cai, S.-Y. Li, and R. Yeung. Network information flow. *IEEE Transactions on Information Theory*, 46(4):1204–1216, 2000.

[13] T. André, M. Cagnazzo, M. Antonini, and M. Barlaud. JPEG2000-compatible scalable scheme for wavelet-based video coding. *EURASIP Journal on Image and Video Processing*, 2007(1):9–19, 2007.

[14] T. André, M. Cagnazzo, M. Antonini, M. Barlaud, N. Bozinovic, and J. Konrad. (N,0) motion-compensated lifting-based wavelet transform. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 121–124, Montreal, Canada, May 2004.

[15] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret. The DISCOVER codec: Architecture, techniques and evaluation. In *Proceedings of Picture Coding Symposium*, 2007.

[16] G. Bjontegaard. Calculation of average PSNR differences between RD-curves. In *VCEG Meeting*, Austin, USA, Apr. 2001.

[17] S. Boltz, É. Debreuve, and M. Barlaud. A joint motion computation and segmentation algorithm for video coding. In *Proceedings of European Signal Processing Conference*, Antalya, Turkey, Sept. 2005.

[18] E. Bosc, V. Jantet, M. Pressigout, L. Morin, and C. Guillemot. Bit-rate allocation for multi-view video plus depth. In *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*, pages 1–4. IEEE, 2011.

[19] C. Cafforio and F. Rocca. Methods for measuring small displacements of television images. *IEEE Transactions on Information Theory*, IT-22(5):573–579, Sept. 1976.

[20] C. Cafforio and F. Rocca. The differential method for motion estimation. In T. S. Huang, editor, *Image Sequence Processing and Dynamic Scene Analysis*, pages 104–124, 1983.

[21] C. Cafforio, F. Rocca, and S. Tubaro. Motion compensated image interpolation. *IEEE Transactions on Communications*, 38(2):215–222, Feb. 1990.

[22] M. Cagnazzo, M. Agostini, M. Antonini, G. Laroche, and J. Jung. Motion vector quantization for efficient low-bitrate video coding. In *Proceedings of International Symposium on Visual Communication and Image Processing*, San Jose, California, 2009.

[23] M. Cagnazzo, M. Antonini, and M. Barlaud. Mutual information-based context quantization. *Signal Processing: Image Communication (Elsevier Science)*, 25(1):64–74, Jan. 2010.

[24] M. Cagnazzo, F. Castaldo, T. André, M. Antonini, and M. Barlaud. Optimal motion estimation for wavelet video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(7):907–911, July 2007.

[25] M. Cagnazzo, L. Cicala, G. Poggi, and L. Verdoliva. Low-complexity compression of multispectral images based on classified transform coding. *Signal Processing: Image Communication (Elsevier Science)*, 21(3):850–861, Nov. 2006.

[26] M. Cagnazzo, F. Delfino, L. Vollero, and A. Zinicola. Trading off quality and complexity for a low-cost video codec on portable devices. *Elsevier Journal of Visual Communication and Image Representation*, (3), June 2006.

[27] M. Cagnazzo, T. Maugey, and B. Pesquet-Popescu. A differential motion estimation method for image interpolation in distributed video coding. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 1861–1864, Taiwan, 2009.

[28] M. Cagnazzo, W. Miled, T. Maugey, and B. Pesquet-Popescu. Image interpolation with edge-preserving differential motion refinement. In *Proceedings of IEEE Internantional Conference on Image Processing*, Cairo, Egypt, 2009.

[29] M. Cagnazzo, S. Parrilli, G. Poggi, and L. Verdoliva. Costs and advantages of object-based image coding with shape-adaptive wavelet transform. *EURASIP Journal on Image and Video Processing*, 2007:Article ID 78323, 13 pages, 2007. doi:10.1155/2007/78323.

[30] M. Cagnazzo, S. Parrilli, G. Poggi, and L. Verdoliva. Improved class-based coding of multispectral images with shape-adaptive wavelet transform. *IEEE Geoscience and Remote Sensing Letters*, 4(4): 566–570, Oct. 2007.

[31] M. Cagnazzo and B. Pesquet-Popescu. Introducing differential motion estimation into hybrid video coders. In *SPIE Visual Communications and Image Processing Conference*, volume 1, pages 1–4, Huang Shan, An Hui, China, 2010.

[32] M. Cagnazzo and B. Pesquet-Popescu. Perceptual impact of transform coefficients quantization for adaptive lifting schemes. In *International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, AZ, 2010.

[33] M. Cagnazzo and B. Pesquet-Popescu. Depth map coding by dense disparity estimation for mvd compression. In *IEEE Digital Signal Processing*, Corfu, Greece, 2011.

[34] M. Cagnazzo, G. Poggi, and L. Verdoliva. Region-based transform coding of multispectral images. *IEEE Transactions on Image Processing*, 16(12):2916–2926, Dec. 2007.

[35] M. Cagnazzo, G. Poggi, L. Verdoliva, and A. Zinicola. Region-oriented compression of multispectral images by shape-adaptive wavelet transform and SPIHT. In *Proceedings of IEEE Internantional Conference on Image Processing*, volume 4, pages 2459–2462, Singapore, Oct. 2004.

[36] M. Castella, J.-C. Pesquet, and A. Petropulu. A family of frequency- and time-domain contrasts for blind separation of convolutive mixtures of temporally dependent signals. *IEEE Transactions on Signal Processing*, 53(1):107–120, Jan. 2005.

[37] J. Chen. Context modeling based on context quantization with application in wavelet image coding. *IEEE Transactions on Image Processing*, 13(1):26–32, Jan. 2004.

[38] Y. Chen, Y. K. Wang, K. Ugur, M. M. Hannuksela, J. Lainema, and M. Gabbouj. The emerging MVC standard for 3D video services. *EURASIP Journal on Advances in Signal Processing*, 2009, 2009.

[39] G. Cheung, A. Ortega, and N. Cheung. Interactive streaming of stored multiview video using redundant framestructures. *IEEE Transactions on Image Processing*, 20(3):744 –761, Mar. 2011.

[40] S. J. Choi and J. W. Woods. Motion-compensated 3-D subband coding of video. *IEEE Transactions on Image Processing*, 8(2):155–167, Feb. 1999.

[41] P. Chou, Y. Wu, and K. Jain. Practical network coding. In *Proceedings of Allerton Conference on Communication, Control, and Computing*, Monticello, IL, 2003.

[42] R. L. Claypoole, G. M. Davis, W. Sweldens, and R. G. Baraniuk. Nonlinear wavelet transforms for image coding via lifting. *IEEE Transactions on Image Processing*, 12(12):1449–1459, Dec. 2003.

[43] P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 1st edition, 2010.

[44] S. Corrado, M. Agostini, M. Cagnazzo, M. Antonini, G. Laroche, and J. Jung. Improving h.264 performances by quantization of motion vectors. In *Proceedings of Picture Coding Symposium*, Chicago, IL, 2009.

[45] T. Cover and J. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[46] I. Daribo, M. Kaaniche, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu. Dense disparity estimation in multiview video coding. In *Proceedings of IEEE Workshop on Multimedia Signal Processing*, Rio de Janeiro, Brazil, 2009.

[47] I. Daubechies and W. Sweldens. Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.*, 4(3):245–267, 1998.

[48] C. D'Elia, G. Poggi, and G. Scarpa. A tree-structured markov random field model for bayesian image segmentation. *IEEE Transactions on Image Processing*, 12(10), oct 2003.

[49] D. Donoho, M. Vetterli, R. A. DeVore, and I. Daubechies. Data compression and harmonic analysis. *IEEE Transactions on Information Theory*, 44(6):2435–2476, 1998.

[50] F. Dufaux, B. Pesquet-Popescu, and M. Cagnazzo, editors. *Emerging Technologies for 3D Video: Creation, Coding, Transmission and Rendering*. John Wiley & Sons, Ltd, 2013.

[51] O. Egger, P. Fleury, T. Ebrahimi, and M. Kunt. High-performance compression of visual information–a tutorial review-part i: Still pictures. *Proceedings of the IEEE*, 87(6):976–1011, June 1999.

[52] A. El Fawal, J. Le Boudec, and K. Salamatian. Multi-hop broadcast from theory to reality: practical design for ad-hoc networks. In *Proceedings of International Conference on Autonomic Computing and Communication Systems*, 2007.

[53] A. El Gamal and T. Cover. Achievable rates for multiple descriptions. *IEEE Transactions on Information Theory*, 28(6):851–857, Nov. 1982.

[54] J. Fabrizio, S. Dubuisson, and D. Béréziat. Motion compensation based on tangent distance prediction for video compression. *Signal Processing: Image Communication (Elsevier Science)*, 27:153–171, 2012.

[55] J. E. Fowler. Shape adaptive coding using binary set splitting with k-d trees. In *Proceedings of IEEE Internantional Conference on Image Processing*, pages 1301–1304, Singapore, Oct. 2004.

[56] O. N. Gerek and A. E. Çetin. Adaptive polyphase subband decomposition structures for image compression. *IEEE Transactions on Image Processing*, 9(10):1649–1659, Oct. 2000.

[57] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic, Jan. 1992.

[58] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero. Distributed video coding. *Proceedings of the IEEE*, 93(1):71–83, Jan. 2005.

[59] V. Goyal. Multiple description coding: Compression meets the network. *IEEE Signal Processing Magazine*, 18(5):74–93, Sept. 2001.

[60] C. Greco and M. Cagnazzo. A cross-layer protocol for cooperative content delivery over mobile ad-hoc networks. *International Journal of Computer Networks and Distributed Systems*, 7(1–2):49–63, 2011.

[61] C. Greco, M. Cagnazzo, and B. Pesquet-Popescu. H.264-based multiple description coding using motion compensated temporal interpolation. In *Proceedings of IEEE Workshop on Multimedia Signal Processing*, Saint-Malo, France, Oct. 2010.

[62] C. Greco, M. Cagnazzo, and B. Pesquet-Popescu. ABCD : Un protocole cross-layer pour la diffusion vidéo dans des réseaux sans fil ad-hoc. In *Groupe de Recherche sur le Traitement des Signaux et des Images*, Bordeaux, France, Sept. 2011.

[63] C. Greco, M. Cagnazzo, and B. Pesquet-Popescu. Low-latency video streaming with congestion control in mobile ad-hoc networks. *IEEE Transactions on Multimedia*, 14(4):1337 –1350, Aug. 2012.

[64] C. Greco, I. Nemoianu, M. Cagnazzo, and B. Pesquet-Popescu. A network coding scheduling for multiple description video streaming over wireless networks. In *European Signal Processing Conference*, volume 1, pages 1915–1919, Bucharest, Romania, 2012.

[65] C. Greco, G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu. An MDC-based video streaming architecture for mobile networks. In *Proceedings of IEEE Workshop on Multimedia Signal Processing*, Hangzhou, PRC, Oct. 2011.

[66] D. Green, F. Yao, and T. Zhang. A linear algorithm for optimal context clustering with application to bi-level image coding. In *Proceedings of IEEE Internantional Conference on Image Processing*, volume 1, pages 508–511, Oct. 1998.

[67] C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi, R. Leonardi, and J. Ostermann. Distributed monoview and multiview video coding: Basics, problems and recent advances. *IEEE Signal Processing Magazine*, pages 67–76, Sept. 2007.

[68] P. Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2):388–404, Mar. 2000.

[69] H.264/AVC JM reference software. Website. `http://iphome.hhi.de/suehring/tml/`.

[70] H. J. A. M. Heijmans, B. Pesquet-Popescu, and G. Piella. Building nonredundant adaptive wavelets by update lifting. *Applied Computational Harmonic Analysis*, 18(3):252–281, May 2004.

[71] T. Ho, M. Medard, R. Koetter, D. Karger, M. Effros, J. Shi, and B. Leong. A random linear network coding approach to multicast. *IEEE Transactions on Information Theory*, 52(10):4413–4430, Oct. 2006.

[72] Call for proposals on 3D video coding technology. Technical report, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Mar. 2011. Doc. N12036.

[73] ISO/IEC JTC 1. *Coding of audio-visual objects—Part 2: Visual*. ISO/IEC 14496-2 (MPEG-4 Visual), Version 1: Apr. 1999, Version 3: May 2004.

[74] Joint Video Team (JVT). *H.264 JM KTA software coordination, K. Suehring*.

[75] S. Katti, H. Rahul, W. Hu, D. Katabi, M. Médard, and J. Crowcroft. XORs in the air: practical wireless network coding. *ACM SIGCOMM Computer Communication Review*, 36:243–254, Aug. 2006.

[76] J. Katto and Y. Yasuda. Performance evaluation of subband coding and optimization of its filter coefficients. In *Proc. SPIE Visual Communications and Image Processing*, volume 1605, pages 95–106, Nov. 1991.

[77] A. Kawanaka and V. R. Algaz. Zerotree coding of wavelet coefficients for image data on arbitrarily shaped support. In *Proceedings of Data Compression Conference*, page 534, Mar 1999.

[78] S. Li and W. Li. Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 725–743, Aug. 2000.

[79] Z. Liu and L. J. Karam. Mutual information-based analysis of JPEG2000 contexts. *IEEE Transactions on Image Processing*, 14(4):411–421, Apr. 2005.

[80] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[81] M. K. Marina, G. D. Kondylis, and U. C. Kozat. RBRP: A robust broadcast reservation protocol for mobile ad-hoc networks. In *Proceedings of IEEE International Conference on Communications*, 2001.

[82] D. Marpe, H. Schwarz, S. Bosse, B. Bross, P. Helle, T. Hinz, H. Kirchhoffer, H. Lakshman, T. Nguyen, S. Oudin, et al. Video compression using nested quadtree structures, leaf merging, and improved techniques for motion representation and entropy coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 20(12):1676–1687, Dec. 2010.

[83] T. Maugey and P. Frossard. Interactive multiview video system with non-complex navigation at the decoder. *IEEE Transactions on Multimedia*, 15, 2013.

[84] T. Maugey, J. Gauthier, M. Cagnazzo, and B. Pesquet-Popescu. Evaluation of side information effectiveness in distributed video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2014. To appear.

[85] T. Maugey, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu. Fusion schemes for multiview distributed video coding. In *Proceedings of European Signal Processing Conference*, Glasgow, Scotland, 2009.

[86] T. Maugey, W. Miled, M. Cagnazzo, and B. Pesquet-Popescu. Méthodes denses d'interpolation de mouvement pour le codage vidéodistribué monovue et multivue. In *Groupe de Recherche sur le Traitement des Signaux et des Images*, Dijon (France), 2009.

[87] T. Maugey, C. Yaacoub, J. Farah, M. Cagnazzo, and B. Pesquet-Popescu. Side information enhancement using an adaptive hash-based genetic algorithm in a Wyner-Ziv context. Saint-Malo, France, Oct. 2010.

[88] N. Mehrseresht and D. Taubman. Spatially continuous orientation adaptive discrete packet wavelet decomposition for image compression. In *Proceedings of IEEE Internantional Conference on Image Processing*, pages 1593–1596, Atlanta, GA (USA), Oct. 2006.

[89] W. Miled, T. Maugey, M. Cagnazzo, and B. Pesquet-Popescu. Image interpolation with dense disparity estimation in multiview distributed video coding. In *International Conference on Distributed Smart Cameras*, Como, Italy, Sept. 2009.

[90] W. Miled and J. Pesquet. Disparity map estimation using a total variation bound. In *Computer and Robot Vision, 2006. The 3rd Canadian Conference on*, page 48, june 2006.

[91] G. Minami, Z. Xiong, A. Wang, and S. Mehrotra. 3-d wavelet coding of video with arbitrary regions of support. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(9):1063–1068, Sept. 2001.

[92] M. Miyahara, K. Kotani, and V. Algazi. Objective picture quality scale (pqs) for image coding. *IEEE Transactions on Communications*, 46(9):1215–1226, Sept. 1998.

[93] E. G. Mora, C. Greco, B. Pesquet-Popescu, M. Cagnazzo, and J. Farah. Cedar: An optimized network-aware solution for P2P video multicast. In *International Conference on Telecommunications*, Beirut, 2012.

[94] E. G. Mora, J. Jung, M. Cagnazzo, and B. Pesquet-Popescu. Codage de vidéos de profondeur basé sur l'héritage des modes intra de texture. In *Compression et Représentation des Signaux Audiovisuels*, volume 1, pages 1–4, Lille, France, 2012.

[95] E. G. Mora, J. Jung, M. Cagnazzo, and B. Pesquet-Popescu. Depth video coding based on intra mode inheritance from texture. *APSIPA Transactions on Signal and Information Processing*, Jan. 2013. Submitted.

[96] E. G. Mora, J. Jung, M. Cagnazzo, and B. Pesquet-Popescu. Initialization, limitation and predictive coding of the depth and texture quadtree in 3d-hevc video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, Feb. 2013. Submitted.

[97] E. G. Mora, J. Jung, M. Cagnazzo, and B. Pesquet-Popescu. Modification of the merge candidate list for dependent views in 3D-HEVC. In *Proceedings of IEEE Internantional Conference on Image Processing*, Melbourne, Australia, 2013.

[98] E. G. Mora, B. Pesquet-Popescu, M. Cagnazzo, and J. Jung. *Modification of the Merge Candidate List for Dependant Views in 3DV-HTM*. ISO/IEC - ITU, Shangai, PRC, October 2012. Document JCT3V-B0069 for Shanghai meeting (MPEG number m26793).

[99] H. Nagel and W. Enkelmann. An investigation of smoothness constraints for the estimation of displacement vector fields from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(5):565–593, Sept. 1986.

[100] I. Nemoianu, C. Greco, M. Cagnazzo, and B. Pesquet-Popescu. A framework for joint multiple description coding and network coding over wireless ad-hoc networks. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, Mar. 2012.

[101] I. Nemoianu, C. Greco, M. Cagnazzo, and B. Pesquet-Popescu. Multi-view video streaming over wireless networks with rd-optimized scheduling of network coded packets. In *SPIE Visual Communications and Image Processing Conference*, San Diego, CA (USA), 2012.

[102] I. Nemoianu, C. Greco, M. Castella, B. Pesquet-Popescu, and M. Cagnazzo. On a practical approach to source separation over finite fields for network coding applications. In *International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 2013.

[103] The Network Simulator — ns-2. Website. `http://www.isi.edu/nsnam/ns/`.

[104] S. Parrilli, M. Cagnazzo, and B. Pesquet-Popescu. Distortion evaluation in transform domain for adaptive lifting schemes. In *Proceedings of IEEE Workshop on Multimedia Signal Processing*, pages 200–205, Cairns, Australia, 2008.

[105] S. Parrilli, M. Cagnazzo, and B. Pesquet-Popescu. Distortion evaluation in transform domain for adaptive lifting schemes. In *Visual Signal Processing and its Application*, Paris, France, 2008.

[106] S. Parrilli, M. Cagnazzo, and B. Pesquet-Popescu. Estimation of quantization noise for adaptive-prediction lifting schemes. In *Proceedings of IEEE Workshop on Multimedia Signal Processing*, Rio de Janeiro, Brazil, Oct. 2009.

[107] G. Pau, C. Tillier, and B. Pesquet-Popescu. Optimization of the predict operator in lifting-based motion-compensated temporal filtering. In *SPIE Visual Communications and Image Processing*, volume 5308, pages 712–720, San Jose, CA (USA), Jan. 2004.

[108] B. Pesquet-Popescu and V. Bottreau. Three-dimensional lifting schemes for motion compensated video compression. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1793–1796, 2001.

[109] G. Petrazzuoli, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu. Using distributed source coding and depth image based rendering to improve interactive multiview video access. In *Proceedings of IEEE Internantional Conference on Image Processing*, volume 1, pages 605–608, Bruxelles, Belgium, Sept. 2011.

[110] G. Petrazzuoli, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu. Wyner-Ziv coding for depth maps in multiview video-plus-depth. In *IEEE International Conference on Image Processing*, volume 1, pages 1857–1860, Bruxelles, Belgium, 2011.

[111] G. Petrazzuoli, M. Cagnazzo, F. Dufaux, and B. Pesquet-Popescu. Enabling immersive visual communications through distributed video coding. *IEEE MMTC E-Letters*, pages 17–18, May 2013.

[112] G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu. Fast and efficient side information generation in distributed video coding by using dense motion representation. In *Proceedings of European Signal Processing Conference*, Aalborg, Denmark, 2010.

[113] G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu. High order motion interpolation for side information improvement in dvc. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2342–2345, Dallas, TX, 2010.

[114] G. Petrazzuoli, M. Cagnazzo, and B. Pesquet-Popescu. Novel solutions for side information generation and fusion in multiview distributed video coding. 2013. In preparation.

[115] G. Petrazzuoli, T. Maugey, M. Cagnazzo, and B. Pesquet-Popescu. Side information refinement for long duration GOPs in DVC. In *Proceedings of IEEE Workshop on Multimedia Signal Processing*, volume 1, Saint-Malo, France, 2010.

[116] G. Petrazzuoli, T. Maugey, M. Cagnazzo, and B. Pesquet-Popescu. A novel dvc framework for multiple video plus depth coding. *Journal of Advances in Signal Processing*, 2013.

[117] G. Piella, B. Pesquet-Popescu, and H. J. A. M. Heijmans. Gradient-driven update lifting for adaptive wavelets. *Signal Processing: Image Communication (Elsevier Science)*, 20(9-10):813–831, Oct.-Nov. 2005.

[118] V. Ratnakar. RAPP, lossless image compression with runs of adaptive pixel patterns. In *32nd Asilomar Conf. Signals, System and Computers*, pages 1251–1255, Pacific Grove, California, Nov. 1998.

[119] E. Royer, P. Melliar-Smith, and L. Moser. An analysis of the optimum node density for ad-hoc mobile networks. In *Proceedings of IEEE International Conference on Communications*, 2001.

[120] A. Said and W. Pearlman. A new, fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(3):243–250, June 1996.

[121] E. Setton, T. Yoo, X. Zhu, A. Goldsmith, and B. Girod. Cross-layer design of ad-hoc networks for real-time video streaming. *IEEE Transactions on Wireless Communications*, 12(4):59–65, 2005.

[122] F. Shao, G. Jiang, M. Yu, K. Chen, and Y.-S. Ho. Asymmetric coding of multi-view video plus depth based 3-d video for view rendering. *IEEE Transactions on Multimedia*, 14(1):157–167, 2012.

[123] A. Skodras and T. Ebrahimi. The JPEG2000 still image compression standard. *IEEE Signal Processing Magazine*, 18(5):36–58, Sept. 2001.

[124] D. Slepian and J. K. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19:471–480, July 1973.

[125] D. Taubman. High performance scalable image compression with EBCOT. *IEEE Transactions on Image Processing*, 9(7):1158–1170, July 2000.

[126] N. Thomos, J. Chakareski, and P. Frossard. Prioritized distributed video delivery with randomized network coding. *IEEE Transactions on Multimedia*, 13(4):776–787, 2011.

[127] J. Tourrilhes. Robust broadcast: Improving the reliability of broadcast transmissions on CSMA/CA. In *IEEE International Symposium on Personal, Indoor and Mobile Radio Communications*, 1998.

[128] B. Usevitch. Optimal bit allocation for biorthogonal wavelet coding. In *Proceedings of Data Compression Conference*, pages 387–395, Snowbird, USA, Mar. 1996.

[129] B. E. Usevitch. A tutorial on modern lossy wavelet image compression: foundations of JPEG2000. *IEEE Signal Processing Magazine*, 18(5):22–35, Sept. 2001.

[130] G. Valenzise, G. Cheung, R. Galvao, M. Cagnazzo, B. Pesquet-Popescu, and A. Ortega. Motion prediction of depth video for depth-image-based rendering using don't care regions. In *Proceedings of Picture Coding Symposium*, pages 93–96, Krakow, Poland, May 2012.

[131] A. Vetro, T. Wiegand, and G. Sullivan. Overview of the stereo and multiview video coding extensions of the H.264/MPEG-4 AVC standard. *Proceedings of the IEEE*, 99(4):626–642, Apr. 2011. Invited Paper.

[132] D. Vukobratović and and V. Stanković. Unequal error protection random linear coding for multimedia communications. In *Proceedings of IEEE Workshop on Multimedia Signal Processing*, pages 280–285, Saint-Malo, France, Oct. 2010.

[133] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[134] X. Wu, P. A. Chou, and X. Xue. Minimum conditional entropy context quantization. In *Proceedings of IEEE International Symposium on Information Theory*, pages 43–53, Sorrento, Italy, June 2000.

[135] X. Wu, P. A. Chou, and X. Xue. Minimum conditional entropy context quantization, 2000.

[136] X. Wu and N. D. Memon. Context-based, adaptive, lossless image coding. *IEEE Transactions on Communications*, 45:437–444, Apr. 1997.

[137] A. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the receiver. *IEEE Transactions on Information Theory*, 22:1–11, Jan. 1976.

[138] C. Zhu and M. Liu. Multiple description video coding based on hierarchical B pictures. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(4):511–521, Apr. 2009.

Selected Publications

1. **Marco Cagnazzo**, Filippo Castaldo, Thomas André, Marc Antonini, and Michel Barlaud. Optimal Motion Estimation for Wavelet Motion Compensated Video Coding. IEEE Transactions on Circuits and Systems for Video Technology, vol. 17, no. 7, pp. 907-911, July 2007.

2. **Marco Cagnazzo**, Giovanni Poggi, and Luisa Verdoliva. Region-Based Transform Coding of Multispectral Images. IEEE Transactions on Image Processing, Dec. 2007

3. **Marco Cagnazzo**, Marc Antonini, Michel Barlaud. Mutual information-based context quantization. Signal Processing: Image Communication, vol. 25, no. 1, pp. 64-74, Jan. 2010.

4. Marie Andrée Agostini-Vautard, **Marco Cagnazzo**, Marc Antonini, Guillaume Laroche, and Joël Jung. A New Coding Mode for Hybrid Video Coders Based on Quantized Motion Vectors. IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 7, pp. 946-956, July 2011.

5. Claudio Greco, **Marco Cagnazzo**, Béatrice Pesquet-Popescu. Low-Latency Video Streaming with Congestion Control in Mobile Ad-Hoc Networks. In IEEE Transactions on Multimedia, vol. 14, no. 4, pp. 1337-1350, Aug. 2012.

6. Abdalbassir Abou-El Ailah, Frédéric Dufaux, Joumana Farah, **Marco Cagnazzo**, and Béatrice Pesquet-Popescu. Fusion of Global and Local Motion Estimation for Distributed Video Coding. IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 1, pp. 158-172, Jan. 2013.