Distributed Compressed Representation of Correlated Image Sets

THÈSE N^{O} 5264 (2012)

PRÉSENTÉE LE 27 JANVIER 2012

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE TRAITEMENT DES SIGNAUX 4
PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Vijayaraghavan THIRUMALAI

acceptée sur proposition du jury:

Prof. J.-Ph. Thiran, président du jury
Prof. P. Frossard, directeur de thèse
Prof. J. E. Fowler, rapporteur
Dr C. Guillemot, rapporteur
Prof. P. Vandergheynst, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2012

What you have learned is a mere handful; What you haven't learned is the size of the world.

- Avvaiyar

Acknowledgements

I take the pleasure of thanking everyone who made this thesis possible. Foremost, I owe my deepest gratitude to Prof. Pascal Frossard who has guided me throughout this thesis with his valuable suggestions and advices, most importantly for giving me the flexibility to work on my own research directions. I had an opportunity to learn from his experience and vision while working closely with him. I could not have imagined a better advisor and mentor for my PhD. I thank him for offering me the great opportunity to work in his laboratory. I also greatly appreciate his dedication in reading and correcting this manuscript.

I am thankful to all my thesis committee members Prof. J. E. Fowler, Dr C. Guillemot and Prof. P. Vandergheynst, as well as the President of the committee Prof. J.-Ph. Thiran for the time spent in carefully reading this manuscript and for their valuable suggestions in improving the final version of the thesis. My special thanks to Dr C. Guillemot and Prof. J. E. Fowler for their valuable time spent on traveling for my sake. The work presented here has been funded by the Swiss National Science Foundation under grant 200021-118230. I am grateful for their support.

I thank all my present and former colleagues in the LTS4 laboratory-*best lab in the world*: Ivana R, Dan, Ivana T, Jean-Paul, Efi, Zorana, Tamara, Zafer, Thomas, Luigi, Ana, David, Eirina, Elif, Xiaowen, Sofia, Dorina, Nikos, Jacob, Yannick, Hyunggon and Jari for creating such a friendly working environment and for the many great moments spent together. I thank my office mate Tamara for the exciting and fruitful discussions we had together. Thanks Tamara for tolerating me all these years! My sincere thanks to Xiaowen, Tamara, Elif, Dorina, Eirina, Sofia and Thomas for proof reading this thesis and giving their valuable comments. My special thanks to Thomas for helping me in translating the abstract to French. I also thank the LTS staffs, in particular Marianne Marion and Rosie De Pietro for making the administrative work less painful.

I thank all my Indian friends in Switzerland: Aravind, Deepu, Sai, Venkat, Subbu, Karthik, Raja, Bharghav and Bhavani for the great times we spent together including mountain hiking, week-end parties and trips. They made Switzerland a fun place to live. I thank all my close friends both in India and abroad for their support, enthusiasm, and affection. I have enjoyed the company and support of many others whom I fail to mention here, and would like to thank them all for their help and friendship.

I thank my family members, my father Thirumalai, my mother Padmavathy and my brother Varadharajan for their love and affection towards me in all my life. My special thanks to my father-in-law Srinivasan, my mother-in-law Sujatha, and my sister-in-laws Divya and Hasitha for their unconditional and never-ending support. Their support and love have made me what I am and helped me through the most difficult periods of my life.

Huge thanks to my wife, Ramya, for her patience, endless love and support towards me during the hardest year of my PhD studies. Staying with PhD student is not an easy task. She has raised my confidence whenever necessary and she has always believed in me more than myself.

Abstract

Vision sensor networks and video cameras find widespread usage in several applications that rely on effective representation of scenes or analysis of 3D information. These systems usually acquire multiple images of the same 3D scene from different viewpoints or at different time instants. Therefore, these images are generally correlated through displacement of scene objects. Efficient compression techniques have to exploit this correlation in order to efficiently communicate the 3D scene information. Instead of joint encoding that requires communication between the cameras, in this thesis we concentrate on distributed representation, where the captured images are encoded independently, but decoded jointly to exploit the correlation between images. One of the most important and challenging tasks relies in estimation of the underlying correlation from the compressed correlated images for effective reconstruction or analysis in the joint decoder.

This thesis focuses on developing efficient correlation estimation algorithms and joint representation of multiple correlated images captured by various sensing methodologies, e.g., planar, omnidirectional and compressive sensing (CS) sensors. The geometry of the 2D visual representation and the acquisition complexity vary for each sensor type. Therefore, we need to carefully consider the specific geometric nature of the captured images while developing distributed representation algorithms. In this thesis we propose robust algorithms in different scene analysis and reconstruction scenarios.

We first concentrate on the distributed representation of omnidirectional images captured by catadioptric sensors. The omnidirectional images are captured from different viewpoints and encoded independently with a balanced rate distribution among the different cameras. They are mapped on the sphere which captures the plenoptic function in its radial form without Euclidean discrepancies. We propose a transform-based distributed coding algorithm, where the spherical images initially undergo a multi-resolution decomposition. The visual information is then split into two correlated partitions. The encoder transmits one partition after entropy coding, as well as the syndrome bits resulting from the Slepian-Wolf encoding of the other partition. The joint decoder estimates a disparity image to take benefit of the correlation between views and uses the syndrome bits to decode the missing information. Such a strategy proves to be beneficial with respect to the independent processing of images and shows only a small performance loss compared to the joint encoding of different views.

The encoding complexity in the previous approach is non-negligible due to the visual information processing based on Slepian-Wolf coding and its associated rate parameter estimation. We therefore discard the Slepian-Wolf encoding and propose a distributed coding solution, where the correlated images are encoded independently using transform-based coding solutions (e.g., SPIHT). The central decoder now builds a correlation model from the compressed images, which is used to jointly decode a pair of images. Experimental results demonstrate that the proposed distributed coding solution improves the rate-distortion performance of the separate coding results for both planar and omnidirectional images. However, this improvement is significant only at medium to high bit rates. We therefore propose a rate allocation scheme that identifies and transmits the necessary visual information from each image to improve the correlation estimation accuracy at low bit rate. Experimental results show that for a given bit budget the proposed encoding scheme permits to compute an accurate correlation estimation comparing to the one obtained with SPIHT, JPEG 2000 or JPEG coding schemes. We show however that the improvement in the correlation estimation comes at the price of penalizing the image reconstruction quality; therefore there exists an interesting trade-off between the accurate correlation estimation and image reconstruction as encoding optimization objectives are different in both cases.

Next, we further simplify the encoding complexity by replacing the classical imaging sensors with the simple CS sensors, that directly acquire the compressed images in the form of quantized linear measurements. We now concentrate on the particular problem, where one image is selected as the reference and it is used as a side information for the correlation estimation. We propose a geometry-based model to describe the correlation between the visual information in a pair of images. The joint decoder first captures the most prominent visual features in the reconstructed reference image using geometric functions. Since the images are correlated, these features are likely to be present in the other images too, possibly with geometric transformations. Hence, we propose to estimate the correlation model with a regularized optimization problem that locates these features in the compressed images. The regularization terms enforce smoothness of the transformation field, and consistency between the estimated images and the quantized measurements. Experimental results show that the proposed scheme is able to efficiently estimate the correlation between images for several multi-view and video datasets. The proposed scheme is finally shown to outperform DSC schemes based on unsupervised disparity (or motion) learning, as well as independent coding solutions based on JPEG 2000.

We then extend the previous scenario to a symmetric decoding problem, where we are interested to estimate the correlation model directly from the quantized linear measurements without explicitly reconstructing the reference images. We first show that the motion field that represents the main source of correlation between images can be described as a linear operator. We further derive a linear relationship between the correlated measurements in the compressed domain. We then derive a regularized cost function to estimate the correlation model directly in the compressed domain using graph-based optimization algorithms. Experimental results show that the proposed scheme estimates an accurate correlation model among images in both multi-view and video imaging scenarios. We then propose a robust data fidelity term that improves the quality of the correlation estimation when the measurements are quantized. Finally, we show by experiments that the proposed compressed correlation estimation scheme is able to compete the solution of a scheme that estimates a correlation model from the reconstructed images without the complexity of image reconstruction.

Finally, we study the benefit of using the correlation information while jointly reconstructing the images from the compressed linear measurements. We consider both the asymmetric and symmetric scenarios described previously. We propose joint reconstruction methodologies based on a constrained optimization problem which is solved using effective proximal splitting methods. The constraints included in our framework enforce the reconstructed images to satisfy both the correlation and the quantized measurements consistency objectives. Experimental results demonstrate that the proposed joint reconstruction scheme improves the quality of the decoded images, when compared to a scheme where the images are handled independently.

In this thesis we build efficient distributed scene representation algorithms for the multiple correlated images captured in planar, omnidirectional and CS cameras. The coding rate in our symmetric distributed coding solution stays balanced between the encoders and stays close to the joint encoding solutions. Our novel algorithms lead to effective correlation estimation in different sensing and coding scenarios. In addition, we provide innovative solutions for robust correlation estimation from highly compressed images in simple sensing frameworks. Our CS-based joint reconstruction frameworks effectively exploit the inter-view correlation, that permits to achieve high compression gains compared to state-of-the-art independent and distributed coding solutions.

Keywords: Distributed scene representation, multi-view images, video images, correlation estimation, random projections, joint reconstruction, quantization.

Résumé

De nombreuses applications actuelles s'appuient sur des systèmes comme les réseaux de capteurs et les caméras vidéos, et visent à élaborer des représentations efficaces des scènes ou l'analyse de scènes 3D. Ces systèmes effectuent généralement l'acquisition de plusieurs images d'une même scène 3D depuis différents points de vue ou à différents instants. Ainsi, les images capturées se retrouvent corrélées par les déplacements des objets de la scène. Les images corrélées ne peuvent cependant pas être transmises directement, cela serait trop coûteux en terme de taille et de ressource de communication. Il est ainsi nécessaire de faire appel à des techniques efficaces de compression afin de transmettre l'information de la scène 3D en profitant de la corrélation. Dans cette thèse, au lieu d'envisager un encodage conjoint qui nécessiterait une communication entre les caméras, nous nous focalisons sur une représentation distribuée, où les images capturées sont encodées indépendamment, mais décodées conjointement afin d'exploiter leur corrélation. Un des défis les plus importants est celui qui consiste à estimer le modèle de corrélation sous-jacent à partir des images compressées, afin que de garantir un traitement ou un codage efficace de la scène 3D.

Les travaux de cette thèse s'attèlent à développer des algorithmes efficaces d'estimation de la corrélation et des reconstructions jointes des différentes images capturées par des technologies de capteurs variées, p.ex., plans, omnidirectionnels, basés compressive sensing (CS). La géométrie de la représentation visuelle 2D ainsi que la complexité d'acquisition varient en fonction du type de capteur. Ainsi, nous devons considérer de manière attentive la nature géométrique spécifique des images capturées pour développer des algorithmes de représentation distribuée. Nous proposons dans cette thèse des algorithms robustes pour l'analyse et la représentation distribuée d'images multiples dans différents scénarios.

Nous nous concentrons d'abord sur une représentation distribuée des images omnidirectionnelles dont l'acquisition est effectuée par des capteurs catadioptriques. En particulier, nous supposons que les images capturées à partir de différents points de vues sont encodées indépendamment avec un partage du débit équilibré entre les différentes caméras. Les images omnidirectionnelles sont représentées sur la sphère, où la fonction plénoptique est décrite en coordonnées radiales plutôt qu'euclidiennes. Nous proposons un algorithme de codage distribué, où les images sphériques subissent d'abord une décomposition multi-résolution. L'information visuelle est ainsi divisée en deux partitions corrélées. L'encodeur transmet une partition après un codage entropique, ainsi que des bits de syndromes résultant du codage Slepian-Wolf de l'autre partition. Le décodeur conjoint utilise une estimation de la disparité afin d'exploiter la corrélation entre les vues, et utilise également les bits de syndromes pour décoder l'information manquante. Une telle stratégie se montre plus performante que celle consistant à traiter les images indépendamment, et n'occasionne qu'une faible perte de performance par rapport à un encodage conjoint des différentes vues.

La complexité d'encodage de l'approche précédente est non négligeable à cause du traitement de l'information visuelle par le codage Slepian-Wolf et du paramétrage des débits associés. Nous abandonnons ainsi l'encodage Slepian-Wolf et proposons une solution de codage distribué, dans laquelle les images corrélées sont compressées indépendamment en utilisant des solutions de codage classiques à base de transformées (e.g., SPIHT). Le décodeur central construit désormais lui-même un modèle de corrélation pour les images compressées, qui est utilisé pour décoder conjointement les deux images. Les résultats expérimentaux ont démontré que la solution proposée de codage distribué améliore les performances débit-distorsion par rapport aux résultats du décodage disjoint pour le cas des images planes et omnidirectionnelles. Cependant, cette amélioration est significative seulement à moyen et haut débit. Nous proposons ainsi un schéma d'allocation de débit qui identifie et transmet l'information visuelle nécessaire pour chaque image afin d'améliorer la

précision de l'estimation de la corrélation à bas débit. Les résultats expérimentaux montrent que, pour un débit donné, le schéma d'encodage proposé permet d'effectuer une estimation de la corrélation plus précise que celle obtenue avec des schémas de codage classique comme SPIHT, JPEG 2000, et JPEG. Nous montrons cependant que l'amélioration de l'estimation de la corrélation s'effectue au détriment de la qualité de reconstruction des images, et ainsi qu'il existe un compromis intéressant entre la précision de l'estimation de la corrélation et la qualité de la reconstruction des images.

Par la suite, nous simplifions encore plus la complexité d'encodage en remplaçant les capteurs d'image classiques par de simples capteurs CS, qui capturent directement les images compressées par des mesures linéaires quantifiées. Nous nous concentrons maintenant sur le problème où une image est désignée comme référence et est utilisée comme une information adjacente pour l'estimation de la corrélation. Nous proposons un modèle fondé sur la géométrie afin de décrire la corrélation entre l'information visuelle de deux images. Le décodage conjoint capture dans un premier temps les caractéristiques visuelles les plus proéminentes dans l'image de référence reconstruite en utilisant des fonctions géométriques. Comme les images sont corrélées, ces caractéristiques ont de grandes chances d'être présentes dans les autres images aussi, avec de possibles transformations géométriques. Ainsi, nous proposons d'estimer le modèle de corrélation par le biais d'un problème d'optimisation régularisé qui localise ces caractéristiques dans l'image compressée. Les termes de régularisation renforcent l'aspect lisse du champ de transformation, et la consistance entre l'image prédite et les mesures quantifiées. Les résultats expérimentaux montrent que le schéma proposé est capable d'estimer efficacement la corrélation entre les images de plusieurs bases de données vidéo et multi-vues. Le schéma proposé améliore de plus les résultats des schémas de codage de source distribué fondés sur un apprentissage non supervisé de la disparité (ou du mouvement), ainsi que ceux des solutions de codage indépendant basées sur JPEG 2000.

Nous étendons ensuite le scénario précédent à un problème de décodage symétrique, où nous nous attelons à estimer le modèle de corrélation directement depuis les mesures linéaires quantifiées sans reconstruire explicitement les images de référence. Nous montrons d'abord que le champ de mouvement qui représente la source principale de corrélation entre les images peut être efficacement décrit comme un opérateur linéaire. Nous en tirons une relation linéaire entre les mesures corrélées dans le domaine compressé. Nous obtenons alors une fonction de coût régularisée qui permet d'estimer le modèle de corrélation directement dans le domaine compressé en utilisant des algorithmes d'optimisation par graphes. Les résultats expérimentaux montrent que le schéma proposé estime un modèle de corrélation précis entre les images dans les scénarios multi-vue et vidéo. Nous proposons aussi un terme de fidélité qui améliore la qualité de l'estimation de la corrélation quand les mesures sont quantifiées. Enfin, nous montrons par les expériences que le schéma proposé pour l'estimation de la corrélation est compétitif par rapport à la solution d'un schéma qui estimerait un modèle de corrélation à partir des images reconstruites, mais qui nécessiterait une étape coûteuse de reconstruction des différentes images.

Enfin, nous étudions l'intérêt d'utiliser l'information de corrélation, durant la reconstruction conjointe des images à partir des mesures linéaires compressées. Nous considérons les scénarios asymétriques et symétriques décrits précédemment. Nous proposons des méthodologies de reconstruction conjointe fondées sur un problème d'optimisation avec contrainte, qui est résolu par des méthodes efficaces de séparation proximale. Les contraintes inclues dans notre système forcent les images reconstruites à satisfaire à la fois des objectifs de corrélation et de consistance entre les mesures quantifiées. Les résultats expérimentaux démontrent que le schéma de reconstruction conjointe améliore la qualité des images décodées, lorsqu'on les compare à un schéma où les images sont traitées indépendamment.

En résumé, dans cette thèse, nous construisons des algorithmes efficaces de représentation distribuée de scène pour des nombreuses images corrélées capturées avec des caméras planes, omnidirectionnelles ou CS. Les débits de codage dans notre solution symétrique de codage distribué restent équilibrés entre les différents encodeurs et sont proches de ceux obtenus par des solutions d'encodage conjoint. Notre nouveau schéma d'allocation de débit a prouvé qu'il parvenait à calculer une information de corrélation précise à partir de vues fortement compressées. Nos nouveaux algorithmes offrent la possibilité d'estimer efficacement la corrélation entre des images dans différents scenarios à vues multiples ou video. En plus, nous proposons

des solutions novatrices pour une analyse d'image efficace à partie d'images fortement compressées dans des schémas de capture trés simples. Notre système de reconstruction conjointe basée CS exploite efficacement les corrélations inter-vues, ce qui permet d'atteindre d'importants gains de compression par rapport aux solutions de codage indépendant et de codage distribué de l'état de l'art.

Mots clefs : représentation distribuée de scène, image multi-vues, image vidéo, estimation de corrélation, reconstruction conjointe, quantification.

Contents

Chapter 1

Introduction

1.1 Motivation

Camera vision sensor is a device that captures the visual information from the real world physical environments. Such sensors usually collect the intensity and color light field information of objects in the scene. These vision sensors however provide only a 2D view of the scene information, and hence the visual information acquired by a single camera cannot provide the complete knowledge of the scene, as well as the 3D position of the objects. However, in numerous image analysis applications, it is beneficial to have the 3D structural information of the scene. In this context, more details of the scene can be provided by using a single moving camera or multiple cameras which sample the same 3D scene at different time instants or different viewpoints respectively. The multiple images captured by these cameras carry enough information for an accurate representation of the 3D information. The 3D information is usually extracted by processing the correlated multiple images based on the fundamental principles of structure from motion or multi-view geometry [1, 2, 3].

The multiple images acquired by vision sensor networks or by moving cameras are highly correlated. In particular, there exists a geometrical relationship among images as the same scene objects are observed in different viewpoints or at different time instants. Due to the enormous volume of visual information these images cannot be transmitted directly, as it could quickly exceed the available bandwidth resources. Therefore, it is necessary to build an efficient compression scheme that exploits the inter-view redundancies in order to reduce the bandwidth and communication resources. In the literature, efficient coding strategies for compressing the multi-view images have been proposed, e.g., [4, 5]. These conventional methods are based on joint encoding, that processes all the visual information at a central encoder in order to exploit the correlation between images. Unfortunately, such an approach demands bulk computing resources at the encoder that further translates into significant power consumption. Also the inter-sensor communication is undesirable due to the limited availability of communication and bandwidth resources or due to network topology for example. Due to the high power consumption and encoding complexity, the joint encoding approaches are not attractive for compressing the correlated images captured in a low power vision sensor network and a video camera.

The distributed coding paradigm becomes particularly attractive in such settings as it involves a low complexity encoding stage that further permits to get rid of inter-sensor communication. In this case, the images captured by one or several image sensors are encoded independently but decoded jointly by a central decoder that exploits the underlying correlation between views, as illustrated in Fig. 1.1. The computational complexity of the visual information in this representation is thus shifted from the encoder to the joint decoder. The joint decoder eventually reconstructs the visual information from the compressed images by exploiting the correlation between the samples, which permits to achieve a good rate-distortion

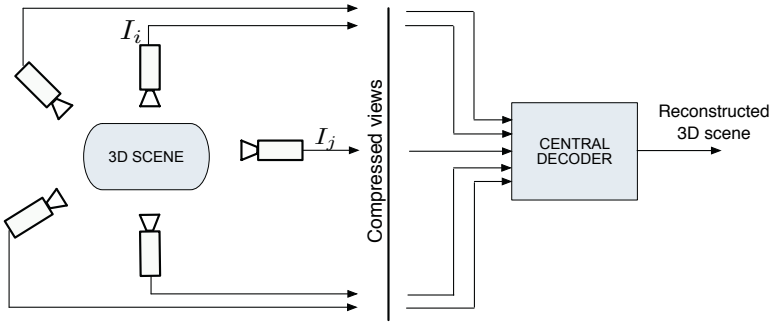


Figure 1.1: Distributed compression of multiple images acquired in a vision sensor network. The images are encoded independently by the respective cameras, but decoded jointly at a central decoder.

trade-off in the representation of video or multi-view information.

In general, one can use different vision sensors to acquire the visual information from the 3D scene. For example, it could be a perspective planar sensor, an omnidirectional sensor and a compressive sensor (CS). The main difference among these vision sensors are the complexity and the geometrical representation of the acquired visual information. Now, the interesting and challenging task is to distributively compress the multiple images captured by these different cameras by properly considering the specific geometry of the imaging system. The scope of this thesis is to design such distributed coding architectures in order to efficiently communicate the 3D scene information provided by the different imaging cameras.

1.2 Challenges

The theory behind the distributed compression was postulated by Slepian and Wolf in 1970's [6]. Though the theory exists for more than four decades, the application of distributed compression for the multi-view and time varying images has not been fully matured yet, due to the lack of accurate correlation modeling and estimation. The rate-distortion (RD) performance in the distributed coders is highly sensitive to the incorrect correlation modeling and estimation. Therefore, it is desirable to estimate an accurate correlation information at the joint decoder, so that the visual information can be efficiently processed, coded or rendered.

In distributed compression framework, the captured images are compressed independently due to the limited availability of bandwidth and communication resources. The joint decoder therefore has access to the compressed images and not the original images, as shown in Fig. 1.1. Now, the decoder is in a situation where it has to estimate the underlying 3D scene geometry from the multiple compressed views, in order to perform the joint reconstruction of multiple images. In the literature, plethora of tools are available to estimate the underlying geometrical relationship between images [7, 8]. However, these state-of-the-art techniques cannot efficiently handle the compressed images and often fail to capture the actual correlation from compressed views, which results in rate-distortion penalization in distributed scenarios.

In this thesis, we need to consider the different geometrical and representation challenges related to the processing of visual information collected by different sensors that include planar, omnidirectional and compressive sensing (CS) sensors. It should be noted that in order to provide an accurate representation of a 3D scene, we need to efficiently account the geometry of the particular imaging system while processing the visual information. This thesis hence addresses the following issues.

1. How to properly consider the compression artifacts and to build a robust correlation model that estimates an accurate scene geometry from the compressed images ?

2. How to properly take advantage of the appropriate geometrical structure and the visual representation of a particular sensor in order to build an efficient distributed scene representation ?

In this thesis, we propose a novel distributed coding solution that processes the omnidirectional images on the sphere in order to take advantage of the geometry of the imaging system. Also, we propose a novel rate allocation methodology to compress the correlated images (e.g., planar and omnidirectional) that facilities to estimate an accurate correlation information at the decoder from highly compressed images. Finally, our novel CS-based distributed joint representation framework estimates the underlying 3D geometry directly in the compressed domain by enforcing quantization consistency constraint; this provides robust correlation estimation from the compressed measurements.

1.3 Thesis organization

This thesis focuses on developing distributed scene representation algorithms in different multi-view sensing scenarios. We perform correlation estimation and joint reconstruction for different types of sensor networks and video cameras. This thesis is organized as follows.

In Chapter 2, we provide the background theory on the projective geometry and processing of the visual information acquired using perspective, omnidirectional and compressive sensing cameras. We also highlight the geometrical properties of the visual information captured by these sensors. We then describe the geometrical relationship between multiple images captured in multi-view and video imaging scenarios. Finally, we review the state-of-the-art distributed image representation and compression techniques for encoding the correlated images acquired in both video and multi-view imaging frameworks. At last, we summarize the drawbacks in the existing distributed coding frameworks.

In Chapter 3, we propose a distributed scene representation algorithm where a central decoder reconstructs the 3D scene information from the distributively encoded two omnidirectional images. In particular, we concentrate on the problem where the bandwidths between the sensors and the joint decoder are constrained with balanced rate allocation. The distributed coding is built on a multi-resolution representation and partitioning of the visual information in each camera. The encoder then transmits one partition after entropy coding, as well as the syndrome bits resulting from the channel encoding of the other partition. The decoder exploits the intra-view correlation and attempts to reconstruct the source images by combining the entropy-coded partition and the syndrome information. At the same time, it exploits the inter-view correlation using disparity estimation between the views from different cameras. Experiments demonstrate that the proposed distributed coding solution performs better than a scheme where the images are handled independently.

In Chapter 4, we seek a distributed coding paradigm, where the acquired planar or omnidirectional images are progressively encoded (e.g., SPIHT) and are further transmitted to the joint decoder without Slepian-Wolf coding which might necessitate complex channel estimation. We propose to first estimate the underlying correlation model from the indepedently compressed images. Then, the estimated correlation information is used for the joint signal recovery. The joint reconstruction is cast as a convex optimization problem that enhances the quality of the reconstructed image pairs that have been compressed independently. We show by experiments that the joint reconstruction enhances the quality of the indepedently compressed images only at medium bit rates and not at low bit rates; this is due to the poor correlation information estimated from the highly compressed images. In the sequel, we propose a *smart encoder* that identifies and encodes the visual information such that it leads to a better correlation estimation for a given target bit rate. We finally show by experiments that there exists an actual trade-off between the quality of image reconstruction and the correlation estimation due to the encoding of different image characteristics.

We then replace the traditional imaging cameras in the previous distributed scheme with the low cost CS-based imaging cameras. The CS-based cameras directly acquire the compressed image in the form of quantized linear measurements using simple inner product operations, and therefore they permit low complexity encoding. In Chapter 5, we focus on the problem of estimating the correlation information from

one reference image and highly compressed images given in the form of random projections, that are further quantized and entropy coded. We propose a geometry-based model to describe the correlation between images, which is mostly driven by the translational motion of objects or vision sensors. We first estimate prominent visual features by computing a sparse approximation of a reference image with a dictionary of geometric basis functions. We then pose a regularized optimization problem to estimate the corresponding features in correlated images given by quantized linear measurements. Eventually, a dense depth or motion field is generated from the local transforms of the geometric features. Experimental results show that the proposed algorithm effectively estimates the depth and optical flow between images, respectively in multi-view and video datasets. In addition, when the depth or motion field is used for image prediction, the resulting rate-distortion performance becomes better than the independent coding solutions, as well as state-of-the-art distributed coding schemes based on disparity learning.

In Chapter 6, we concentrate on the correlation estimation between images, where all the images are represented in the form of quantized linear measurements with simple CS sensors. This scenario therefore provides a low complexity encoding compared to the previous scenarios and the entire computational complexity is shifted to the joint decoder. In these settings, we propose to estimate a dense depth or optical flow model directly in the compressed domain by jointly processing the quantized linear measurements without explicit image reconstruction. We first show that the correlated images can be efficiently related using a linear operator. Using this linear relationship, we then describe the dependencies between images in the compressed domain. We further cast a regularized optimization problem where the correlation is estimated in order to satisfy both data consistency and motion smoothness objectives. Extensive experiments in multi-view and video imaging applications show that our novel solution stays competitive with methods that implement complex image reconstruction steps prior to correlation estimation. In addition to this, we experimentally show that the accuracy of the correlation estimation can be improved by properly handling the measurement quantization noise.

In Chapter 7, we propose joint image reconstruction algorithms for decoding images based on the correlation estimation results in Chapters 5 and 6. The joint reconstruction is cast as a convex optimization problem where the reconstructed images have to satisfy either total variation (TV) or sparsity priors, as well as consistency with both the quantized measurements and the correlation estimates. We solve this joint reconstruction problem by effective proximal methods. We show experimentally that the accurate correlation estimation in distributed image representation permits to outperform independent decoding solutions in terms of reconstructed image quality.

In Appendix A.1, we present our framework that estimates a dense transformation field from the sets of geometrically transformed atoms, i.e., the local transforms between geometric atoms are fused to estimate the correlation between pixels in the images. We propose to estimate the underlying dense transformation field based on a regularized optimization framework that is solved using Graph Cuts. The data fidelity term finds a solution that best fits with the local transforms between the corresponding atoms. We also penalize the transformations among the neighboring pixels in order to ensure consistency. We show experimentally that the proposed solution estimates an accurate transformation field when compared to the one that finds a dense field based on most confident mapping.

Finally, in Chapter 8, we draw conclusions and the future directions of research related to this thesis.

1.4 Thesis contributions

The contributions in this thesis are listed as follows,

- We propose a novel rate balanced distributed coding scheme for compressing the correlated spherical images captured in omnidirectional sensor networks.
- We propose a novel joint reconstruction algorithm that enhances the quality of the distributively compressed images by effectively exploiting the inter-view correlation. We show that our reconstruction

algorithm is convex, which is further effectively solved using proximal methods.

- We propose a novel rate allocation scheme that permits to estimate an accurate disparity information from the highly compressed planar or omnidirectional images.
- We propose a novel algorithm to compute the geometrical relationship between correlated images from a reference image and highly compressed linear measurements. We further propose a methodology to efficiently handle the measurement quantization noise.
- We propose a new optimization algorithm based on Graph Cuts that estimates a dense correlation model between images from sets of geometrically transformed atoms.
- We propose a new algorithm to estimate the correlation model directly from the set of compressed images that are given in the form of quantized linear measurements. Furthermore, we propose a robust correlation estimation scheme that properly handles the non-linearities of measurement quantization.
- We propose novel joint reconstruction algorithms based on a convex optimization framework for decoding correlated images from the quantized linear measurements. Our joint reconstruction solution is applied to both asymmetric and symmetric distributed coding frameworks.

Chapter 2

Related Work

This chapter summarizes state-of-the-art works from the literature that are closely related to the problems addressed in this thesis. First, in Section 2.1 we give a brief overview of the general imaging sensors that include perspective, omnidirectional, and compressed sensing (CS) cameras. Next, in Section 2.2 we present the image representation and coding tools to encode the visual information captured from various imaging sensors. We then move from the single imaging to multi-imaging frameworks, where the multiple images are captured by a video camera or a vision sensor network. In the sequel, in Section 2.3, we present various correlation models that describe the geometrical relationship among images, followed by a brief review of joint encoding schemes in Section 2.4. The rest of this chapter focuses on distributed coding. In Section 2.5, we first introduce the theory behind the distributed compression of correlated sources proposed by Slepian-Wolf and Wyner-Ziv. Then, we present the methodologies proposed in the literature for the distributed compression of correlated sources and images. Finally, in Section 2.6 we summarize the drawbacks in the existing distributed representation schemes.

2.1 Camera models

In this section, we give a brief overview of the projective geometry for three common camera models namely perspective, omnidirectional and compressive sensing sensors.

2.1.1 Perspective vision sensors

The perspective imaging system is based on the pinhole camera model [1, 2] and it is widely used in several imaging applications. The geometric model of a pinhole camera is shown in Fig. 2.1 and it consists of an image plane I and a point C. The point C is known as the optical center or the focal point of the camera. The line from the camera center C that is perpendicular to the image plane is known as the principal axis, and it intersects with the image plane I at the principal point p. The distance between the principal point p and camera center C represents the focal length f of the camera. In this scenario, the rays of light reflected or emitted from the objects pass through the camera center C and intersect at the image plane I to form an image. As shown in Fig. 2.1, a point in the 3D space with coordinates $\mathbf{M} = (X, Y, Z)$ is projected to a point $\mathbf{z} = (m, n)$ on the image plane I. The scene point $\mathbf{M} = (X, Y, Z)$ and the image point $\mathbf{z} = (m, n)$ can be related as

$$m = \frac{fX}{Z}, \quad n = \frac{fY}{Z}. \tag{2.1}$$

From Eq. (2.1), we observe a non-linear relationship between the 3D scene point $\mathbf{M}$ and its corresponding 2D projected point $\mathbf{z}$. To avoid this non-linearity, one could represent both the scene and the image points

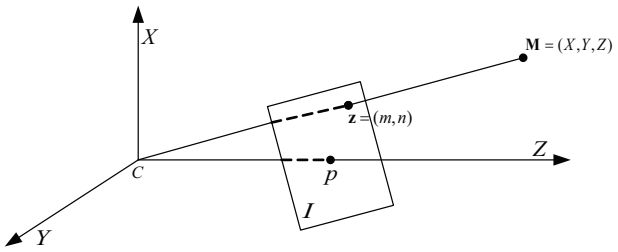


Figure 2.1: Pinhole camera model that describes the relationship between a 3D point (X, Y, Z) and its corresponding 2D projection (m, n) onto the image plane I [1, 2].

in homogeneous coordinates. The latter leads to a linear relationship between the 3D and 2D points, thus making the image projection process much easier to handle mathematically. Using homogenous vector representation, Eq. (2.1) can be rewritten as

$$\lambda \begin{bmatrix} m \\ n \\ 1 \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}, \tag{2.2}$$

where $\lambda \neq 0$ is the homogenous scaling factor.

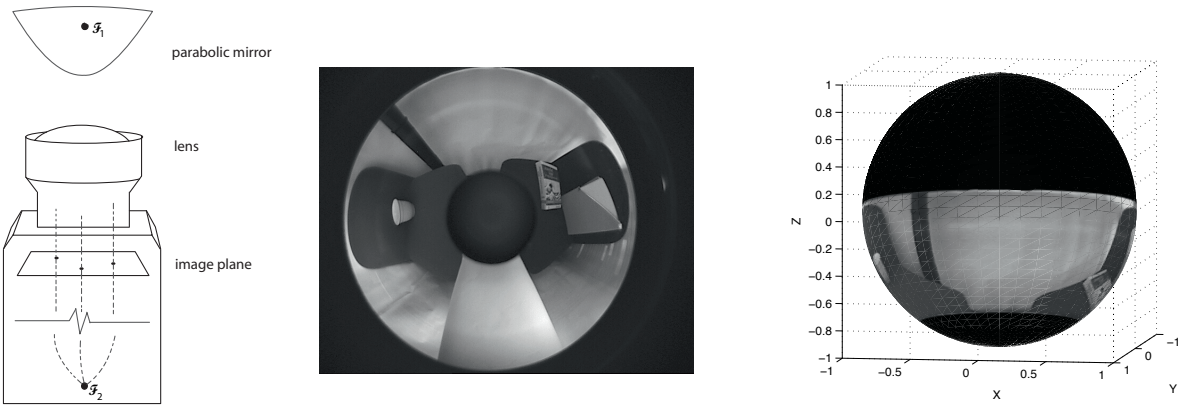
In essence, the perspective pinhole camera model projects the 3D scene points onto a 2D rectangular sensor grid to form the image. Despite its wide availability, such camera models provide only a limited view of the 3D scene due to the parallel ray assumption. Furthermore, rectangular sampling grid is not ideal for image representation, as the photoreceptors in our human fovea is organized on the non-planar surfaces considering the radial form of light.

2.1.2 Omnidirectional vision sensors

The plenoptic function [9] provides a mathematical model for early vision that measures the visual information from every possible position and direction in the 3D space. By dropping the chromatic and time components in the function, the viewing angle at a given position is represented in two dimensions. In particular, due to its radial form, the viewed intensity distribution is best described on the sphere S^2 and not in the rectangular grid. Therefore, the plenoptic function can be considered as a perfect model for a vision sensor, and hence the spherical image formation is interesting for 3D scene representation. In this context, we describe the omnidirectional cameras whose output can be uniquely mapped on the sphere.

Omnidirectional cameras typically use catadioptric system in order to capture 360^0 view of the scene [11]. Such a camera system contains a mirror (e.g., parabolic, elliptical) placed in front of a perspective camera, where the optic axis of the lens is aligned with the mirror's axis [10, 12]. The catadioptric camera systems constructed using quadratic mirror are of particular interest, as they acquire images with a single center of projection. Such systems are also known as central catadioptric systems. The mirrors satisfying this property are parabolic, hyperbolic and elliptical.

Fig. 2.2(a) shows a central catadioptric system with a parabolic mirror. In such a case, the ray of light incident with the focus of the parabola is reflected in a ray of light parallel to the parabola's axis. This construction is equivalent to a purely rotating perspective camera. A typical omnidirectional image captured by this system is shown in Fig. 2.2(b). We observe from Fig. 2.2(b) that the vertical straight lines of the 3D objects form radial lines or conic surfaces in the catadioptric plane [10, 13, 14]. This indicates that it is not straightforward to analyze the omnidirectional images, and one has to take into account the inherent imaging geometry while processing such data. However, due to the single centre projection property of



(a) (b) (c)

Figure 2.2: (a) Central catadioptric system with a parabolic mirror [10]. The parabolic mirror is placed at the parabolic focus $\mathcal{F}_1$ and the other focus $\mathcal{F}_2$ is at infinity. (b) An omnidirectional image captured by such a system. (c) The captured image is mapped on the sphere.

the central catadioptric cameras, it is possible to map the catadioptric images on the sphere and process these spherical images efficiently. In particular, Geyer *et al.* [10] showed that the central catadioptric system which captures the omnidirectional images is composed of two mappings on the sphere. The first mapping represents spherical projection, where the center of the sphere coincides with the focal point of the mirror; the second mapping represents the projection of a point on the sphere's principal axis to the plane perpendicular to that axis. It should be noted that the first mapping is independent of the mirror's shape. On the other hand, the second mapping depends on the mirror's shape which changes the projection point on the sphere's axis as a function of the eccentricity ϵ of the conic. Using the equivalence between the composite mapping on the sphere and the catadioptric projection, the catadioptric images can be mapped on the sphere through inverse stereographic projection [10]. For example, Fig. 2.2(c) shows the spherical representation of the catadioptric image in Fig. 2.2(b).

2.1.3 Compressive sensing camera

In the standard imaging sensors, i.e., perspective and omnidirectional, the images are usually acquired, and compression is performed later in order to reduce the transmission rate. Instead of acquiring the entire image, it is possible to directly acquire the compressed image in the form of linear measurements using compressed sensing (CS) cameras [15, 16]. Such cameras compute only few linear projections of the visual information and thereby significantly reduces the computational cost at the sensor nodes.

In more details, standard imaging sensors measure the per-pixel brightness and chromatic values of the scene objects. The measured brightness and chromatic values form an image that usually has some typical structure, e.g., piecewise smooth. This structural information can be exploited by an encoder that efficiently compresses the data without significant distortion. Though these sensors finally store or transmit only the most important visual information, they acquire and process the entire data. As a result, the number of samples they need to capture and process grows linearly with the resolution of the image.

Instead, the compressed sensing (CS) cameras simultaneously acquire and compress the scene view. That is, they directly acquire the compressed scene information without explicit image sampling or acquisition stage [15, 16]. Unlike standard imaging sensors, the CS sensors provide a representation of the image or signal data in the form of linear measurements, which are usually of lower dimension than the image resolution. These small number of linear measurements, that represent the compressed image contain enough

information for the image reconstruction. In more words, let Y be the linear measurements computed from the image I by projecting on a random set of coding vectors $\{\phi_i\}$. In matrix notation, these measurements are computed as

$$Y = \Phi I, \tag{2.3}$$

where Φ is the measurement matrix formed as $\Phi = [\phi_1|\phi_2|\dots|\phi_K]^T$. This set of measurement vectors $\{\phi_i\}$ can be constructed using random realizations of Gaussian or Bernoulli random variables. The compressed data is directly acquired in the form of linear measurements that are computed using simple inner product operations, i.e., $Y_i = \langle I, \phi_i \rangle$. Therefore, the encoder stage becomes computationally cheap which further translates into power saving. For example, Mamaghanian *et al.* [17] have developed a compressed sensing based acquisition system that is shown to have low power consumption and low complexity compared to the systems that implement transform-based image/signal compression.

While the encoding is very simple, efficient reconstruction strategies have to be activated at decoder to recover the visual information. Donoho [15] and Candes *et al.* [16] showed that a sparse signal in a particular basis Ψ can be recovered from a small number of linear measurements taken on the measurement basis Φ, if Φ and Ψ are incoherent. Especially, if the signal is K sparse then one need approximately βK linear measurements (typically $\beta = 3$ *or* 4) to reconstruct the signal with high probability [18]. The sparse coefficients $c = \Psi^* I$ of the underlying image can be estimated as a solution to the following optimization problem:

$$\hat{c} = arg\min_{c} \|c\|_1 \quad \text{s.t.} \quad \|Y - \Phi\Psi c\|_2 = 0. \tag{2.4}$$

The estimated coefficient vector $\hat{c}$ is used to reconstruct the underlying image as $\hat{I} = \Psi\hat{c}$. Instead of using a sparsity prior, Candes *et al.* [18] proposed an alternative recovery model that estimates a smooth image based on solving the following optimization problem:

$$\hat{I} = arg\min_{I} \|I\|_{TV} \quad \text{s.t.} \quad \|Y - \Phi I\|_2 = 0, \tag{2.5}$$

where $\|I\|_{TV}$ represents the sum of magnitudes of the discrete gradients of $I(m,n)$ computed as

$$\|I\|_{TV} = \sum_{m,n} \sqrt{(I(m+1,y) - I(m,n))^2 + (I(m,n+1) - I(m,n))^2}. \tag{2.6}$$

In the literature, plethora of tools are available to solve the optimization problems given in Eq. (2.4) and Eq. (2.5). We refer the reader to the overview papers [19, 20, 21] that summarize various optimization methodologies.

In rest of this section, we briefly explain some of the CS hardwares available in the literature. The first CS imaging sensor was developed in Rice and it uses optical principles to directly compute the linear measurements [22, 23]. The block diagram is shown in Fig. 2.3, where the light field from the object is focused on a digital micro-mirror device (DMD) that consists of several tiny mirrors. Each tiny mirror corresponds to a particular pixel in the image that can be further independently oriented in one of two directions. In one of the directions, there is another photodetector that collects the reflected light from the DMD, as shown in Fig. 2.3. Therefore, depending on the random configuration of the DMD mirrors, the photodetector actually measures the linear combinations of the image intensity values with the binary basis vectors $\{\phi_i\}$. Before computing the next measurements, the configuration of the DMD mirror pattern is varied based on the random generator RNG. The measurements are therefore computed serially by adapting the DMD mirror patterns.

In the MIT random lens imaging camera [24], the usual lens is replaced by a random lens using reflective or refractive elements. The random lens distributes the light intensity of a particular pixel to several pixel locations of the measurement array. Contrary to the single pixel camera, the random lens imaging system allows to record all the measurements simultaneously. However, the main drawback of this approach is that

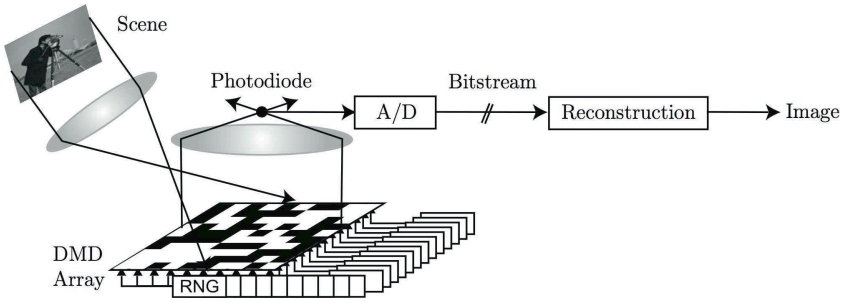


Figure 2.3: Block scheme of the single pixel camera developed in Rice [22].

the random lens requires pre-calibration to find out the measurement matrix Φ. Following similar ideas, Marcia *et al.* [25] have proposed a compressive imaging architecture based on coded aperture imaging. In this imaging architecture, a binary aperture is used to modulate the light field emitted from the source. The aperture pattern can be conveniently represented by a binary matrix whose value is 1 if that particular aperture element is transparent, and 0 if it is opaque. The resulting modulated intensity that represents the weighted linear combinations of image intensity values are measured by a photo detector.

The compressive sensing imagers in [26, 27] are based on Random convolution that first convolves the underlying image with a random pulse and then sub-samples the set of results at random locations [28]. Romberg has proposed an optical CS imager [26] that implements Random convolution in the focal plane using lens and spatial modulators. The equivalent electronic implementation has also been proposed in [29, 27]. In [26], the light field from the source with unknown distribution first passes through a lens that computes the spatial Fourier transform of the incoming light. The resulting Fourier coefficients are spatially modulated using a random mask and the inverse Fourier transform is eventually computed. Finally, the measurements are taken at random sample points using individual photo detectors; it actually represents the weighted linear combinations of the light distribution emanating from the objects in the scene.

Finally, CMOS convolution electronic imager has been proposed in [27] to compute linear measurements of visual information, where the convolution is performed electronically using flip-flop memory and shift registers. The imaging architecture consists of a CMOS pixel array and a one-bit flip-flop memory per pixel that stores the binary filter coefficients. The input and the output of the flip-flop memories of the neighboring pixels are connected to form a shift register that circularly shifts the binary filter coefficients. Initially, the flip-flop memories are set to random binary values that are driven by a pseudo-random generator. The net effect is that the binary weight available in the flip-flop memory is convolved with the analog image data that is available in the CMOS array to provide a measurement value. The second measurement value is computed by adapting the binary coefficients in the flip-flop memory, which can be easily done by shifting the contents in the shift register. Therefore, the measurements are computed serially by appropriately varying the memory contents. The main shortcoming of this camera is that it can compute the measurements only with the binary coefficients, but not with the analog weighted coefficients. The latter representation permits to compute the measurements using FFT or block-based DCT basis vectors to speed-up the sampling process [30, 31]. In this context, Robocci *et al.* [29] have proposed a CS camera based on CMOS architecture that can perform measurement operations with both binary and analog weights. Additionally, it can be used for block-based sensing to speed-up the sampling process. In essence, the CS imagers permit to directly capture the compressed information in the form of linear measurements with a reduced power consumption and low complexity compared to the traditional imaging systems.

2.2 Image representation and coding

The visual information acquired from these sensors cannot be efficiently stored in raw format due to high redundancy in the visual information. In this section, we provide an overview of the image coding tools that reduce the storage and transmission costs of the three systems described in the previous section.

Coding of Planar Images

For decades, the problem of coding planar images has been widely studied in the literature. Earlier attempts on image compression are based on vector quantization (VQ) that quantizes group of pixels together [32]. Despite its optimal compression performance, VQ suffers from high computational cost and delay that grows exponentially with the image dimensions. Due to these difficulties, practical coding algorithms have been designed based on transform coding. These algorithms reduce the redundancy among the samples by first representing the signal in a different domain followed by a scalar quantization. The rate-distortion performance of transform-based coding approaches competes with the VQ scheme with the additional advantage of low computational complexity.

The coding efficiency of the transform-based solutions depends on the ability of the basis functions $\{\psi_i\}$ to sparsely represent the features of the image or signal. In this context, the two most successful representations are the Discrete Cosine Transform (DCT) and the Wavelet Transform [33] that have been used for building JPEG [34] and JPEG 2000 [35] image coding standards respectively. In these standards, the image I is represented as a linear combination of basis functions decomposed as

$$I = \sum_{i=1}^{K} \psi_i c_i = \Psi c, \tag{2.7}$$

where Ψ is an orthonormal basis function (e.g., DCT, Wavelets), and c is the coefficient vector. The computed coefficient vector c is then quantized and entropy coded. At the decoder, the quantized coefficient $\hat{c}$ is finally used to reconstruct an image by inverting the transform given as $\hat{I} = \Psi^* \hat{c}$, where Ψ^* represents the conjugate transpose of the matrix Ψ. More details are available in [34, 35]. The orthonormal basis vectors however, are inefficient in sparsely approximating the edges or singletons in the images, because they cannot deal with the image geometrical regularities. These features can be sparsely represented by including anisotropy and orientation in the basis vectors, e.g., bandelets [36], curvelets [37] or over-complete dictionaries [38]. By including directionality and orientation properties, the orthogonality condition among the basis functions might be relaxed. Non-orthogonal or redundant transforms provide a richer image representation that captures most of the signal or image energy in few coefficients and leads to good coding performance [38].

Coding of Omnidirectional Images

Coding of omnidirectional images has been largely overlooked in the literature. One simple solution is to consider the omnidirectional image as a planar image, and then the unwrapped images can be processed using the planar transformation tools (e.g., Wavelet). Such an approach however fails to take advantage of the geometry of the imaging system and therefore results in degraded performance. Instead, by using the equivalence between the catadioptric projection and composite mapping onto the sphere (described in Section 2.1), the omnidirectional images can be mapped on the sphere. This mapping is equivalent to an inverse stereographic projection, and the resulting image lies on the sphere. The compression of the spherical images can be then carried out using transform-based coding approaches, in which the images are transformed using spherical decomposition tools (e.g., Spherical Laplacian Pyramid [39] or Spherical Wavelet Transform [40, 41, 42]) followed by scalar quantization. However, as described previously, these decomposition techniques fail to capture efficiently the edges in the spherical images. To overcome this

difficulty, Tosic *et al.* [43] have proposed a spherical compression framework using a structured over-complete dictionary that leads to redundant representation of the images.

Coding of Linear measurements

As discussed above, the recovery of an underlying image from the respective low dimensional linear measurements is highly complex and non-linear. This is contrary to the transform-based image coding schemes described earlier, where an image is reconstructed by simply inverting the transform matrix Ψ. Therefore, it is clear that in CS frameworks the encoder is very simple and the entire computational complexity is shifted to the decoder.

Though the CS camera directly acquires the compressed image in the form of linear measurements, the linear measurements cannot be transmitted directly due to high entropy. It is necessary to quantize and entropy code the measurements in order to reduce the transmission or storage costs. In this context, few works have been presented in the literature that study the rate-distortion performance of the linear measurements, where the encoding of measurements is carried out using quantization and entropy coder [44, 45, 46]. Their results show that the RD performance of linear measurement is inferior to the RD performance of state-of-the-art image coding solutions, e.g., JPEG, JPEG 2000 [46]. This is because the linear measurements are highly sensitive to quantization noise. Moreover, the reconstructed signal from quantized measurements fail to satisfy the consistent reconstruction property [47]. Hence, it is essential to develop adapted quantization techniques and reconstruction algorithms that reduce the distortion in the reconstructed signal, such as [48, 49]. The authors in [44, 50] have also studied the asymptotic reconstruction performance of the signal under uniform and non-uniform quantization schemes, and they have shown that the non-uniform quantization schemes usually give smaller distortion in the reconstructed signal, comparing to uniform quantization schemes. An optimal quantization strategy for the random measurements has been designed based on distributed functional scalar quantizers [51]. Recently, Wang *et al.* have proposed a progressive scalar quantization technique for lossy compression of linear measurements that exploits the hidden correlation between the linear measurements [52]. These frameworks however mostly concentrate on decoding an image from the quantized measurements and have not been extended to the joint decoding or scene analysis tasks. In this thesis, we propose to properly handle the non-linearities of the measurement quantization for the latter scenarios.

2.3 Correlation models

Now, we shift our attention from single to multiple image scenarios where the J images $I_j, j \in \{1, 2, \ldots, J\}$ represent the same scene acquired at different time instants or viewpoints. In such scenarios, one of the key characteristics is the high correlation between multi-view images or successive images in a video sequence. The correlation between images can be exploited for effective reconstruction of image sets or for joint analysis tasks. However, in practical applications the correlation model among multiple images is not known a priori; it has to be estimated from the correlated images. Therefore, the first step in the joint representation or joint analysis tasks is to accurately model and estimate the actual correlation among multiple images. In the rest of this section, we describe few correlation models that are commonly used in the literature.

In this context, Duarte *et al.* [53, 54] proposed three joint sparsity models (JSM) that define the inter-signal correlation between sources. We briefly describe here the three JSM's.

- JSM-1: This sparsity model assumes that all the correlated signals share a common sparse component with each signal containing a sparse innovative component. The signal at sensor j is represented as

$$I_j = z + z_j, \quad j \in \{1, 2, \ldots, J\}, \tag{2.8}$$

with

$$z = \Psi c \quad \text{and} \quad z_j = \Psi c_j,$$

where Ψ is an orthonormal basis, $\|c\|_0 = K$ and $\|c_j\|_0 = K_j$. Thus, the signal $z = \Psi c$ is common to all J signals with sparsity K in basis Ψ. The innovation signal $z_j = \Psi c_j$ is unique to the specific signal I_j with sparsity K_j in basis Ψ. This sparsity model holds well in a practical scenario where a group of sensors measure a physical phenomenon such as temperature, humidity in a given region. The global effect in the region could have the same effect on all the sensors which is captured by the common part. On the other hand, the local factors attributed to the particular sensor is reflected in the innovative component.

- JSM-2: This model assumes that all the signals share a common sparse support, but with different coefficients. The signal at sensor j is represented as

$$I_j = \Psi c_j, \quad j \in \{1, 2, \ldots, J\}, \tag{2.9}$$

where the number of non-zero coefficients in vector c_j is K, i.e., $\|c_j\|_0 = K$. This sparsity model holds well in a practical scenario where a set of sensors (e.g., acoustics) acquire replicas of a same signal, but with different amplitudes and phases caused by different attenuations due to signal path effects.

- JSM-3: This model assumes that all the signals have a non-sparse common signal with an individual sparse innovation associated with each signal. Mathematically, it is represented as

$$I_j = z + \Psi c_j, \quad j \in \{1, 2, \ldots, J\}, \tag{2.10}$$

where Ψ is an orthonormal basis, $\|c_j\|_0 = K_j$ and the signal z is the common part that is not necessarily sparse in the basis Ψ. Therefore, JSM-3 model is an extension of the JSM-1 model, where the common part z is not sparse. A practical situation well-modeled by JSM-3 is where different sensors record the static or dynamic scenes. In such scenarios, the signal captured by each sensor is not necessarily sparse, while the ensemble of images captured by different sensors is highly correlated with sparse variations. For example, in a given video sequence each video frame may not be sparse, but the differences between the video frames can be sparse in some basis Ψ.

These three sparsity models however are not ideal in modeling the correlation between multi-view images and successive images in a video sequences due to the large motion. In such scenarios, the inter-pixel geometrical relation between frames is given by the epipolar and optical flow constraints respectively. In the next paragraphs, we review both constraints starting with the epipolar geometrical constraint in multi-view imaging systems.

Epipolar geometry constraint

When a 3D scene is viewed from several angles, the geometrical relationship between multiple images and the observed 3D scene is provided by the epipolar geometry. In particular, it establishes a geometrical relationship between a 3D point and its corresponding projection onto the 2D image plane. The epipolar geometry for two cameras is illustrated in Fig. 2.4, where a 3D point $\mathbf{M} \in \mathbb{R}^3$, their respective projected points $\mathbf{z}_1 \in \mathbb{R}^2$ and $\mathbf{z}_2 \in \mathbb{R}^2$ (in the views I_1 and I_2) and the cameras C_1 and C_2 are coplanar and lie in the same plane [1, 2]. More formally, the epipolar constraint is given by the following mathematical equation:

$$\mathbf{z}_2^T \hat{T} R \mathbf{z}_1 = 0, \tag{2.11}$$

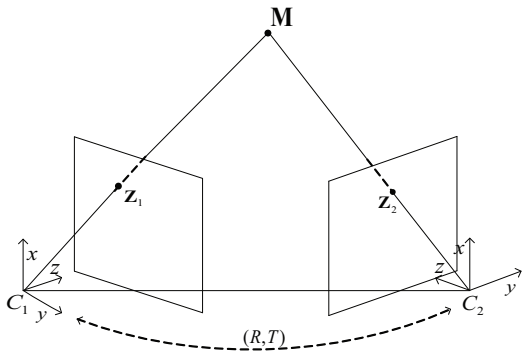


Figure 2.4: Epipolar geometry for the perspective cameras C_1 and C_2 [1, 2]. The relative pose between cameras C_1 and C_2 is represented by a rotation matrix R and a translation vector T.

where R and T respectively represent the rotation and translations between cameras C_1 and C_2. Given the translation vector $T = [t_1\ t_2\ t_3]$, the skew symmetric matrix $\hat{T}$ can be expressed as

$$\hat{T} = \begin{pmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{pmatrix}. \tag{2.12}$$

The difference between the two projected points $\mathbf{z}_1$ and $\mathbf{z}_2$ corresponding to the same scene point $\mathbf{M}$ is usually denoted as *disparity*. It should be noted that, by knowing the relative pose (R, T) between cameras and the two projected points $\mathbf{z}_1$, $\mathbf{z}_2$ that satisfy Eq. (2.11), the 3D coordinates of the scene point can be calculated by triangulation [1, 2]. Therefore, the correlation estimation problem in multi-view imaging networks consists in finding the corresponding projection points between images. In other words, we need to estimate the difference between the pixel coordinates $\mathbf{z}_1$ and $\mathbf{z}_2$ corresponding to the same scene point. Finally, the epipolar equation given in Eq. (2.11) can be used to describe the relationship between spherical images, by defining the projected points $\mathbf{z}_1 \in S^2$ and $\mathbf{z}_2 \in S^2$ in the Spherical coordinate system rather in Euclidean coordinates [55, 56].

Optical flow constraint

The optical flow constraint describes the geometrical relationship between time varying images acquired by a video camera, that is, the relationship between pixels in successive images when a dynamic scene is captured by a camera. Let us assume that two images I_1 and I_2 are sampled at time t and $t+1$ respectively. We assume that the position of a 3D point at time t is $\mathbf{M} \in \mathbb{R}^3$ and it is projected at a 2D position $\mathbf{z}_1 = (m, n)$ in view I_1. Due to the motion of scene objects, the same scene point is moved to the position $\mathbf{M}' \in \mathbb{R}^3$ at time $t+1$, and therefore it is projected at location $\mathbf{z}_2 = (m+u, n+v)$ in the image I_2, where u and v represent the horizontal and vertical displacement components respectively. The relationship between the points $\mathbf{z}_1$ and $\mathbf{z}_2$ is given by the following optical flow constraint equation [57, 58] that is formally expressed as

$$u\frac{\partial I}{\partial m} + v\frac{\partial I}{\partial n} + \frac{\partial I}{\partial t} = 0, \tag{2.13}$$

where $\frac{\partial I}{\partial m} = I_1(m,n) - I_1(m+1,n)$, $\frac{\partial I}{\partial n} = I_1(m,n) - I_1(m,n+1)$ and $\frac{\partial I}{\partial t} = I_1(m,n) - I_2(m,n)$. Eq. (2.13) says that the temporal intensity variation is explained by the spatial intensity variation multiplied by the velocity or displacement of the scene point moving with respect to the camera. Moreover, Eq. (2.13) is derived by assuming that the rate of change of intensity along the motion trajectory is null. In other words,

it is derived based on the constant brightness assumption, expressed as $I_1(\mathbf{z}_1, t) = I_2(\mathbf{z}_2, t)$. The difference between the pixel coordinates $\mathbf{z}_1$ and $\mathbf{z}_2$ corresponding to the same scene point is usually called *motion* or *optical flow*. The motion field can also be used to estimate the underlying 3D scene geometry with structure from motion principles [3]. The same analogy also applies for the motion field estimation to omnidirectional images, where we need to find the correspondences between images defined on the sphere rather on the rectangular grid [39].

Finally, in both multi-view and video imaging scenarios, the correlation estimation problem boils down in finding the differences between corresponding pixels in different images. In other words, we need to solve the correspondence problem that finds the patches or pixels in multiple images corresponding to the projection of the same scene element. The correspondence problem can be solved either pixel-wise or block-wise, where the former representation provides an accurate model at the expense of computational complexity. For more details, we refer the reader to the overview papers on correlation estimation [7, 8]. In the next section, we describe the usage of these correlation models for encoding multiple images.

2.4 Joint encoding of correlated images

In this section, we briefly review the joint coding approaches for encoding sets of correlated images that are captured by a video camera or multiple cameras. This compression problem is quite different from the independent image coding schemes described in Section 2.2, due to the presence of correlation components between the sets of images. Therefore, one has to take benefit of both the intra- and inter-view redundancies in order to maximize the RD performance. In the joint encoding approach, the central encoder usually collects the multiple images and it removes the inter-view redundancy between images based on disparity or optical flow models, as described in Section 2.3. This problem has been widely studied in the literature.

In this context, several joint encoding frameworks for compressing stereo image pairs have been proposed, e.g., [59, 60, 61, 62], where the underlying correlation between images is exploited using a block-based disparity estimation and compensation. These schemes attempt to minimize the mean-square error (MSE) between the original stereo image pair and the reconstructed stereo image pair for a given target bit rate. Contrary to this, Perkins [63] studied the effect of stereo image compression on the user's ability to perceive the disparity. In particular, he has proposed a multi-resolution coding algorithm that is perceptually justified when the compressed stereo pair is viewed by a human.

Few coding standards have been proposed for compressing the time varying images based on block-based motion estimation and compensation, e.g., H.264 [64]. Recently, multi-view video coding schemes based on H.264 standards have been proposed in [4, 65], where the inter-view correlation is exploited using block-based disparity models. Later, similar ideas have been used for the compressive video acquisition based on linear measurements [66, 67, 68, 69], in effort to reduce the sampling rate of the video acquisition. Rather than encoding all the images independently, these frameworks employ block-based motion compensation and estimation techniques to improve the sparse representations of the images. However, block-based translation models that are commonly used for correlation estimation fail to efficiently capture the geometry of scene objects. A better correlation model can be estimated by comparing the most prominent features in different images, where the most prominent features are captured by dictionary functions in a redundant dictionary [70].

The joint coding frameworks however, necessitate communication among sensors that is highly undesirable or expensive for low power applications. Furthermore, the complexity at the encoder is high, as it computes motion or disparity at the central encoder. In addition, communication between cameras is often highly undesirable due to power constraints or due to network topology (for example). We therefore focus from now on distributed coding solutions where the correlated sources are encoded independently, but decoded jointly.

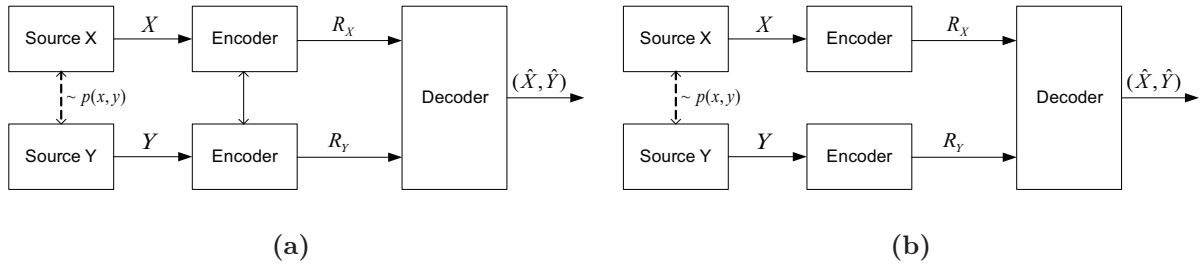


Figure 2.5: Two different coding methods to compress the correlated sources X and Y, where the dependency between sources is given p.d.f. $p(x, y)$. (a) Joint encoding. (b) Distributed coding. In both cases, the encoded sources are jointly decoded to exploit the correlation between the samples X and Y.

2.5 Distributed coding

Distributed source coding (DSC) refers to the independent coding of two or more physically separated correlated sources that are not communicating with each other; the compressed streams are sent to a central unit and decoded jointly. The distributed compression finds its theoretical foundation in two papers, the first one by Slepian and Wolf [6] and second one by Wyner and Ziv [71]. We briefly state the Slepian-Wolf and Wyner-Ziv thereoms without any mathematical proof. For more details we refer the reader to [6, 71], and to the introductory articles on distributed coding [72, 73].

2.5.1 Theoretical foundations of distributed coding

Consider $\{(X_i, Y_i)\}_{i=1}^{n}$ as the sequence of independent drawings of a pair of correlated random variables X, Y from a given distribution $p(x, y)$. The sequences X and Y are statistically dependent and the dependency between them can be described by the conditional probability mass function $P(X|Y)$. Let us first consider the joint encoding and joint decoding of the correlated sources X and Y as shown in Fig. 2.5(a). The question to ask is: what is the minimum encoding rate R required to perfectly recover (without any errors) the two sources X and Y at the joint decoder ? The answer to this question derived from information theory is the joint entropy $R = H(X, Y) = H(Y) + H(X|Y) = H(X) + H(Y|X)$, where $H(.)$ is the entropy function. For example, this can be achieved by encoding Y at a rate $R_Y = H(Y)$ and based on the complete knowledge of Y at both the encoder and the decoder, the source X is encoded at a rate $R_X = H(X|Y)$ where $H(X|Y)$ is the conditional entropy of X given Y. Next, we consider the distributed compression of sources X and Y as shown in Fig. 2.5(b), where the correlated sources X and Y are separately encoded but the decoding process is performed jointly. Certainly, encoding the sources X and Y at rates $R_X \geqslant H(X)$ and $R_Y \geqslant H(Y)$ guarantees error free reconstruction at the joint decoder. However, for the correlated sources X and Y, the total rate $R \geqslant H(X) + H(Y)$ is greater than the joint entropy $H(X, Y)$. The real question is: is it possible to perfectly recover the sequences at the decoder, by encoding them at a total rate smaller than the sum of individual entropies ?

In 1970's, Slepian and Wolf studied this distributed coding problem and showed that a total rate $R = H(X, Y)$ is sufficient to perfectly reconstruct the sources. In particular, they showed that there is no loss in coding efficiency in separate encoding compared to joint encoding, as long as joint decoding is performed. More specifically, for the two correlated sources X and Y the achievable rate region is given by

$$R_X \geqslant H(X|Y), R_Y \geqslant H(Y|X), R_X + R_Y \geqslant H(X, Y). \tag{2.14}$$

The achievable rate region or the Slepian-Wolf region for two correlated sources X and Y is shown in Fig. 2.6. The corner points U and V in the Slepian-Wolf region represent a special case of the distributed

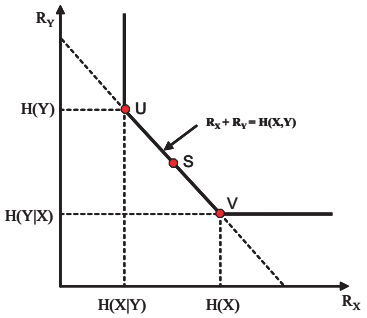


Figure 2.6: Achievable Slepian-Wolf rate region for the two correlated sources X and Y.

source coding, often referred as *coding with side information* or asymmetric coding. For this special case, Pradhan *et al.* [74] proposed a practical and constructive design approach to realize Slepian-Wolf theorem based on channel codes. We illustrate the relationship between the channel codes and the Slepian-Wolf theorem for the corner point U in Fig. 2.7, where the source X is compressed based on the knowledge of Y that is available at the decoder. In a channel coding framework, in order to recover the sequence X from the noisy observation Y one has to do channel coding for X. The encoder therefore transmits parity bits that are used by the decoder to perfectly recover X from its noisy observation Y. From the Slepian-Wolf coding point of view the correlated sources X and Y can be modeled by a virtual dependency channel with X as the channel input and Y as the *noisy* output. The virtual dependency channel motivates the authors in [74] to employ channel coding techniques in DSC. Therefore, the error between sources X and Y can be corrected by applying channel coding to the source X. Hence, the encoder shown in Fig. 2.7 is actually a channel encoder that generates the parity bits from source X. The transmitted parity bits are later used by the decoder to reconstruct X from Y.

Therefore, it is clear that the Slepian-Wolf encoding problem is actually a channel coding problem, and the Slepian-Wolf bounds can be achieved by using capacity approaching channel codes. Practical DSC schemes have been first designed by establishing a relation between the Slepian-Wolf theorem and channel coding [74]. They brought channel coding ideas in the Slepian-Wolf source coding problem and they used Trellis channel code to yield a practical Slepian-Wolf system. This further motivated a lot of researchers to use better channel codes. In this context, systems with better compression performance were developed using Turbo codes [75, 76]. Liveris *et al.* [77] introduced the encoding strategy with LDPC codes and they showed that the LDPC code is a better alternative to the Turbo code. However, both the LDPC and Turbo codes reach the channel capacity asymptotically, and therefore it requires large block sizes in order of 10^5 to reach the Slepian-Wolf bound. In practice, the source lengths are usually small in the order of few hundreds to few thousands. Recently, M. Grangetto *et al.* [78] built an asymmetric distributed coding scheme using the Arithmetic codes that provide a good performance at small block lengths compared to LDPC and Turbo codes. For more details, we refer the reader to the overview articles [72, 73].

In 1976, Wyner and Ziv extended the problem of coding with side information to the lossy coding scenario [71]. It is the problem of lossy compression of one source (e.g., source X) when the other source (e.g., source Y) is available at the decoder, as shown in Fig. 2.8. Wyner and Ziv considered an average distortion D between the original X and the decoded $\hat{X}$ versions, and they computed the minimum rate $R^{WZ}(D)$ required to encode X under the constraint that the average distortion between X and $\hat{X}$ is $E\{d(X,\hat{X})\} \leqslant D$. In this scenario, Wyner and Ziv proved that the transmission rate increases, when the statistical dependency between sources is exploited only at the decoder compared to the case where the dependency is exploited at both the encoder and the decoder. Mathematically, the Wyner-Ziv theorem is stated by the following equation,

$$R^{WZ}(D) \geqslant R_{X|Y}(D), D \geqslant 0, \tag{2.15}$$

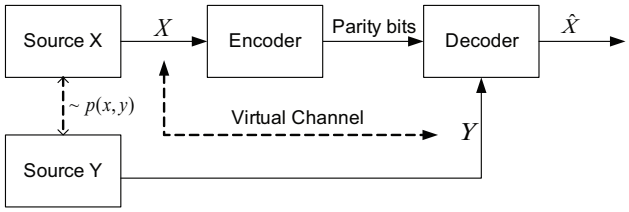


Figure 2.7: Relationship between Slepian-Wolf coding and channel coding. The correlation between sources can be modeled by a virtual channel with X as the channel input and Y as the output. The source X can be perfectly recovered from the noisy observation Y using parity bits.

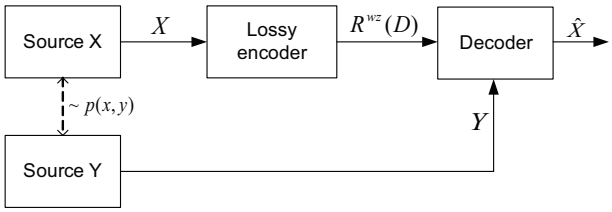


Figure 2.8: Wyner-Ziv coding scheme: Lossy compression of source X with side information Y available at the decoder [71].

where $R^{WZ}(D)$ and $R_{X|Y}(D)$ are rate-distortion functions with the average distortion D. $R^{WZ}(D)$ represents the minimum encoding rate for X when Y is available only at the decoder, and $R_{X|Y}(D)$ represents the minimum encoding rate for X when Y is simultaneously available at both the encoder and the decoder. They also showed that the equality sign in Eq. (2.15) holds when X and Y are jointly Gaussian. Finally, as a special case when $D = 0$, Eq. (2.15) becomes $R^{WZ}(0) = R_{X|Y}(0)$, i.e., the Slepian-Wolf theorem is a special case of the Wyner-Ziv theorem. This shows that it is possible to reconstruct X with an arbitrarily small probability of error, when the side information Y is available only at the decoder. For more details, we refer the reader to [71].

2.5.2 DSC for traditional imaging systems

The Slepian-Wolf and Wyner-Ziv theorems show that the correlated sources can be distributively compressed with a minimum coding loss when compared to the joint encoding. However, it is not straightforward to apply these theoretical results for multi-view or video compression. One has to solve the following issues in order to design a practical distributed coding system. The statistical dependency between sources given in terms of conditional mass function $P(X|Y)$ is not an ideal model for describing the correlation between multi-view images and video sequences. In such scenarios, as described in Section 2.3, the correlation model for accurate scene representation is effectively described by the disparity or motion models. Furthermore, the Slepian-Wolf and Wyner-Ziv results come under the assumption that the statistical dependency (or equivalently the correlation model parameters) between the Wyner-Ziv encoded source and the side information is perfectly known at the encoder. In practical distributed scenarios, the side information is available only at the decoder and not at the encoder. Therefore, while designing a practical distributed coding solution, one has to accurately model the characteristics of the channel and estimate the channel model parameters at the encoder for controlling the Slepian-Wolf coding rate.

The first practical distributed coding scheme for compressing the time varying images captured by a video camera has been proposed in [79, 80], where the video frames are independently encoded and jointly decoded by exploiting the correlation between images using block-based motion compensation. In

[79], the video frames are categorized into key frames and Wyner-Ziv frames with a GOP size of 2, i.e., $I_j, j \in \{1, 3, \ldots, (2\lceil J/2 \rceil) - 1\}$ represent the key frames and $I_j, j \in \{2, 4, \ldots, 2\lfloor J/2 \rfloor\}$ represent the Wyner-Ziv frames, where J represents the number of frames in the video sequence. The key frames are encoded independently using standard coding solutions, e.g., JPEG 2000 or H.264. The Wyner-Ziv frames are first transform coded (e.g., DCT) followed by a scalar quantization with 2^M levels. Then, the quantized coefficients are represented in M bitplanes, and each bitplane is finally channel coded (e.g., Turbo codes). The resulting parity bits are stored in the buffer and transmitted to the joint decoder upon request. The joint decoder first estimates a side information $\tilde{I}_j$ using block-based motion compensation and interpolation from the decoded key frames, denoted as $\hat{I}_{j-1}$ and $\hat{I}_{j+1}$. The side information $\tilde{I}_j$ efficiently captures the low frequency components, but not the high frequency components, as the motion compensation usually fails to efficiently capture the visual information along the edges and in texture regions. Therefore, the side information can be considered as a noisy version of the original Wyner-Ziv frame, and the noise can be corrected using the parity bits transmitted from the encoder. Using a feedback channel, the decoder then requests the encoder to transmit the parity bits, and this process is repeated until the error in the side information is corrected. The Slepian-Wolf encoding rate in this scenario is controlled in an accurate manner by exchanging the channel statistics using the feedback messages.

The PRISM architecture [80] is very similar to the previous one, except that it has no feedback channel. Instead, the rate control is achieved at the encoder by allowing limited communication between the encoders. Furthermore, in addition to the parity bits, the encoder also transmits a cyclic redundancy check word (CRC) computed from the quantized Wyner-Ziv frames in order to assist the motion estimation/compensation at the decoder. The decoder first selects a set of candidate motion compensated side information blocks from the previously decoded reference frames. The transmitted CRC is then compared with the CRC generated from the decoded block. If there is no deviation, the decoding is reported as successful, otherwise the decoder chooses another candidate block. This process is repeated until successful decoding is achieved. Inspired by these frameworks [79, 80], significant research efforts have been carried out in the past decade in order to improve the compression performance. In particular, a lot of research has been focused on the side information improvement and on the accurate correlation noise modeling ([81, 82]). More details related to recent advances in distributed video coding can be found in the overview articles [83, 84].

Now, we discuss few studies reported in the literature about the application of distributed coding principles to camera networks. The works reported in the literature are generally based on coding with side information, where one camera is used as a reference to decode the information from the other cameras. For example, in [85, 86] the cameras are categorized into reference and Wyner-Ziv cameras and the correlation among views is exploited at the joint encoder using disparity estimation based on epipolar geometry, which usually requires camera parameters. When camera parameters are not available and calibration is not possible, the joint decoder can rather use block-based disparity estimation to exploit the redundancy between images [87, 88]. Gehrig *et al.* [89, 90] have proposed a geometry-based distributed coding scheme for compressing the multi-view images. The authors proposed to represent each view using a piecewise polynomial model. Using this polynomial representation of image, the correlation model is built by relating the locations of discontinuities among different views. However, they consider a special camera arrangement where all the cameras are placed in a straight line. Super-resolution techniques have been applied for the distributed coding in camera networks in [91]. In their framework, each sensor transmits a low resolution image to the decoder. At the decoder, these low resolution images are registered with respect to a reference image, where the image registration is performed by shape analysis and image warping. The registered images are then jointly processed to decode a high resolution image. Distributed compression has also been applied for the multiple images captured in omnidirectional sensor networks [92]. The correlation between omnidirectional images is first estimated that is modeled using the local transformations of the sparse features captured by an over-complete structured dictionary. Then, the estimated correlation model is used for Wyner-Ziv image coding based on partitioning the dictionary into several cosets.

Most of state-of-the-art schemes reported in the literature are based on coding with side information,

i.e., they target the corner points U or V in the Slepian-Wolf rate region shown in Fig. 2.6. It is clear that these frameworks do not balance the transmission rate between the encoders. In practice, it may however be desirable to have more flexibility in the transmission rate among the encoders. The first work that addresses balanced rate allocation in distributed coding is based on time sharing mechanism [93], which is however hard to implement, due to node synchronization issues. The first practical scheme for symmetric coding based on channel code partitioning has been proposed in [94]. This scheme has been later extended to multiple sources using systematic channel codes [95]. It is based on splitting the generator matrix of the channel code into sub-generator matrices. Codewords are then generated using the sub-matrices, and are assigned to each encoder. The compression rate of each encoder is determined by the number of rows retained in the corresponding sub-matrix. The advantage of this system is the need for only one channel code. However, this framework is limited to systematic channel codes. The authors in [96] have developed a symmetric DSC solution using a general linear channel code framework based on algebraic binning concept. Simulation results have shown that almost the entire Slepian-Wolf region can be covered with this coding algorithm. Symmetric distributed coding can also be achieved by information partitioning. In this context, Sartipi *et al.* [97] have considered the compression of two sources by information partitioning, where half of the source bits are transmitted directly, while the corresponding syndrome bits are generated on the other half (complementary part) of the source bits. Similar to [96], the authors show that they can approach the entire Slepian-Wolf region, and thus the decoding error becomes insensitive to an arbitrary rate allocation among the encoders. However, both schemes are based on capacity approaching channel codes, which usually approach the Slepian-Wolf bound only for long source length (typically 10^4). Grangetto *et al.* [98] have proposed a balanced coding scheme for small block length binary sources. The algorithm is based on a time sharing version of distributed Arithmetic codes, which performs better than the Turbo code solution in the considered framework.

DSC with rate allocation has also been considered in imaging applications. The authors in [99] proposed a rate balanced DSC scheme for video sequences. In this scheme, each frame is divided into two partitions and one partition is then transmitted directly. In addition, each frame is Wyner-Ziv encoded and the side information is eventually generated using motion estimation. This scheme permits to avoid hierarchical relations between frames. However, it results in high coding rates, since one of the partitions in each frame is encoded using both Wyner-Ziv and independent coding. Finally, a balanced distributed coding scheme for camera networks has been proposed in [100], based on linear channel code construction that can achieve any point in the Slepian-Wolf region. The proposed linear codes have however not been applied to the actual coding of images in camera networks.

2.5.3 DSC in CS framework

In Section 2.1, we have already described that the CS-based image acquisition provides a low complexity encoding, since it computes the linear measurements with simple inner products. Therefore, it is advantageous to merge the distributed coding and CS-based image acquisition, so that the encoder becomes simple with the entire complexity shifted to the joint decoder. Distributed compressive schemes for the time varying images have been proposed in [101, 102, 103]. Similar to the DVC scheme [79] described in the previous section, these schemes split the video sequences into key frames and compressed sensing frames (instead of Wyner-Ziv frame) with a GOP size 2. The linear measurements computed from the key frames are intra-coded and the respective images are reconstructed at the decoder by solving an optimization problem of the form of Eq. (2.4). The side information $\tilde{I}_j$ is then generated from the reconstructed reference images using block-based motion compensation and interpolation. The procedure to estimate the side information $\tilde{I}_j$ is summarized in Algo. 1, where the reference images I_{j+1} and I_{j-1} are reconstructed from the respective linear measurements Y_{j+1} and Y_{j-1} by solving an l_2-l_1 minimization problem. The reconstructed images are then used to build a side information $\tilde{I}_j$ by linearly interpolating the motion vectors computed between images $\hat{I}_{j+1}$ and $\hat{I}_{j-1}$. However, the generated side information $\tilde{I}_j$ fails to efficiently represent the visual information along the edges and texture regions. An additional joint reconstruction stage is therefore nec-

essary to estimate the missing details in the side information, in order to reconstruct the CS frame I_j. It should be noted that in DVC schemes the error in the side information is usually corrected using the parity bits. In the distributed compressive framework the quality of the side information is improved by estimating the missing details from the visual information provided by the linear measurements.

Algorithm 1 Correlation Estimation algorithm proposed in [101, 102, 103]

1: $\tilde{c} = \arg\min\|c\|_1 \quad \text{s.t.} \quad Y_{j-1} = \Phi\Psi c$
2: $\hat{I}_{j-1} = \Psi^T\tilde{c}$
3: $\tilde{c} = \arg\min\|c\|_1 \quad \text{s.t.} \quad Y_{j+1} = \Phi\Psi c$
4: $\hat{I}_{j+1} = \Psi^T\tilde{c}$
5: $\tilde{I}_j = \text{Motion compensation}(\hat{I}_{j+1}, \hat{I}_{j-1})$

In [102], the joint reconstruction model assumes that the error between the original CS image I_j and the side information $\tilde{I}_j$ is sparse in an orthonormal basis Ψ (e.g., DCT). The authors propose to reconstruct the residual error image $I_e = I_j - \tilde{I}_j$ by solving the following optimization problem:

$$\tilde{c} = \arg\min_c \|c\|_1 \quad \text{s.t.} \quad Y_e = \Phi\Psi c, \tag{2.16}$$

where Y_e represents the error between the measurements generated from the original CS frame and the corresponding side information, i.e., $Y_e = \Phi(I_j - \tilde{I}_j)$. Once the residual image $I_e = \Psi^T\tilde{c}$ is estimated, it is added to the side information $\tilde{I}_j$ in order to reconstruct the CS frame denoted as $\hat{I}_j$.

In [103], the CS frame is reconstructed by assuming that it can be sparsely represented in a block-based adaptive dictionary Ψ_b constructed using the side information. The CS image is reconstructed by solving the following optimization problem:

$$\tilde{c} = \arg\min_c(\lambda\|c\|_1 + \|Y_j - \Phi\Psi_b c\|_2^2), \tag{2.17}$$

where $Y_j = \Phi I_j$, and λ is the regularization constant. Kang *et al.* [101] proposed to reconstruct the CS frame by solving Eq. (2.17) under the assumption that the frame to be reconstructed is sparse in a DCT basis rather than a block-based adaptive dictionary.

Distributed compressive solution have also been proposed in multi-view framework [104, 105]. The linear measurements are computed independently from each view and they are transmitted to joint decoder. The joint decoder first reconstructs all the views by solving an l_2-l_1 optimization problem of the form of Eq. (2.4), assuming that the images are sparse in a particular basis Ψ, e.g., Dual-tree Wavelet, Contourlet. The joint decoder then attempts to improve the reconstruction quality of the independently decoded images by exploiting the underlying correlation between views.

In more details, the authors propose to predict the current view by bi-directionally interpolating the two neighboring reconstructed views $\hat{I}_{j-1}$ and $\hat{I}_{j+1}$. The generated predicted image is denoted as $\tilde{I}_j$. The residual image I_e between the original view I_j and the predicted view $\tilde{I}_j$ is then estimated by solving an l_2-l_1 optimization problem in an assumption that the residual error I_e is sparse in Dual-tree Wavelet or Contourlet basis Ψ. Once the residual image is estimated, it is added to the predicted view $\tilde{I}_j$ to reconstruct $\hat{I}_j$. These steps are summarized in lines $1-4$ of Algo. 2. The authors further propose to improve the reconstruction quality of $\hat{I}_j$ using the disparity vectors estimated from the left and right reference images, denoted as $\hat{I}_{j-1}$ and $\hat{I}_{j+1}$ respectively. The estimated disparity vectors are further used to generate the predicted image $\tilde{I}_j$ based on disparity compensation. The residual image I_e is then estimated by solving an l_2-l_1 optimization problem, and it is further used to update the prediction results. The disparity compensation, residual image estimation and update steps are repeated till convergence or a predefined number of iterations is reached.

Wakin [106] has proposed an image registration and joint reconstruction algorithm for multi-view images

Algorithm 2 Distributed multi-view compression from linear measurements [104, 105]

1: $\tilde{I}_j$ = Image Interpolation$(\hat{I}_{j+1}, \hat{I}_{j-1})$
2: $Y_e = \Phi(I_j - \tilde{I}_j)$
3: $\tilde{c} : \arg\min\|c\|_1 \quad \text{s.t.} \quad Y_e = \Phi\Psi c$
4: Update: $\hat{I}_j = \tilde{I}_j + \Psi^T\tilde{c} = \tilde{I}_j + I_e$
5: **repeat**
6: $\quad \tilde{I}_j$ = Disparity compensation$(\hat{I}_j, \hat{I}_{j+1}, \hat{I}_{j-1})$
7: $\quad Y_e = \Phi(I_j - \tilde{I}_j)$
8: $\quad \tilde{c} : \arg\min\|c\|_1 \quad \text{s.t.} \quad Y_e = \Phi\Psi c$
9: $\quad$ Update: $\hat{I}_j = \tilde{I}_j + \Psi^T\tilde{c} = \tilde{I}_j + I_e$
10: **until** k times

based on manifold lifting. The underlying correlation among multiple views is exploited by constructing an image appearance manifold, where the images represent sample points on the manifold; these points are controlled by a few camera parameters (e.g., rotation, translation). By knowing the initial camera positions, the images are jointly reconstructed based on an l_2-l_1 optimization framework. The reconstructed scene is then used to refine the camera parameters, and thus both the camera positions and the scene are jointly estimated using alternating minimization techniques. In another framework [107], the authors proposed a joint reconstruction algorithm that exploits the underlying correlation model while reconstructing a pair of images. However, in their framework the disparity-based correlation model is estimated from the reconstructed reference images. Further, they failed to consider the effect of measurement quantization in the correlation estimation and the joint reconstruction stages. Therefore, this framework cannot be applied in practical applications. Recently, the authors in [108] have proposed a joint reconstruction scheme based on a regularized optimization framework. The two regularization terms encourage sparse priors of multi-view images and the difference between images. However, this framework is not well suited for practical multi-view and video imaging scenarios, as the correlation among images is usually given in the form of disparity or motion vectors and not as a sparsity prior of the signal differences.

2.6 Concluding remarks

Based on the facts described in this chapter, we would like to address the following points:

- State-of-the-art symmetric encoding schemes [95, 96, 97, 98] assume that the binary sources X and Y are correlated as $X = Y \oplus Z$, where Z is the additive binary noise. The binary correlation model is certainly not an ideal one for the multi-view images captured in sensor networks, since the correlation model in such scenarios is described by a disparity model and not as the image differences. Therefore, efficient compression techniques are required to distributively encode the images captured in sensor networks. In Chapter 3 we propose a rate-balanced distributed coding scheme for encoding the multi-view omnidirectional images.

- State-of-the-art distributed coding schemes (e.g., [79, 80]) usually rely on Slepian-Wolf and Wyner-Ziv results that exploit the statistical correlation between sources using channel codes. The computational complexity at the encoder certainly becomes important due to the channel rate estimation. Therefore, we propose a distributed coding scheme in Chapter 4 that provides an efficient joint representation directly from the independently compressed images. Hence, our framework neither requires an additional encoder rate control module nor a feedback channel.

- Most of the CS-based distributed schemes estimate a correlation model after reconstructing the reference images that is often based on solving a complex optimization problem of the one given in

Eq. (2.4). Also, the current CS-based distributed representation schemes often fail to estimate an accurate correlation model when the measurements are quantized. In Chapter 5, we present our robust algorithms that estimate a correlation model among multiple images using one compressed reference image; while the remaining compressed images given in the form of quantized linear measurements are directly used in the correlation estimation algorithm without priori reconstruction. In Chapter 6, we go one step further and present our robust correlation estimation algorithm that builds the correlation model among multiple images directly from the compressed quantized linear measurements without explicit image reconstruction steps.

- Distributed joint reconstruction of multiple correlated images from linear measurements has not been addressed thoroughly in the literature, especially when the measurements are quantized. Therefore, state-of-the-art schemes cannot be used in practical coding applications, since the measurements are usually quantized in order to limit the bandwidth resources. In Chapter 7, we present our distributed joint image reconstruction algorithms that efficiently handle the non-linearities of measurement quantization.

Chapter 3

Symmetric Distributed Coding of Omnidirectional Images

3.1 Introduction

In this chapter, we present a rate balanced distributed coding scheme[1] for the representation of 3D scenes captured by multiple omnidirectional cameras, as shown in Fig. 3.1. Each catadioptric camera samples the plenoptic function that represents the entire visual information available to the observer [9]. The catadioptric projective geometry theory described in Chapter 2 shows that these sensors capture the light field in the radial form. Therefore, in order to preserve the geometry of the captured light information, we propose to work directly in the spherical domain by appropriately mapping the omnidirectional images on the sphere.

We design a distributed coding scheme based on the Wyner-Ziv results presented in Chapter 2, where the images are encoded using a transform-based coding solution followed by a Slepian-Wolf encoding. In more words, the spherical images initially undergo a multi-resolution decomposition based on the Spherical Laplacian Pyramid (SLP) that provides shift invariance. The resulting sets of coefficients are quantized and then split into two correlated partitions. The quantized coefficients of the first partition are entropy coded, and sent to the decoder. The second partition is encoded using Nested Scalar Quantization (NSQ) [109], which is a binning scheme that encompasses a scalar quantizer and a coset encoder. It outputs the coset bin indexes and permits to reduce the coding rate compared to the direct encoding of quantized coefficients. The coset bin indexes are further encoded using a Slepian-Wolf encoder based on multi-level LDPC codes [110, 111], in order to achieve further compression of the distributed coded images. The resulting syndrome bits are finally transmitted to the joint decoder.

The joint decoder estimates the quantized coefficients of the second partition from the quantized coefficients of the first partition by exploiting intra-view correlation. Furthermore, the joint decoder takes benefit of the correlation between views by performing block-based disparity estimation on the sphere [39], which matches similar blocks of solid angles from two omnidirectional images in the spherical domain. Therefore, the proposed scheme efficiently combines the intra and inter Wyner-Ziv image coding, which allows for a balanced coding rate between cameras. Such a strategy proves to be beneficial with respect to the independent processing of omnidirectional images and shows only a small performance loss compared to the joint encoding of different views. Moreover, we exploit the inter-view correlation by block-based disparity estimation without using epipolar geometry constraints. Hence, our scheme does not require any camera parameters which are usually required in the techniques based on epipolar geometry to perform the correspondence matching (e.g., [112]). This is certainly advantageous in camera networks where the

[1] Part of this work has been published in: V. Thirumalai, I. Tosic and P. Frossard, "Symmetric Distributed Coding of Stereo Omnidirectional Images," Signal Processing: Image Communication, vol. 23(5), pp. 379-390, 2008.

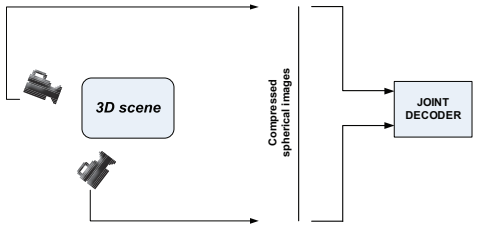


Figure 3.1: Distributed coding of the 3D scenes. The correlated images are compressed independently and are decoded jointly.

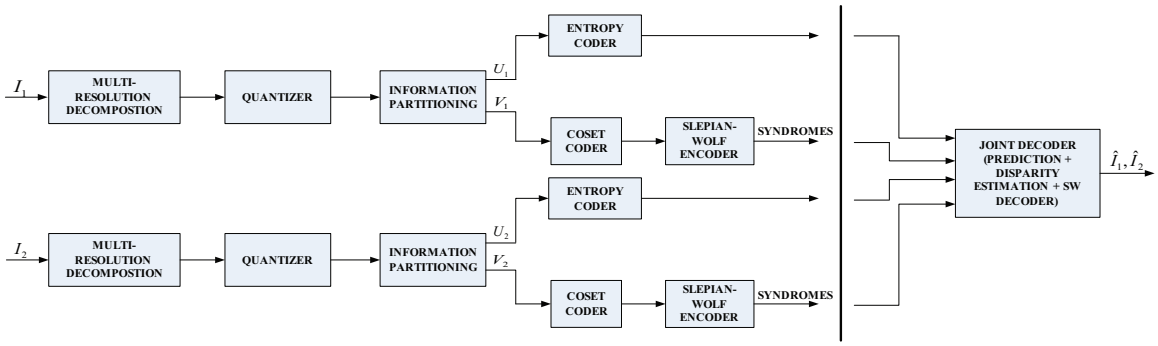


Figure 3.2: Overview of the proposed symmetric distributed coding scheme.

camera parameters are not given or when camera network calibration is not achievable in practice. The proposed scheme therefore provides a low complexity coding solution for the representation of 3D scenes, which neither requires complex setup nor hierarchical organization between vision sensors.

The rest of this chapter is organized as follows. Section 3.2 presents the distributed coding algorithm adapted to omnidirectional images. Section 3.3 presents in more details the Wyner-Ziv coding strategy, while Section 3.4 describes the joint decoding scheme. Section 3.5 finally presents the experimental results that demonstrate the benefits of the proposed solution. Section 3.6 concludes this chapter.

3.2 Distributed coding scheme

3.2.1 Overview

This section presents the overview of the symmetric distributed coding scheme, illustrated in Fig. 3.2. We consider omnidirectional images that can be exactly mapped on the sphere, as those captured by the catadioptric cameras [113]. The captured omnidirectional images I_1 and I_2 undergo multi-resolution decomposition based on Spherical Laplacian Pyramid (SLP), which handles the images directly on the sphere; this properly considers the geometry of the omnidirectional cameras. The computed Spherical Laplacian Pyramid coefficients are then quantized. As we target a balanced rate allocation between cameras, the information of both images is partitioned in a similar way. The quantized coefficients are split into two subsets or partitions: U_1 and V_1 from image I_1, and U_2 and V_2 from image I_2, as shown in Fig. 3.2. These partitions are generally correlated due to the simple partitioning process, which puts quantized coefficients alternatively in both partitions. The first partitions from both images (e.g., partitions U_1 and U_2) are transmitted directly to the joint decoder after entropy coding. The second partitions V_1 and V_2 are coset encoded and the resulting coset indexes are Slepian-Wolf encoded with a multilevel LDPC code. Hence,

each encoder transmits one half of the quantized coefficients, and only the syndrome bits for the quantized coefficients of the second partition.

The joint decoder tries to exploit intra- and inter-view correlation for improved decoding performance. It estimates the quantized coefficients of the partitions V_1 and V_2 by using the coefficients of the partitions U_1 and U_2 respectively. Under the assumption that the images I_1 and I_2 are correlated, the predicted result is further refined using block-based disparity estimation on the sphere. Disparity estimation (DE) permits to compensate for the displacement of the objects captured from different viewpoints. Prediction and disparity estimation together lead to effective side information, which permits to reduce the channel rate of the Slepian-Wolf encoder. The coset indexes corresponding to the partitions V_1 and V_2 are further recovered after correcting the virtual channel noise in the side information using the corresponding syndromes. The SLP coefficients are then estimated from the recovered coset index, and the images are finally reconstructed by inverting the SLP transform.

3.2.2 Spherical Laplacian pyramid

Multi-resolution analysis is an efficient tool that permits to decompose a signal at successive resolutions and perform coarse to fine computations on the data. The two most successful embodiments of this paradigm are the Wavelet decompositions [33] and the Laplacian Pyramid (LP) [114]. As the shift invariance represents an interesting property for distributed coding in camera networks, we have chosen to use the Laplacian Pyramid in our scheme. It proves to be beneficial for predictive coding based on disparity estimation. Furthermore, since we work with omnidirectional images, we propose to use the Laplacian Pyramid on the sphere which is well suited to analyze the spherical data. In the rest of this section, we briefly describe SLP.

Let $L^2(S^2, d\mu)$ denote the Hilbert space of the square integrable signals on the 2D sphere S^2, where $d\mu(\theta, \varphi) = sin\theta \; d\theta \; d\varphi$ represents the rotation invariant Lebesgue measure on S^2. Any spherical signal $F \in L^2(S^2)$ can be expanded using the spherical harmonics $Y_{p,q}$ [115], whose Fourier coefficients are given by

$$\check{F}(p,q) = \int_{S^2} d\mu(\theta,\varphi) Y^*_{p,q}(\theta,\varphi) F(\theta,\varphi) \; , \tag{3.1}$$

where $Y^*_{p,q}$ is the complex conjugate of the spherical harmonic of order (p, q). The omnidirectional images are sampled on the nested equi-angular grids on the sphere described as

$$\mathcal{G}_v = \{(\theta_{vm}, \varphi_{vn}) \in S^2 : \theta_{vm} = \tfrac{(2m+1)\pi}{4W_v}, \varphi_{vn} = \tfrac{n\pi}{W_v}\}, \tag{3.2}$$

$m, n \in \mathcal{N}_v \equiv \{q \in \mathbb{N} : q < 2W_v\}$, for a range of bandwidth $W = \{W_v \in 2\mathbb{N}, v \in \mathbb{Z}\}$. These grids permit to perfectly sample any band-limited function $F \in L^2(S^2)$ of bandwidth W_v, such that $\check{F}(p, q) = 0$ for all $p > W_v$. This class of sampling grids are advantageously associated to a Fast Spherical Fourier Transform [116], which permits rapid transformation of images.

Similar to the classical Laplacian Pyramid decomposition [114], the Spherical Laplacian Pyramid (SLP) proceeds first by low pass filtering the spherical signal in the Fourier domain for speeding up the computations. Let us assume that the original data F is bandlimited, i.e, $\check{F}(p, q) = 0$, $\forall p > W_0$ and sampled on grid $\mathcal{G}_0$. We capture the low frequency information by using a half-band axisymmetric filter $\check{H}$. A spherical function is said to be axisymmetric if it is invariant to rotation with respect to the principal spherical axis. Such a function is independent of q and is represented by $\check{H}(p)$. The bandwidth of the filter $\check{H}(p)$ is chosen such that it is numerically close to a perfect half band filter. The signal F is low pass filtered using $\check{H}(p)$, and the filtered data is then downsampled on the nested sub-grid $\mathcal{G}_1$, which gives the low-pass channel F_1 of our spherical laplacian pyramid. The high-pass channel of the pyramid F_0 is computed as usual, that is by first up-sampling F_1 on the finer grid $\mathcal{G}_0$, low-pass filtering it with $\check{H}$ and taking the difference with F. Coarser resolutions are computed by iterating this algorithm on the low-pass channel F_1. By repeating this step k times, the original spherical signal F can be decomposed into the LL subband F_k and detailed subbands $F_{k-1}, F_{k-2}, \ldots, F_0$. The parameter k denotes the level of decomposition used to process the image

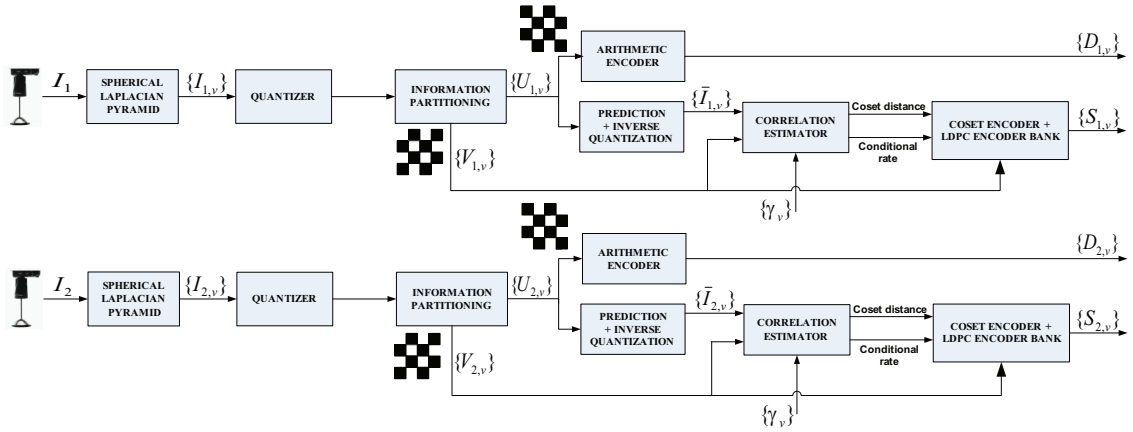


Figure 3.3: Detailed block scheme of the Wyner-Ziv encoder with correlation estimation.

F.

The coefficients of the SLP need to be quantized with an efficient rate distribution among the subbands. We follow the algorithm proposed for the Laplacian Pyramid in [117]. The rate allocation can be computed by Lagrange's multipliers method when the quantizers are uniform. Unsurprisingly, the rate in the different subbands is chosen to be proportional to the variance of the coefficients.

3.3 Slepian-Wolf encoder

We describe now in more details the proposed Wyner-Ziv encoding scheme illustrated in Fig. 3.3. The omnidirectional images I_1 and I_2 undergo multiresolution decomposition based on SLP using k decomposition levels. The generated subband coefficients are represented by $\{I_{j,v}\}$, $\forall j = \{1, 2\}$, $v = \{0, \ldots, k\}$, where j represents the image (view) index and k represents the corresponding subband index. For example, $I_{j,v}$ represents the set of coefficients in the v^{th} subband of the image I_j. The generated subband coefficients are then quantized uniformly with optimal rate allocation among the subbands, as described above.

Coefficient partitioning

The quantized coefficients of each image are then distributed alternatively into two correlated partitions that form a kind of a checkerboard pattern. Let (m, n) denote the position indexes of a point $(\theta_{vm}, \varphi_{vn})$ on the equi-angular spherical grid $\mathcal{G}_v$, as defined in Eq. (3.2). The quantized coefficient $I_{j,v}$ at a point $(\theta_{vm}, \varphi_{vn})$ on the spherical grid is put in the partition $U_{j,v}$ if $((m \mod 2)\ XOR\ (n \mod 2)) = 0$. Otherwise it is put in the partition $V_{j,v}$. For illustration, Fig. 3.4 shows the partitioning strategy used in our scheme for the subband $I_{j,v}$ of the bandwidth $W_v = 2$ (this subband is of the size 4×4). It is clear that the coefficients at positions marked in white and black color belong to the partitions $U_{j,v}$ and $V_{j,v}$ respectively. Finally, the quantized coefficients from each subband $v = \{0, \ldots, k\}$ of the pyramid are split into two partitions $U_{j,v}$ and $V_{j,v}$, following the same partitioning strategy.

The quantized coefficients in the partitions $\{U_{j,v}\}$, $\forall j = \{1, 2\}$, $v = \{0, \ldots, k\}$ are compressed using the Arithmetic encoder and the compressed bits $\{D_{j,v}\}$ are transmitted directly to the joint decoder. The partitions $\{V_{j,v}\}$ however further undergo coset and Slepian-Wolf coding to save bit rate in the distributed coding scheme. The quantized coefficients in partition $V_{j,v}$ are put in different cosets. The cosets group coefficients from $V_{j,v}$ that are separated by a distance $d_{j,v}$. Only the coset indexes are eventually encoded, which provides some significant rate savings. The coset distance is estimated as $d_{j,v} = 2^{\lceil \log_2(2E_{j,v}+1) \rceil}$, where

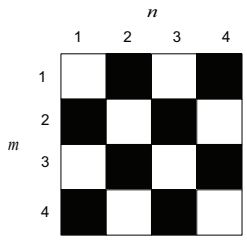


Figure 3.4: Checkerboard partition strategy for the subband $I_{j,v}$ of the bandwidth $W_v = 2$. The top most left quantized coefficient is indexed by $m = 1, n = 1$. The partitions $U_{j,v}$ and $V_{j,v}$ are marked by the white and black color respectively.

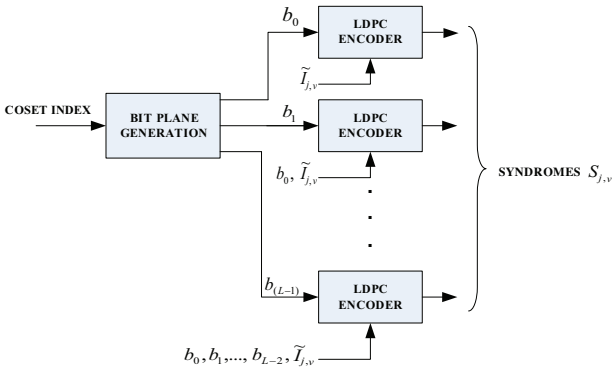


Figure 3.5: Block scheme of the LDPC encoder bank, where the bitplanes are encoded starting from b_0. Each bit plane b_l $(0 \leq l \leq L-1)$ is encoded based on the previously encoded $l-1$ bitplanes and the side information $\tilde{I}_{j,v}$.

$E_{j,v}$ is the maximum error between the original and the side information images in the v^{th} subband. Scalar quantization and coset encoding together behave similarly to nested scalar quantization.

Multilevel LDPC coding

Even after coset encoding, some correlation still exists between the coefficients and the side information available at the decoder. We propose to achieve further compression by encoding the coset indexes with multilevel LDPC codes [118]. In other words, instead of sending the coset indexes to the decoder, the encoder only transmits the syndrome bits $\{S_{j,v}\}$ resulting from the LDPC encoding. We propose to use irregular LDPC codes and we follow the procedure described in [111, 119] in order to construct the parity check matrix. We describe now in more details the multilevel LDPC coding, and the channel rate estimation.

For the image I_j, the coset indexes generated from each subband v are first decomposed into L bit planes $b_0, b_1, \ldots, b_{L-1}$, where b_0 represents the most significant bit plane (MSB) and b_{L-1} represents the least significant bit plane (LSB). Each bitplane b_l $(0 \leq l \leq L-1)$ is encoded by the LDPC encoder, starting from the bit plane b_0, as illustrated in Fig. 3.5. The LDPC encoding rate is chosen by assuming that the error between the SLP subbands and the corresponding side information follows a Laplacian distribution. The subsequent bitplanes are also encoded with LDPC codes, where the code rate is however adapted based on the previously encoded bitplanes.

In more details, the coding rate for encoding the l^{th} bitplane of the subband $I_{j,v}$ is estimated as follows.

First, the conditional probability $P(i) = Pr(b_l(i) = 1|\tilde{I}_{j,v}(i), b_0(i), b_1(i), \ldots, b_{l-1}(i))$ is calculated for each bit i, where $\tilde{I}_{j,v}$ denotes the side information for decoding the subband $I_{j,v}$. The rate of the LDPC encoder is then chosen to be equal to the following conditional entropy,

$$H(b_l|\tilde{I}_{j,v}, b_0, b_1, \ldots, b_{l-1}) \simeq \frac{1}{M}\sum_{i=1}^{M} H(P(i)), \tag{3.3}$$

where M denotes the number of bits in the bit plane [120]. Unfortunately, the side information is not available at the encoder, and the conditional probability P generally has to be estimated as described below.

Channel rate estimation

One of the main difficulties in distributed coding is the estimation of the correlation between sources, or equivalently the construction of noise models at the encoder for the proper design of the Slepian-Wolf encoder. Our encoder has to estimate the noise distribution in order to determine the coset distance and the LPDC coding rate. Unfortunately, the side information that is used for joint decoding is only present at the decoder, and the encoder can only predict the noise distribution.

We assume that the error $E_{j,v}$ between the SLP subband $I_{j,v}$ and the corresponding side information subband $\tilde{I}_{j,v}$ follows a Laplacian distribution, of the form $f_{E_{j,v}}(e) = \frac{1}{2\lambda_{j,v}}\ exp(-\frac{|e|}{\lambda_{j,v}})$. The Laplacian distribution is a common assumption in such a scenario, and it provides a good approximation of the actual distribution of the error. In this case, the rate $R_{j,v}$ necessary to code the error $E_{j,v}$ is equivalent to the conditional entropy $H(I_{j,v}|\tilde{I}_{j,v})$. When the quantization is uniform with step size $\delta_{j,v}$ for the subband $I_{j,v}$, the rate depends only on the variance of the Laplacian distribution [121]. It can be written as

$$R_{j,v} = H(I_{j,v}|\tilde{I}_{j,v}) = \alpha_v\ \log_2(\frac{\lambda_{j,v}}{\delta_{j,v}}) + \beta_v, \tag{3.4}$$

where α_v and β_v are constants that can be estimated offline on test image sets, and therefore are not dependent of the image I_j. The construction of the noise model for a proper choice of the coding parameters therefore consists in estimating the parameters $\{\lambda_{j,v}\}$ of the Laplacian distribution, for all $j = \{1, 2\}$ and $v = \{0, \ldots, k\}$.

In our scheme, the side information is actually built on prediction and disparity estimation steps. We can thus model separately the effect of spatial prediction of the coefficients and the benefit of disparity estimation. In the first case, the encoder can estimate the rate $R'_{j,v}$ that is necessary to correct the error due to spatial prediction of the coefficients. The encoder can implement the coefficient prediction step, since it does not depend on the information from the other sensors. The residual error between the coefficients in the subband $I_{j,v}$ and the corresponding subband computed by coefficient prediction $\bar{I}_{j,v}$ can then be modeled with a Laplacian distribution. The parameter of the distribution $\lambda'_{j,v}$ is finally estimated from the prediction error. The rate $R'_{j,v}$ needed to code the prediction error can also be computed by the conditional entropy as

$$R'_{j,v} = H(I_{j,v}|\bar{I}_{j,v}) = \alpha_v\ \log_2(\frac{\lambda'_{j,v}}{\delta_{j,v}}) + \beta_v. \tag{3.5}$$

However, the side information is not only built on coefficient prediction; disparity estimation is used at the decoder in order to exploit the correlation between images from different sensors. We propose to compute a conservative approximation of the gain due to disparity estimation, expressed as $\gamma_v = R'_{j,v}/R_{j,v}$. It can be computed by offline encodings of several test images, where the complete side information is made available at the encoder. The offline estimation of γ_v finally permits to estimate at the encoders the parameter of the

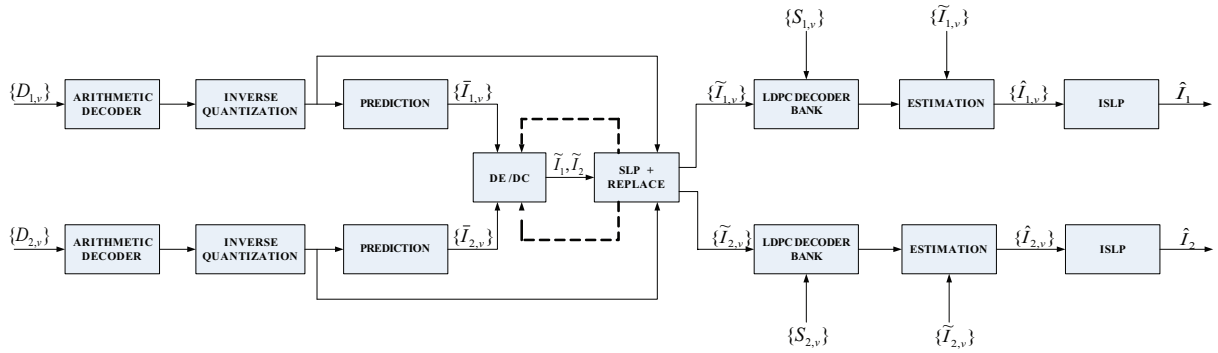


Figure 3.6: Detailed block scheme of the Wyner-Ziv decoder.

complete noise model. The parameter $\lambda_{j,v}$ of the Laplacian distribution can be expressed as

$$\lambda_{j,v} = \lambda'_{j,v}\ 2^{(-\frac{R'_{j,v}(1-\frac{1}{\gamma_v})}{\alpha_v})}, \tag{3.6}$$

by combinations of Eq. (3.4) and Eq. (3.5). As we have now an approximation of the distribution of the error induced by the side information, we can estimate the side information for subband $\tilde{I}_{j,v}$ at the encoder. It permits to estimate the error $E_{j,v}$ and hence the coset distance $d_{j,v}$ for coding the quantized coefficients. Finally, we can estimate the LDPC coding rate by computing the probability $P(i)$, and the conditional entropy given in Eq. (3.3). The complete encoding scheme is illustrated in Fig. 3.3.

3.4 Joint decoding

3.4.1 Overview

The joint decoder exploits the correlation between images in order to reconstruct the views of the 3D scenes. The decoding scheme is illustrated in Fig. 3.6. The quantized coefficients from the partitions $\{U_{j,v}\}$, $\forall j = \{1,2\}$, $v = \{0,\ldots,k\}$, are easily recovered by arithmetic decoding and inverse quantization. The quantized coefficients of the partitions $\{V_{j,v}\}$, $\forall j = \{1,2\}$, $v = \{0,\ldots,k\}$, however, have to be reconstructed by Slepian-Wolf decoding. These missing quantized coefficients are first predicted in each subband $V_{j,v}$, with the help of the coefficients from the corresponding partition $U_{j,v}$. They are simply predicted by interpolation from the four nearest neighbors in the partition $U_{j,v}$. This simple spatial prediction exploits the correlation among neighbor coefficients and the v^{th} predicted subband is denoted by $\bar{I}_{j,v}$. The side information is then built by refining the value of the predicted coefficients by disparity compensation between the approximations of the different images. The decoder implements disparity estimation on the sphere to exploit the redundancy between the omnidirectional images from different cameras. Next, the coset indexes that correspond to the coefficients in partitions $\{V_{j,v}\}$ are recovered by using the syndrome bits of the LDPC code, as well as the side information created by prediction and disparity estimation. Finally, the SLP coefficients are recovered by coset decoding with the help of side information, and the images are reconstructed after inverting the SLP. The main steps of the joint decoder algorithm are detailed in the rest of this section.

3.4.2 Side information generation

The subbands built on the spatial prediction of the missing coefficients are refined by the multi-resolution block-based disparity estimation on the sphere. We summarize here the key ideas of the disparity estimation

on the sphere, while more details are available in [39]. The low resolution subband $\bar{I}_{1,k}$ is divided into blocks of uniform solid angles. For each block in $\bar{I}_{1,k}$, a best matching block in the mean square sense is selected in $\bar{I}_{2,k}$. The displacement between the corresponding blocks represents the disparity vector. The generated disparity vectors from the lower resolution are upsampled and are used as the initial estimate for the higher resolution. The initial estimate is further refined using the subband coefficients at higher resolution. This process is iterated up to the finest resolution and eventually outputs the disparity vectors. The resulting disparity vectors are then used for constructing an estimate $\tilde{I}_1$ of the image I_1 that serves as side information for decoding the image I_1. In particular, the side information image is first constructed by applying disparity compensation from the image $\bar{I}_2$. It then undergoes a SLP decomposition, similar to the transform implemented at the encoder. The coefficients corresponding to the partition $\{U_{1,v}\}$ for $v = \{0, \ldots, k\}$ are then substituted by the coefficients that have been correctly received from the encoder, in order to reduce the estimation error. The same process is implemented to generate the side information image $\tilde{I}_2$ with disparity compensation based on the predicted image $\bar{I}_1$.

The exploitation of the correlation between images by disparity estimation permits to refine the values of the predicted result from the partition $V_{j,v}$. The disparity estimation process can be repeated on the images $\tilde{I}_1$ and $\tilde{I}_2$ in order to further improve the image approximation. We have observed empirically that it is advantageous to repeat the disparity compensation a second time. This step is represented with a dashed line on the block scheme in Fig. 3.6. Further iterations however, do not improve significantly the side information. The resulting side information subbands $\{\tilde{I}_{j,v}\}, \forall j = \{1, 2\}, v = \{0, \ldots, k\}$ form a side information that are used for decoding the coset indexes.

3.4.3 Coefficient decoding

The coefficients from partition $\{V_{j,v}\}$, $\forall j = \{1, 2\}$ and $v = \{0, \ldots, k\}$, are recovered by Slepian-Wolf decoding. The side information $\tilde{I}_{j,v}$ is used by the LDPC decoder together with the syndromes bits $S_{j,v}$ to decode the coset indexes in each subband $I_{j,v}$. A LDPC decoder bank uses L LDPC decoders to decode each bit plane successively, starting from the MSB bitplane. While decoding the bitplane b_l $(0 \leq l \leq L-1)$, the previously decoded $l-1$ bitplanes $b_0, b_1, \ldots, b_{l-1}$ are used as side information by the LDPC decoder. LDPC decoding is implemented with a Belief propagation algorithm, where the confidence level is initialized at the variable node using the following log likelihood ratio (LLR),

$$LLR = \log\left(\frac{P(b_l = 0|\tilde{I}_{j,v}, b_0, b_1, \ldots, b_{l-1})}{P(b_l = 1|\tilde{I}_{j,v}, b_0, b_1, \ldots, b_{l-1})}\right). \tag{3.7}$$

The coset indexes of each v^{th} subband $I_{j,v}$ are reconstructed when all the bit planes are decoded. The coefficients in each subband are finally computed by decoding the coset indexes. The decoded coefficient corresponds to the coefficient in the coset that is one closest to the side information $\tilde{I}_{j,v}$. Once all the subband coefficients are decoded, the image is reconstructed by inverting the SLP transform.

3.5 Experimental results

3.5.1 Setup

We evaluate the performance of our system on both synthetic and natural spherical images. Synthetic spherical image set *Room* is shown in Fig. 3.7 and the natural spherical image set *Lab* is shown in Fig. 3.8.

The SLP is implemented using an axisymmetric filter $\check{H}(p)$ constructed from a 7-tap digital filter $h(s) = \{-0.0625 \quad 0 \quad 0.5625 \quad 1 \quad 0.5625 \quad 0 \quad -0.0625\}$. The filter $\check{H}(p)$ is constructed by computing the Fourier transform of $h(s)$ and replicating it for each q such that $\check{H}(p, 1) = \check{H}(p, q), \forall q$. Obviously such a construction results in an axisymmetric filter that is independent of the variable q, and is completely determined by $\check{H}(p)$.

(a) I_1 (b) I_2

Figure 3.7: Room spherical dataset: (a) original left view I_1; (b) original right view I_2.

(a) I_1 (b) I_2

Figure 3.8: Lab spherical dataset: (a) original left view I_1; (b) original right view I_2.

The SLP decomposition is further carried out in the Fourier domain in order to speed up the computations. The SLP is implemented with four levels of decomposition ($k = 4$) in the results presented below. The multi-resolution block-based disparity estimation at the decoder is carried out on blocks of size 4×4. Finally, the performance is measured in terms of PSNR, where the mean square error is evaluated using the inner product on the sphere.

3.5.2 Channel model evaluation

Before analyzing the performance of the distributed coding scheme, we propose to evaluate the channel model that is used for designing the Slepian-Wolf encoder. We first show in Fig. 3.9, the distribution of the error between the subband $I_{2,3}$ and the corresponding side information subband $\tilde{I}_{2,3}$ in the Room image dataset. The error is computed only on the coefficients of the partition $V_{2,3}$. We can see that the error follows zero mean Laplacian distribution with $\lambda_{2,3} = 0.0178$, as expected.

Then, we estimate the constants α_v and β_v that are used to compute the conditional entropy in Eq. (3.4) for both Room and Lab image sets. These constants are the same for all image sets and differ only in the respective subbands. We have obtained $\alpha_4 = 1.04$ and $\beta_4 = 2.44$ for the LL subband and $\alpha_v = 0.54$ and $\beta_v = 1.92$ $(0 \leq v \leq 3)$ for the detail subband.

Finally, we evaluate the constant γ_j for all the subbands $(0 \leq j \leq 4)$ that captures the benefit of the

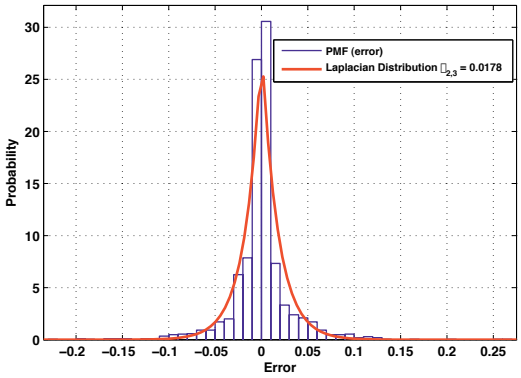


Figure 3.9: Distribution of the error $E_{2,3}$ computed on the partition $V_{2,3}$ between the detail subbands $I_{2,3}$ and $\tilde{I}_{2,3}$ of the Room dataset. The fitting curve shows a zero-mean Laplacian distribution with $\lambda_{2,3} = 0.0178$.

disparity estimation. We have obtained a value of $\gamma_4 = 1.25$ for the LL subband in both image sets. For the detail subbands the parameter γ_j is found to be 1.6, 1.4, 1.2 and 1.1 starting from the lowest to the highest resolution subbands, respectively. The value of γ_j is decreasing when the resolution increases, since the disparity estimation is mostly efficient in capturing the correlation in the low frequency subbands.

3.5.3 Coding performance

We first compare the performance of the proposed DSC solution (using estimated correlation model) with an independent coding scheme and a joint encoding scheme. In the independent coding scheme, the images I_1 and I_2 are encoded independently using a SLP-based strategy. The images I_1 and I_2 are transformed using four SLP decomposition levels. Compression is achieved by first quantizing the coefficients [117] and further the quantized coefficients are entropy coded (i.e., arithmetic coding), similarly to the coefficients of the partition $U_{j,v}$ in the distributed coding scheme. Next, the joint encoding scheme is based on disparity compensated predictive coding. One image is selected as the reference and it is encoded independently, whereas the other image is predicted from the reference image. In our scheme, the image I_2 is selected as the reference image and I_1 is predicted from I_2. The reference image I_2 is encoded using four SLP decomposition levels. Multi-resolution disparity estimation that is used to predict the image I_1 is performed with blocks of size 8×8. The residual error after disparity estimation is also encoded using a SLP based strategy. The disparity vectors of the successive resolution levels are differentially encoded. Finally, the rate allocation between the reference and the predicted images is chosen such that the rate-distortion performance is maximized. The corresponding rate distributions are given in Tables 3.1 and 3.2, where the bits used for the disparity vectors are included in the budget of the predicted frames.

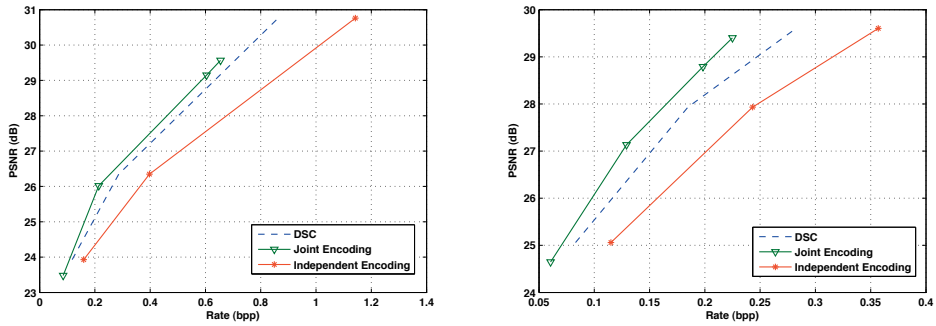
The comparison between the distributed, independent and joint coding schemes are given in Fig. 3.10 in terms of rate-distortion performance. We observe that the distributed coding scheme performs close to joint encoding algorithm that is based on the same representation and coding strategy. We also see that our proposed DSC scheme clearly outperforms independent coding scheme. In particular, the gain reaches 1.5 dB for the Room image set and 1.3 dB for the Lab image set. We further compute the rate savings between DSC and independent coding schemes for the same reconstruction quality. Tables 3.3 and 3.4 tabulate the percentage of rate saving at different reconstructed qualities for the Room and Lab images respectively. We could see that bit saving is approximately 25%, for both image sets. The reconstructed Room image $\hat{I}_1$ is finally shown in Fig. 3.11 for two sample bit rates. The reconstructed images are shown as planar images in the $(\theta,\ \varphi)$ coordinates to show all parts of the spherical images.

Table 3.1: Distribution of bits between the reference image and the predicted image in joint encoding of the Room dataset.

Reference Image (I_2)		Predicted Image (I_1)		Total Rate (bpp)	Mean PSNR (dB)
Bits	PSNR (dB)	Bits	PSNR (dB)		
5293	23.8	297	23.19	0.0853	23.47
13161	26.11	864	25.93	0.214	26
37694	30.5	1887	28.12	0.604	29.15
37694	30.5	5197	28.8	0.6545	29.57

Table 3.2: Distribution of bits between the reference image and the predicted image in joint encoding of the Lab dataset.

Reference Image (I_2)		Predicted Image (I_1)		Total Rate (bpp)	Mean PSNR (dB)
Bits	PSNR (dB)	Bits	PSNR (dB)		
3742	25.26	219	24.11	0.0604	24.64
7759	28.06	707	26.37	0.1292	27.13
11469	29.73	1528	28.02	0.1983	28.79
11469	29.73	3274	29.1	0.2250	29.40



(a) (b)

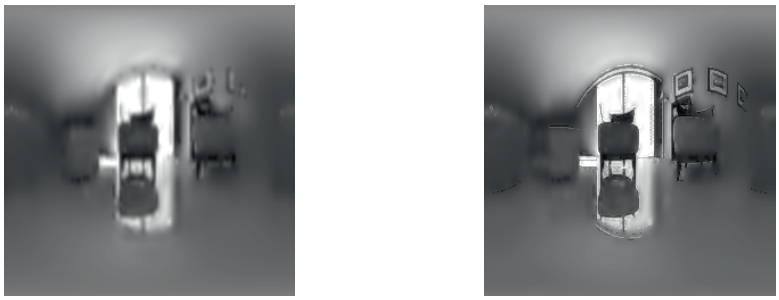
Figure 3.10: Rate-distortion performance comparison between the proposed DSC scheme, joint and independent coding strategies for (a) Room dataset and (b) Lab dataset.

Table 3.3: Bit rate savings for the Room Image I_2.

PSNR (dB)	Bits		Bits saved	% Bit saving
	DSC	Independent		
23.8	3928	5293	1365	25.8
26.1	9526	13161	3635	27.6
30.5	28166	37694	9528	25.3

Table 3.4: Bit rate savings for the Lab Image I_1.

PSNR (dB)	Bits		Bits saved	% Bit saving
	DSC	Independent		
25.26	2852	3797	945	24.9
28.06	6214	8179	1965	24
29.73	9542	11920	2378	20



(a) Rate: 0.14 bpp, PSNR: 26.1 dB (b) Rate: 0.43 bpp, PSNR: 30.5 dB

Figure 3.11: Reconstructed image $\hat{I}_1$ in the Room scene. The images are represented as planar images in (θ, φ) coordinates.

Finally, we compare in Fig. 3.12 the average performance of the distributed coding scheme with independent coding implemented by the JPEG compression standard, for the Room image set. The equiangular grid of the spherical image is represented as a 2D planar image. A baseline JPEG coding scheme[2] is used to encode the unwrapped images I_1 and I_2 independently. We can see that both independent and distributed coding schemes based on the SLP decomposition outperform JPEG at low coding rates, thanks to efficient data processing on the sphere. At higher rates, the mode of representation of the information becomes less critical, and JPEG provides improved performance. The degradation of the RD performance of our scheme with respect to JPEG at higher rates could be explained by the use of a simple encoding scheme of the overcomplete Spherical Laplacian Pyramid, which is based on adaptive quantization. Employing more efficient coding methods for LP, like the one proposed in [122], could result in improved RD performance of the proposed scheme, also at higher rates.

3.5.4 DSC scheme analysis

We finally analyze in more details the behavior of the distributed coding scheme. We first examine the rate balance between the two encoders, by comparing the RD performance of the images I_1 and I_2. Fig. 3.13 shows the RD curves for the test images. As expected, the DSC scheme balances the encoding rates, since the encoding rates between the images I_1 and I_2 are quite similar at a given reconstruction quality.

Next, we study the effect of imprecise estimation of the coding rate in the Slepian-Wolf encoder. For

[2]implemented using VcDemo toolbox available at: http://siplab.tudelft.nl/content/image-and-video-compression-learning-tool-vcdemo

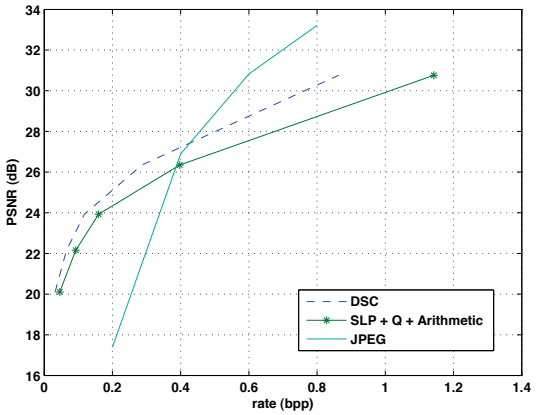
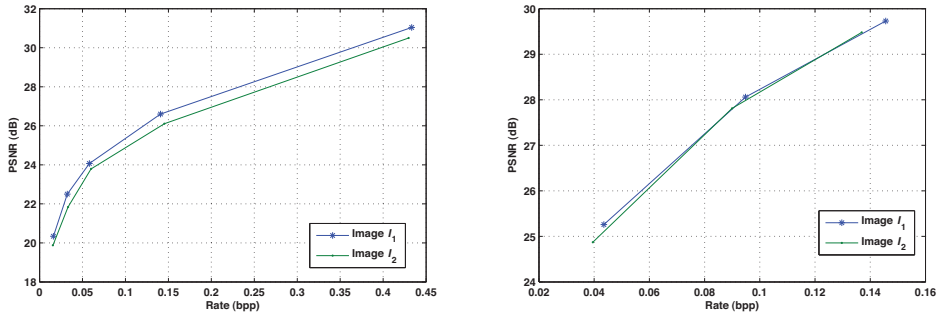


Figure 3.12: Average rate-distortion performance for encoding the Room image set using the distributed coding scheme and independent coding with JPEG.

both image sets, we compare the rate-distortion characteristics between the estimated correlation model described in this paper, and an exact oracle model where the Laplacian distribution parameters $\{\lambda_{j,v}\}$ are known a priori at the encoder for all v, j. The comparison is presented in Fig. 3.14 for the image I_1 of the Room and Lab datasets. We can see that the methodology proposed in this paper for estimating the channel rate performs very similar to the exact model. The performance degradation due to inexact rate estimation stays smaller than 0.2 dB.

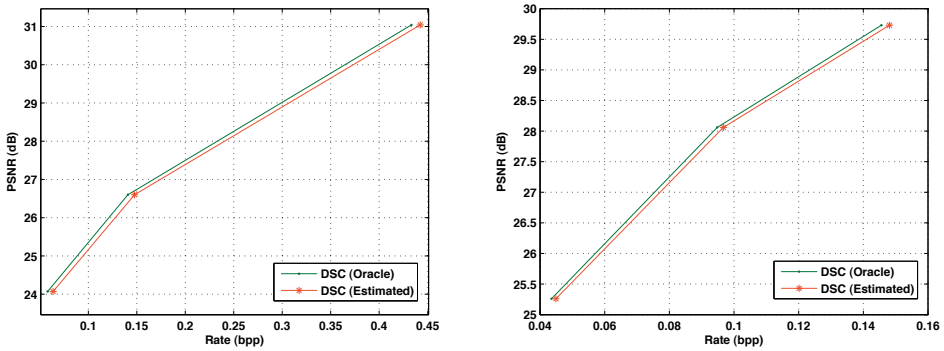


(a) (b)

Figure 3.13: Rate-distortion comparison between the images I_1 and I_2 to examine the rate balance among the encoders: (a) Room dataset; (b) Lab dataset.

3.6 Conclusions

In this chapter we have contributed a rate balanced distributed source coding scheme for the effective representation of 3D scenes captured by stereo omnidirectional cameras. The proposed coding framework processes the visual information on the 2D sphere in order to preserve the intrinsic geometry information of



(a) (b)

Figure 3.14: Rate-distortion comparison between estimated correlation model vs exact correlation model: (a) Room image I_1; (b) Lab image I_1.

the catadioptric imaging systems. The correlated spherical images are decomposed by a Spherical Laplacian Pyramid and coefficients are partitioned and Slepian-Wolf encoded at independent encoders. The joint decoder efficiently exploits both the intra- and inter-view correlation by coefficient prediction and disparity estimation respectively. Our scheme outperforms independent coding and performs close to a joint encoding solution based on similar coding principles. The channel rate parameters has been estimated at the encoder based on a conservative estimation on sets of images; it however shows only a negligible degradation in rate-distortion performance with respect to the oracle based scheme. For a given target bit rate, we have shown experimentally that the image reconstruction qualities are similar between the images. Finally, it should be noted that the ideas presented in this chapter can be adapted for building a rate balanced distributed coding scheme for compressing the stereo planar images; it could be one of the interesting future perspectives of this thesis. In the next chapter, we omit the Slepian-Wolf encoder and propose a joint representation scheme that processes directly the compressed images.

Chapter 4

Distributed Joint Representation from Compressed Images

4.1 Introduction

In the previous chapter, we have proposed a rate balanced distributed coding based on Wyner-Ziv results, where the images are compressed using a multi-resolution partition and Slepian-Wolf encoding. Also, we have proposed a methodology to estimate the channel rate parameters at the encoder. The encoding rate is controlled using an additional rate control module. The overall computational complexity at the encoder becomes important due to this Slepian-Wolf rate control module. One solution to reduce the encoding complexity is to use a feedback channel in order to precisely control the Slepian-Wolf coding rate [79, 82]. This however results in high latency and bandwidth usage due to the multiple requests from the joint decoder, and therefore can hardly be used in real time applications. Hence, we propose to omit the Slepian-Wolf encoder and we present now a rate balanced distributed coding scheme, where the compressed images are directly transmitted to the joint decoder. Due to the absence of Slepian-Wolf encoding, the framework presented in this chapter neither requires an encoder rate control module nor a feedback channel; it therefore reduces the complexity at the encoder. In this context, Schenkel *et al.* [123] have proposed a distributed joint representation of image pairs from the JPEG compressed images. However, this scheme is built on asymmetric frameworks, where one image is considered as the reference for decoding the second image.

In this chapter, we rather propose a symmetric distributed joint scene representation for a pair of correlated images captured in perspective or omnidirectional camera networks. We consider a scenario, where the captured images are compressed independently using standard encoding solutions (e.g., SPIHT) and are transmitted to a central decoder (see Fig. 4.1). The central decoder jointly processes the compressed images and reconstructs an image pair by exploiting the correlation between images. The joint reconstruction is cast as a constrained optimization problem that reconstructs an image pair that satisfies the estimated correlation model. At the same time, we add constraints that force the reconstructed images to be as close as possible to the compressed views. We show by experiments that the proposed joint reconstruction scheme outperforms independent reconstruction in terms of image quality for both planar and spherical datasets. We further observe that the joint reconstruction is effective only at medium bit rates and not at low bit rates due to the significant distortion in the disparity image or the correlation information. In particular, we observe that the disparity values are not precisely estimated in the low rate regime.

In the literature, the effect of compression distortion introduced due to the encoding of images on the quality of disparity estimation has not been studied yet. On the other hand, most of the research works have focussed on studying the effect of geometric distortions in the compressed disparity image on the image prediction quality, e.g., [124, 125, 126, 127]. However, these works fail to consider the effect of image

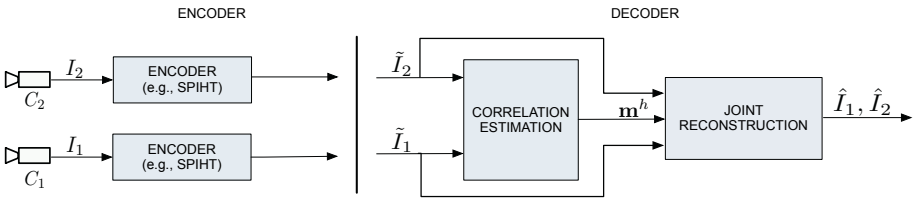


Figure 4.1: Schematic representation of the proposed scheme. The images I_1 and I_2 are correlated through displacement of scene objects due to viewpoint change. The cameras C_1 and C_2 can be either perspective or omnidirectional imaging sensors.

compression on the disparity estimation. In the second part of this chapter, we therefore consider the problem of estimating an accurate disparity image from the low bit rate compressed images. We propose here a rate allocation scheme based on a modified SPIHT algorithm that gives importance to the reconstruction of the visual information in the low contrast regions, and at the same time it controls the penalty of reconstruction in other regions, e.g., texture and strong edges. Such a rate allocation scheme leads to an accurate disparity estimation compared to the one that is estimated from the traditional coding schemes based on JPEG 2000, SPIHT or JPEG. We show experimentally that at low bit rates, the joint reconstruction is effective in improving the quality of the compressed images due to accurate disparity estimation. Finally, we compare the joint reconstruction performance between the distributed coding schemes implemented with SPIHT and modified SPIHT encoding principles. Despite improved correlation estimation with the modified SPIHT coding scheme, we observe that the image reconstruction quality is significantly penalized compared to the one obtained with SPIHT-based joint decoding schemes. This is due to the encoding of different image characteristics for optimizing the image reconstruction and disparity estimation. Therefore, optimizing both the image reconstruction and the disparity estimation are not possible under fixed bit budget.

In summary, the contributions in this chapter are: (1) proposed a distributed joint representation algorithm that exploits the correlation between compressed images captured in perspective or catadioptric cameras; (2) proposed a rate allocation methodology that allows to estimate an accurate disparity model from the compressed images. The rest of this chapter is organized as follows. In Section 4.2, we describe our distributed coding framework and the novel joint reconstruction algorithm based on an optimization framework. Section 4.3 reports the distributed coding performance for both planar and omnidirectional images. Section 4.4 describes the proposed rate allocation methodology and its performances are analyzed in details in Section 4.5. Section 4.6 concludes this chapter.

4.2 Joint reconstruction of compressed images

4.2.1 Framework

In this section, we give an overview of our distributed coding framework. We consider the scenario shown in Fig. 4.1, where a pair of cameras C_1 and C_2 sample a 3D scene in different viewpoints; the cameras can be either perspective or catadioptric sensors. Without loss of generality, we assume that the images I_1 and I_2 (with resolution $N = N_1 \times N_2$) are rectified, so that correlation between images is effectively described by a disparity field. However, our framework can also be extended to estimate a motion field between the time varying images captured by a video camera. The captured images I_1 and I_2 are compressed independently with b bits per view using transform-based coding solutions (e.g., SPIHT [128]). Balanced rate allocation allows us to share the transmission and computational cost equally among the sensors, and thus advantageously avoids any hierarchical relationship among the sensors. The compressed information is transmitted to a central decoder that exploits the underlying correlation between views for improved

decoding quality. In particular, as shown in Fig. 4.1, the joint decoder estimates correlation between images in terms of dense disparity image $\mathbf{m}^h$ from the compressed images $\tilde{I}_1$ and $\tilde{I}_2$. A joint reconstruction stage eventually uses the correlation estimation $\mathbf{m}^h$ and enhances the quality of the independently compressed views $\tilde{I}_1$ and $\tilde{I}_2$. The main steps of the joint decoding algorithm are detailed in the rest of this section.

4.2.2 Disparity models

The first task is to estimate the correlation between images which typically consists in a pixel-wise disparity image. The disparity models are estimated by matching the corresponding pixels between images (see Eq. (2.11)). Several algorithms have been proposed in the literature to compute dense disparity images. For more details, we refer the reader to [7]. In this work, we estimate a dense disparity image $\mathbf{m}^h$ from the compressed images in a regularized energy minimization framework, where the energy model is given as

$$E(\mathbf{m}^h) = E_d(\mathbf{m}^h) + \lambda E_s(\mathbf{m}^h). \tag{4.1}$$

$E_d(\mathbf{m}^h)$ and $E_s(\mathbf{m}^h)$ represent the data and smoothness terms respectively, and λ balances these two terms. The data cost is used to match the pixels across views by assuming that the 3D scene surfaces are Lambertian, i.e., the intensity is consistent irrespective of the viewpoints. In this work, we match the pixels across views using a sampling insensitive Birchfield-Tomasi (BT) cost function [129]. We first measure the disagreement of assigning the real range of disparities $(\mathbf{m}^h(\mathbf{z}) - 1/2, \mathbf{m}^h(\mathbf{z}) + 1/2)$ to the pixel $\mathbf{z} = (m, n)$ by

$$\mathcal{C}_{fwd}(\mathbf{z}, \mathbf{m^h}(\mathbf{z})) = \min_{\mathbf{m}^h(\mathbf{z})-1/2 \leq x \leq \mathbf{m}^h(\mathbf{z})+1/2} |\tilde{I}_1(m, n) - \tilde{I}_2(m + x, n)|. \tag{4.2}$$

For symmetry, we also measure

$$\mathcal{C}_{rev}(\mathbf{z}, \mathbf{m^h}(\mathbf{z})) = \min_{m-1/2 \leq x \leq m+1/2} |\tilde{I}_1(x, n) - \tilde{I}_2(m + \mathbf{m}^h(\mathbf{z}), n)|. \tag{4.3}$$

Then, the cost of assigning the disparity $\mathbf{m}^h$ to the pixel $\mathbf{z} = (m, n)$ is measured by

$$\mathcal{C}(\mathbf{z}, \mathbf{m}^h(\mathbf{z})) = min\{\mathcal{C}_{fwd}(\mathbf{z}, \mathbf{m^h}(\mathbf{z})), \mathcal{C}_{rev}(\mathbf{z}, \mathbf{m^h}(\mathbf{z}), \mathbf{z})\}. \tag{4.4}$$

Finally, the data cost $E_d(\mathbf{m}^h)$ is computed as the cumulative sum of the cost $\mathcal{C}(\mathbf{z}, \mathbf{m}^h(\mathbf{z})), \forall \mathbf{z}$, i.e.,

$$E_d(\mathbf{m}^h) = \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} \mathcal{C}(\mathbf{z}, \mathbf{m}^h(\mathbf{z})), \tag{4.5}$$

where N_1 and N_2 represent the image dimensions.

The smoothness term E_s is used to enforce consistent disparity among the neighboring pixels $\mathbf{z}$ and $\mathbf{z}'$. It is measured as

$$E_s(\mathbf{m}^h) = \sum_{\mathbf{z}, \mathbf{z}' \in \mathcal{N}} V_{\mathbf{z}, \mathbf{z}'}, \tag{4.6}$$

where $\mathcal{N}$ represents the usual four-pixel neighborhood. The term $V(\mathbf{z}, \mathbf{z}')$ is given as

$$V(\mathbf{z}, \mathbf{z}') = min(|\mathbf{m}^h(\mathbf{z}) - \mathbf{m}^h(\mathbf{z}')|, \tau), \tag{4.7}$$

where τ sets an upper level on the smoothness penalty that helps preserving the discontinuities [130]. Several minimization algorithms exist in the literature for solving Eq. (4.1), and they are naturally formulated in the Bayesian framework using MRF, e.g., Simulated annealing [131], Belief Propagation [132], Graph Cuts [133, 134]. Among them, optimization techniques based on Graph Cuts compute the minimum E in polynomial time and they generally give better results compared to the other techniques [7].

Finally, it should be noted that the disparity between spherical images can also be computed by minimizing the energy given in Eq. (4.1). In this case, the disparity is computed as differences between the spherical coordinates of the corresponding pixels, after mapping the omnidirectional images on the sphere. For more details, we refer the reader to [112].

4.2.3 Joint reconstruction

Our novel joint reconstruction algorithm takes benefit of the estimated correlation information in order to reconstruct the images. The joint reconstruction of compressed images has been considered in other applications such as super-resolution, where multiple compressed images are fused to enhance the resolution. Such techniques usually target reconstruction of a high resolution image from multiple images. However, our main target in this chapter is to improve the quality of the compressed views and not to increase the spatial resolution of the reconstructed images.

We propose to reconstruct an image pair $(\hat{I}_1, \hat{I}_2)$ as a solution to the following optimization problem:

$$(\hat{I}_1, \hat{I}_2) = \arg\min_{I_1, I_2}(\|I_1\|_{TV} + \|I_2\|_{TV}) \quad \text{s.t.} \qquad \begin{array}{l} \|I_1 - \tilde{I}_1\|_2 \leq \epsilon_1, \|I_2 - \tilde{I}_2\|_2 \leq \epsilon_1, \\ \|I_2(m,n) - I_1(m + \mathbf{m}^h(m,n), n)\|_2^2 \leq \epsilon_2, \end{array} \tag{4.8}$$

where $\tilde{I}_1$ and $\tilde{I}_2$ represent the compressed views and $\|.\|_{TV}$ represents the total-variation norm defined in Eq. (2.6). From Eq. (4.8) it is clear that we are interested to reconstruct a pair of smooth images that satisfy consistency with both the compressed images and the correlation information $\mathbf{m}^h$. In our framework, we use the total variation (TV) prior, however one could also use a sparsity prior that minimizes the l_1 norm of the coefficients in a sparse representation of the images.

Before solving the optimization problem given in Eq. (4.8), we represent the last constraint $\|I_2(m,n) - I_1(m+\mathbf{m}^h(m,n),n)\|_2^2$ in the matrix format as $\|I_2 - AI_1\|_2^2$. That is, we represent the disparity compensation $I_2(m,n) = I_1(m + \mathbf{m}^h(m,n), n)$ as a linear transformation $I_2 = AI_1$ given as

$$\underbrace{\begin{bmatrix} I_{2,1}^T \\ I_{2,2}^T \\ \vdots \\ I_{2,N_1}^T \end{bmatrix}}_{I_2} = \underbrace{\begin{bmatrix} A^1 & 0 & \dots & 0 \\ 0 & A^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & A^{N_1} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} I_{1,1}^T \\ I_{1,2}^T \\ \vdots \\ I_{1,N_1}^T \end{bmatrix}}_{I_1}, \tag{4.9}$$

where the $I_{1,m}$ represents the m^{th} row of image I_1. The sub-matrix A^m used in Eq. (4.9) is of dimensions $N_2 \times N_2$ that relates the m^{th} row of images.

For planar datasets, the sub-matrix A^m is computed as

$$A^m(p, min(p+\beta, N_2)) = \begin{cases} 1 & \mathbf{m^h}(m,p) = \beta, \\ 0 & \text{otherwise.} \end{cases} \tag{4.10}$$

where $\mathbf{m}^h(m,p)$ represents the disparity value at the p^{th} location in the m^{th} row. If the value of $p+\beta > N_2$ (which might happen at the boundaries) we replace $p+\beta = N_2$, so that the dimensions of the matrix A^m is $N_2 \times N_2$. It is easy to check that the matrix A^m formed using Eq. (4.10) contains only one non-zero value in each row. For example, the matrix A^m corresponding to $\mathbf{m}^h(m,.) = [2\ 2\ 1\ 1]$ is given by

$$A^m = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}. \tag{4.11}$$

Since the matrix A^m contains only one non-zero value in each row it is evident that $I_2(m,i) = I_1(m,j)$ if $A^m(i,j) = 1$. Thus, it is clear that the matrix A^m shifts the pixels in $I_{1,m}$ by the corresponding disparity vector $\mathbf{m}^h(m,.)$, to form $I_{2,m}$.

The disparity compensation between spherical images can also be represented as $I_2 = AI_1$, by adapting the construction of the matrix A. One of the main differences with the perspective cameras is that there is no boundary in the spherical images; instead there exists periodicity in the azimuthal direction. Therefore, we need to consider this periodicity while constructing the matrix A. This can be easily achieved by replacing the boundary condition $min(p+\beta, N_2)$ in Eq. (4.10) with $mod(p+\beta, N_2)$, where mod represents the modulo operator. The sub-matrix A^m for the spherical signals can be constructed as

$$A^m(p, mod(p+\beta, N_2)) = \begin{cases} 1 & \mathbf{m^h}(m,p) = \beta, \\ 0 & \text{otherwise.} \end{cases} \tag{4.12}$$

Using this linear relationship of disparity compensation, Eq. (4.8) can be rewritten as

$$(\hat{I}_1, \hat{I}_2) = \arg\min_{I_1, I_2} (\| I_1 \|_{TV} + \| I_2 \|_{TV}) \quad \text{s.t.} \quad \|I_1 - \tilde{I}_1\|_2 \leq \epsilon_1, \|I_2 - \tilde{I}_2\|_2 \leq \epsilon_1, \|I_2 - AI_1\|_2^2 \leq \epsilon_2. \tag{4.13}$$

4.2.4 Optimization methodology

We propose now a solution for the joint reconstruction problem of Eq. (4.13). We first show that the optimization problem is convex. Then, we propose a solution based on proximal methods.

Proposition 1. *The optimization problem described in Eq. (4.13) is convex.*

Proof: Our objective is to show that all the functions in Eq. (4.13) are convex. However, it is quite easy to check that the functions $\|I_j\|_{TV}$ and $\|I_j - \tilde{I}_j\|_2$, $\forall j \in \{1,2\}$ are convex [135]. So, we have to show that the last constraint $\|I_2 - AI_1\|_2^2$ is a convex function. Let $g(I_1, I_2) = \|I_2 - AI_1\|_2^2$. The function g can be represented as

$$\begin{aligned} g(I_1, I_2) &= \|I_2 - AI_1\|_2^2 \\ &= (I_2 - AI_1)^T (I_2 - AI_1) \\ &= I_2^T I_2 - I_2^T AI_1 - I_1^T A^T I_2 + I_1^T A^T AI_1. \end{aligned} \tag{4.14}$$

The first derivative ∇g and the second derivative $\nabla^2 g$ of the function g are given as

$$\begin{aligned} \nabla g &= (-2A^T I_2 + 2A^T AI_1, 2I_2 - 2AI_1) \\ \nabla^2 g &= \begin{bmatrix} 2AA^T & -2A \\ -2A^T & 2 \end{bmatrix} \\ &= 2C^T C \\ &\succeq 0, \end{aligned} \tag{4.15}$$

where $C = [A^T \quad -I]$ and $2C^T C \succeq 0$ follows from $2x^T C^T Cx = 2\|Cx\|_2^2 \geq 0$ for any x. This means that the Hessian function $\nabla^2 g$ is positive semi-definite and thus $g(I_1, I_2)$ is convex. □

We now describe the proposed optimization methodology to solve the convex problem in Eq. (4.13) based on proximal splitting methods [136]. We first describe our solution for planar images. Later, we explain how to adapt the solution in order to perform joint reconstruction for spherical images. For mathematical convenience, we rewrite the optimization problem given in Eq. (4.13) as

$$\min_X \{\| E_1 X \|_{TV} + \| E_2 X \|_{TV}\} \quad \text{s.t.} \quad \|E_1(Y - X)\|_2 \leq \epsilon_1, \|E_2(Y - X)\|_2 \leq \epsilon_1, \|[-A \quad \mathbb{1}]X\|_2^2 \leq \epsilon_2, \tag{4.16}$$

where $X = \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}$, $Y = \begin{bmatrix} \tilde{I}_1 \\ \tilde{I}_2 \end{bmatrix}$, $E_1 = \begin{bmatrix} \mathbb{1} & 0 \end{bmatrix}$, $E_2 = \begin{bmatrix} 0 & \mathbb{1} \end{bmatrix}$. The optimization problem given in Eq. (4.16) can be visualized as a special case of general convex problem given as

$$\min_{X\in\mathcal{H}}\{f\} = \min_{X\in\mathcal{H}}\{f_1(X) + f_2(X) + f_3(X) + f_4(X) + f_5(X)\}, \tag{4.17}$$

where $\mathcal{H} = \mathbb{R}^{2N}$ is the Hilbert space, and the functions $f_1, f_2, f_3, f_4, f_5 \in \Gamma_0(\mathbb{R}^{2N})$ [48, 136]. $\Gamma_0(\mathbb{R}^{2N})$ is the class of lower semicontinuous convex functions from $\mathbb{R}^{2N}$ to $]-\infty\ +\infty]$ such that the convex function g is not infinity everywhere, i.e., $dom\ g \neq \emptyset$. For the optimization problem given in Eq. (4.16) the functions in the representation of Eq. (4.17) are

1. $f_1(X) = \|E_1X\|_{TV}$
2. $f_2(X) = \|E_2X\|_{TV}$
3. $f_3(X) = i_c(X) = \begin{cases} 0 & X \in c \\ \infty & \text{otherwise} \end{cases}$
   i.e., $f_3(X)$ represents the indicator function of the closed convex set c given as $c = \{X \in \mathbb{R}^{2N} : \|E_1(Y - X)\|_2 \leq \epsilon_1\}$
4. $f_4(X) = i_c(X) = \begin{cases} 0 & X \in c \\ \infty & \text{otherwise} \end{cases}$
   where c is the closed convex set given as $c = \{X \in \mathbb{R}^{2N} : \|E_2(Y - X)\|_2 \leq \epsilon_1\}$
5. $f_5(X) = i_d(X) = \begin{cases} 0 & X \in d \\ \infty & \text{otherwise} \end{cases}$
   where d is the closed convex set given as $d = \{X \in \mathbb{R}^{2N} : \|[-A \quad \mathbb{1}]X\|_2^2 \leq \epsilon_2\}$.

The solution to the problem of Eq. (4.17) can be found by generating the recursive sequence $X^{(t+1)} = prox_{\beta f}(X^{(t)})$, $\beta > 0$, where the function f is given as $f = f_1 + f_2 + f_3 + f_4 + f_5$. The proximity operator is defined as the $prox_f(X) = min_{X\in\mathcal{H}}\ f(Z) + \frac{1}{2}\|X - Z\|^2$. The main difficulty with these iterations are the computations of the $prox_{\beta f}(X)$ operator, as there is no closed form expression to compute the $prox_f(X)$ especially when the function f is the cumulative sum of two or more functions. In such cases, instead of the computing the $prox_f(X)$ directly for the combined function f one can perform a sequence of calculations involving separately the individual operators $prox_{f_1}(X)$, $prox_{f_2}(X)$, $prox_{f_3}(X)$, $prox_{f_4}(X)$ and $prox_{f_5}(X)$. These class of algorithms are popularly known as *splitting methods* as these methods proceed by splitting the combined function f into f_1, f_2, f_3, f_4 and f_5, and hence allows an easily implementable algorithm [136].

We describe in more details the methodology to compute the *prox* for the functions f_1, f_2, f_3, f_4 and f_5. For the function $f_1(X) = \|E_1X\|_{TV}$ the operator $prox_{f_1}(X)$ can be computed iteratively using Chambolle's algorithm [137]. A similar approach can be used to compute the $prox_{f_2}(X)$. The function f_3 can be represented as $f_3 = F \circ G$, where $F = i_{D^2(\epsilon_1)}$ and $G = E_1X - E_1Y$. The set $D^2(\epsilon_1)$ represents the l_2 ball defined as $D^2(\epsilon) = \{y \in \mathbb{R}^{2N} | \|y\|_2 \leq \epsilon_1\}$. The $prox_{f_3}$ can be computed using the following closed form expression given as

$$prox_{f_3}(X) = prox_{F\circ G}(X) = X + (E_1)^*(prox_F - \mathbb{1})(G(X)) \tag{4.18}$$

[138], where $(E_1)^*$ represents the conjugate transpose of E_1. The $prox_F(y)$ with $F = i_{D^2(\epsilon_1)}$ can be computed using radial projection [136] as

$$prox_F(y) = \begin{cases} y & \|y\|_2 \leq \epsilon_1 \\ \frac{y}{\|y\|_2} & \text{otherwise.} \end{cases} \tag{4.19}$$

The *prox* for the function f_4 can also be solved using Eq. (4.18) with $F = i_{D^2(\epsilon_1)}$ and $G = E_2X - E_2Y$. Finally, the function f_5 can be represented with $F = i_{D^2(\sqrt{\epsilon_1})}$ and an affine operator $G = [-A \quad \mathbb{1}]X = \Omega X$, i.e., $f_5 = F \circ G$. As the operator Ω is not a tight frame, the $prox_{f_5}$ can be computed using an iterative

scheme [138]. Let $\mu_t \in (0, 2/c_2)$, and c_1 and c_2 be the frame constants with $c_1\mathbb{1} \leq \Omega\Omega^* \leq c_2\mathbb{1}$. The $prox_{f_5}$ can be calculated iteratively as

$$u^{(t+1)} = \mu_t(\mathbb{1} - prox_{\mu_t^{-1}F})(\mu_t^{-1}u^{(t)} + Gp^{(t)}) \tag{4.20}$$

$$p^{(t+1)} = X - \Omega^* u^{(t+1)}, \tag{4.21}$$

where $u^{(t)} \to u$ and $p^{(t)} \to prox_{F\circ G} = prox_{f_5} = X - \Omega^* u$. It has been shown that both $u^{(t)}$ and $p^{(t)}$ converges linearly and the best convergence rate is attained when $\mu_t = 2/(c_1 + c_2)$ [138].

In our work, we use the parallel proximal algorithm (PPXA) proposed by Combettes *et al.* [136] to solve Eq. (4.17), as the algorithm can be easily implementable on multicore architectures due to its parallel structure. The PPXA algorithm starts with an initial solution $X^{(0)}$ and computes the $prox_{f_1}$, $prox_{f_2}$, $prox_{f_3}, prox_{f_4}$ and $prox_{f_5}$ in each iteration, and the result is used to update the current solution $X^{(0)}$. The iterative procedure of computing the *prox* for functions f_1, f_2, f_3, f_4 and f_5, and the updating steps are repeated until convergence is reached. The authors have shown that the sequence $(X^{(l)})_{l\geq 1}$ generated by the PPXA algorithm is guaranteed to converge to the solution of problems such as the one given in Eq. (4.17).

So far, we have described the *prox* computations for solving Eq. (4.17) on planar images. For spherical images we have to consider the specific geometry on the sphere. The *prox* operator for minimizing the total variation norm $\|.\|_{TV}$ of the functions f_1 and f_2 can be solved with the modified Chambolle's algorithm by defining the spherical signal on a weighted graph [139, 140]. In more words, a weighted graph is constructed with vertices corresponding to the image pixels and edges corresponding to the connections between adjacent pixels. Such a graph-based representation allows us to compute effectively the gradient and divergence differential operators. For more details, we refer the reader to [141]. The *prox* computations for the functions f_3, f_4 and f_5 are obtained by adapting the $prox_F$ operator given in Eq. (4.19) using the l_2 norm definition on the sphere. The l_2 norm for a spherical signal h is defined as

$$\|h\|_{2,s} = \sqrt{\int_\theta \int_\varphi h^2(\theta, \varphi) sin\theta d\theta d\varphi}. \tag{4.22}$$

In the above expression, we use subscript s to denote the norm definition on sphere. Using this definition, the $prox_F$ operator for the spherical signal is computed as

$$prox_F(y) = \begin{cases} y & \|y\|_{2,s} \leq \epsilon_1, \\ \frac{y}{\|y\|_{2,s}} & \text{otherwise.} \end{cases} \tag{4.23}$$

Using the modified $prox_F$ computation given in Eq. (4.23), the *prox* of the functions f_3, f_4 and f_5 are computed with Eq. (4.18) and Eqs. (4.20)-(4.21). Finally, the joint reconstruction problem for the spherical images can be solved using the PPXA algorithm described above, by plugging the adapted *prox* functions in order to cope with the spherical image geometry.

4.3 Joint reconstruction performance

In this section, we verify the performance of our proposed joint reconstruction algorithm on correlated planar and spherical images. We also analyze the effect of image compression on the correlation estimation performance.

4.3.1 Planar images

For the planar images, we carry out experiments on two natural datasets (*Tsukuba* and *Venus*[1] [7]), and one synthetic dataset (*Object* shown in Fig. 4.2). A wavelet transform is applied on the luminance component of the images I_1 and I_2 using *Daub*-9/7 filter [142], followed by the encoding of wavelet coefficients based on SPIHT algorithm [128]. We solve the disparity estimation optimization problem given in Eq. (4.1) using α-expansion algorithm in Graph Cuts [133, 130, 134]. In our experiments, the parameter λ in Eq. (4.1) is heuristically selected such that the disparity error (DE) is minimized, where the DE is computed as

$$DE = \frac{1}{N_1 \times N_2} \sum_{\mathbf{z}=(m,n)} \left\{ |\mathbf{M}^h(\mathbf{z}) - \mathbf{m}^h(\mathbf{z})| > 1 \right\}, \tag{4.24}$$

where $\mathbf{M}^h$ represents the groundtruth disparity image [7]. It should noted that one could also estimate the parameter λ in Eq. (4.1) using the automated method proposed in [143], i.e., parameter λ is estimated without the knowledge of the groundtruth information. However, for simplicity, in our experiments we select the best parameter λ based on trial and error experiments using the groundtruth information. The estimated disparity result is then used to decode an image pair by solving Eq. (4.13). In our experiments, we fix the parameters $\epsilon_1 = 2$ and $\epsilon_2 = 3$ that are selected based on trial and error methods such that it maximizes the quality of the reconstructed images.

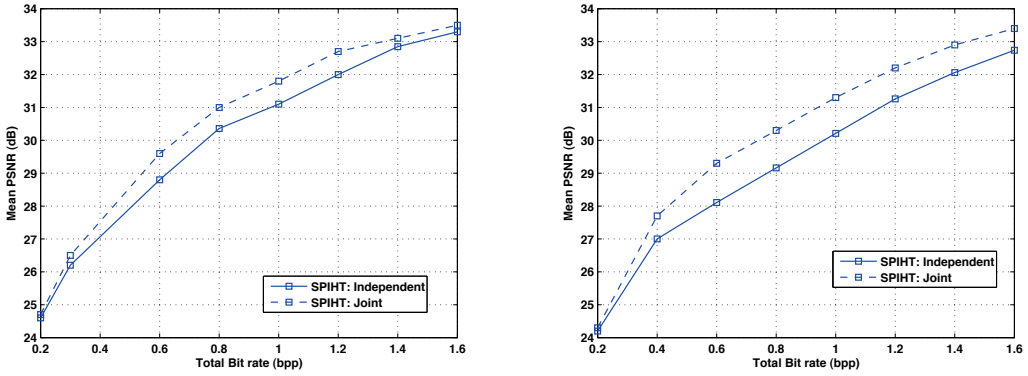
(a) I_1 (b) I_2 (c) $\mathbf{M}^h$

Figure 4.2: Object planar dataset: (a) original left view I_1; (b) original right view I_2; (c) groundtruth disparity image $\mathbf{M}^h$.

Fig. 4.3(a) and Fig. 4.3(b) compare the overall quality of the decoded images between the independent and joint decoding solutions, for Tsukuba and Venus datasets respectively. For total bit rates > 0.6, we see from Fig. 4.3 that the proposed joint reconstruction scheme outperforms the independent reconstruction by a margin of about 0.9 and 1.3 dB for both datasets. This confirms that the proposed joint reconstruction framework effectively exploits the correlation between images. While carrying out the experiments we have also observed that the reconstruction quality of the images $\hat{I}_1$ and $\hat{I}_2$ are quite similar for a given target bit rate. From Fig. 4.3, we further see that the joint reconstruction fails to improve the quality of the compressed images at low rate due to the poor quality of the disparity estimation. In the coming paragraphs, we focus on studying the impact of the compression algorithms on the performance of the disparity estimation.

For the sake of clarity, we first study this impact in the Object dataset shown in Fig. 4.2. In particular, we carry out detailed analysis on the disparity results estimated from the compressed images encoded at a bit rate of 0.1 bpp. The compressed views are available in Fig. 4.4(a) and Fig. 4.4(b) respectively, and Fig. 4.4(c) shows the disparity error, where the white pixels denote an error larger than 1. From Fig. 4.4(c), we observe that the disparity value corresponding to the *big* object is recovered correctly, but not for the *small* object. The reason is that, from the compressed images given in Figs. 4.4(a) and 4.4(b), we see that the contrast or intensity differences in the proximity of the *small* object is not high enough to detect the disparity activity. It can be further noted that the intensity difference in the proximity of the *big* object is

[1] available at http://vision.middlebury.edu/stereo/data/



(a) (b)

Figure 4.3: Comparison of the RD performances between the independent and joint decoding schemes: (a) Tsukuba dataset; (b) Venus dataset.

(a) $\tilde{I}_1$ (b) $\tilde{I}_2$ (c) $|\mathbf{M}^h - \mathbf{m}^h| > 1$

Figure 4.4: Disparity estimation from the SPIHT-based encoded images at a bitrate of 0.1 bpp in the Object dataset. (a) Compressed left view $\tilde{I}_1$; (b) compressed right view $\tilde{I}_2$; (c) error in the disparity map, with DE=4.32%. The white pixels denote an error larger than one.

Table 4.1: Disparity error for the Tsukuba and Venus datasets at various bit rates. Results are tabulated using $k = 4$ and $k = 5$ wavelet decomposition levels respectively.

Bit rate (bpp)		0.4	0.5	0.6	0.7	0.8
$DE(\%)$	Tsukuba	9.4	7.94	7.52	7.04	6.2
	Venus	10.4	9.26	8.5	8.1	7.7

sufficiently large; this leads to the correct disparity estimation. Fig. 4.5 illustrates the intensity distribution in the proximity of the *big* and *small* objects with and without compression of the images used in the disparity estimation. In our experiments, the histogram is generated on a small patch taken around the objects. From Fig. 4.5(a) and Fig. 4.5(c), we see that the intensity distribution generated from the original image I_1 is bimodal, i.e., there exists two peaks separated with a cliff. However, after compression we see that the bimodal distribution is preserved only for the *big* object and not for the *small* object, as shown in Fig. 4.5(b) and Fig. 4.5(d) respectively. This is because, the rate allocation scheme used in SPIHT or in any coding standards allocate more bits to encode the texture and strong edges than the weak edges and smooth regions.

We now carry out similar experiments on the Tsukuba and Venus natural datasets. Table 4.1 tabulates the disparity error for both datasets at various bit rates. As expected, we see in Table 4.1 that there is

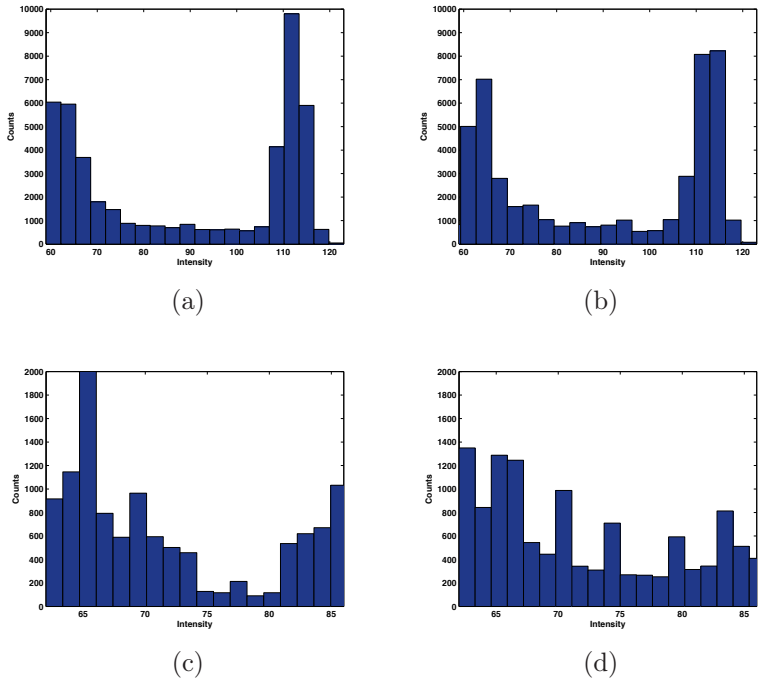
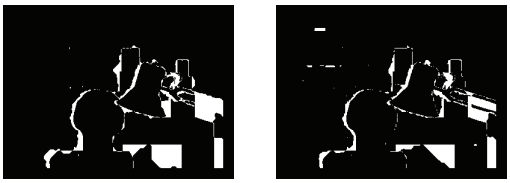


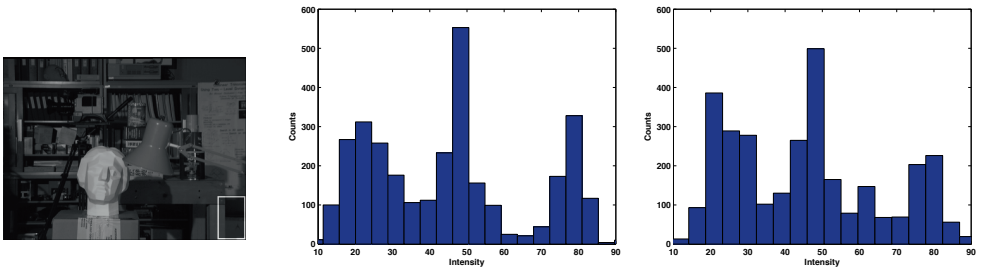
Figure 4.5: Comparison of the intensity distribution in the proximity of the *big* and *small* objects with and without compression in the Object dataset. Intensity distribution in the proximity of the *big* object generated from the (a) original image I_1, and (b) SPIHT-based compressed image $\tilde{I}_1$ at a rate of 0.1 bpp. Intensity distribution in the proximity of the *small* object generated from the (c) original image I_1, and (d) SPIHT-based compressed image $\tilde{I}_1$ at a rate of 0.1 bpp.



(a) $|\mathbf{M}^h - \mathbf{m}^h| > 1$ (b) $|\mathbf{M}^h - \mathbf{m}^h| > 1$

Figure 4.6: Error between the estimated and groundtruth depth image at bit rates (a) 0.4 bpp and (b) 0.6 bpp in the Tsukuba dataset. The images are encoded using SPIHT-based rate allocation scheme. The white pixels correspond to a disparity error larger than 1. The disparity error is (a) 9.52% and (b) 7.52%.

an overall improvement in the quality of the disparity image as the bit rate increases. However, we have observed that the quality of the disparity image is improved only in the regions close to the strong edges, but the disparity value remains unchanged in regions close to the weak edges or in low contrast regions. For example, in Fig. 4.6(a) and Fig. 4.6(b) we show the disparity error for the Tsukuba dataset generated from the compressed images at bit rates 0.4 and 0.6 bpp (per view) respectively. Comparing Fig. 4.6(a) and Fig. 4.6(b) we see that the quality of the disparity value is improved only in the regions close to the strong edges (e.g., structure of the lamp, head) and not in the low contrast regions (e.g., between the table leg and the background at the bottom most right corner). This is because the SPIHT codec [128] minimizes



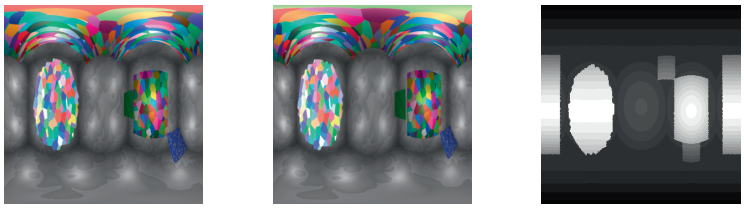
(a) (b) (c)

Figure 4.7: Effect of compression on the bimodal intensity distribution in the Tsukuba dataset. (a) Image patch (marked in white) on which the histogram is generated. (b) Intensity distribution in the original image I_1. (c) Intensity distribution generated from SPIHT-based compressed image $\tilde{I}_1$ encoded at a bit rate of 0.4 bpp.

the MSE in the reconstructed image, which obviously allocates more bits to the strong edges (high contrast regions) comparing to weak edges (low contrast regions). As a result of insufficient bits allocated to the low contrast regions (e.g., between the table leg and the background at the bottom most right corner) the bimodal intensity distribution is not preserved in such regions, as demonstrated in Fig. 4.7. Similar observations are made with other standard encoding schemes like JPEG 2000, JPEG which also minimize the average MSE for a given target bit rate. Therefore, the standard rate allocation schemes fail to provide a compressed representation that could lead to efficient estimation of the disparity between images.

4.3.2 Spherical images

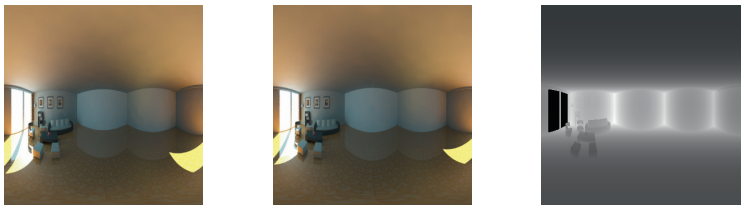
We now carry out similar experiments on spherical images that are captured by a pair of omnidirectional cameras. We study the performance on two synthetic spherical datasets, *Oval* and *Room* shown in Fig. 4.8 and Fig. 4.9 respectively. The spherical images are first transformed using a multi-resolution representation (i.e., the Spherical Laplacian Pyramid (SLP) described in Section 3.2.2), followed by the encoding of transform coefficients based on SPIHT [128]. Then, a disparity image is estimated from the compressed spherical images $\tilde{I}_1$ and $\tilde{I}_2$ using the spherical Graph Cut algorithm proposed by Arican *et al.* [112]. The joint reconstruction experiments are carried out using the parameter set $(\epsilon_1, \epsilon_2) = (1, 3)$.



(a) I_1 (b) I_2 (c) $\mathbf{M}^h$

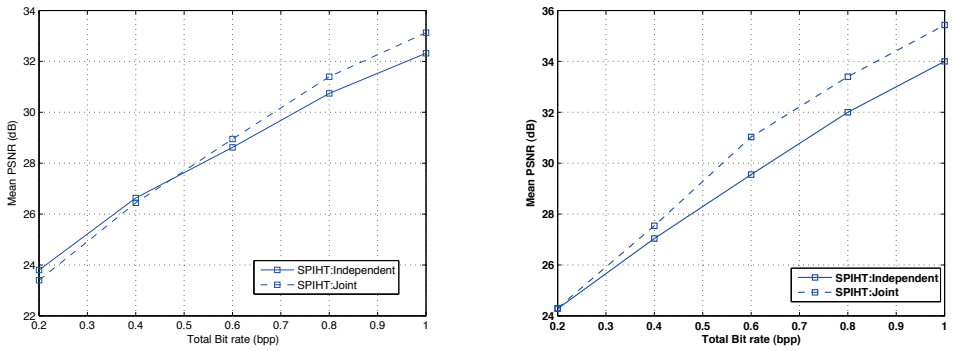
Figure 4.8: Oval Spherical dataset: (a) original left view I_1; (b) original right view I_2; (c) groundtruth disparity image $\mathbf{M}^h$.

Fig. 4.10(a) and Fig. 4.10(b) compare the reconstruction image quality between the proposed and independent decoding schemes for the Oval and Room datasets respectively. We see that the proposed scheme outperforms the independent reconstruction results for total bit rates larger than 0.6, which is consistent



(a) I_1 (b) I_2 (c) $\mathbf{M}^h$

Figure 4.9: Room Spherical dataset: (a) original left view I_1; (b) original right view I_2; (c) groundtruth disparity image $\mathbf{M}^h$.



(a) (b)

Figure 4.10: Comparison of the RD performances between the independent and joint decoding schemes. (a) Oval dataset; (b) Room dataset.

with our earlier observations in the planar datasets. Also, we again see that the benefit of joint reconstruction is not noticeable at low bit rates. Even the quality of the reconstructed images could be lower compared to the quality of the compressed images (see Fig. 4.10(a)). This is due to the poor disparity estimation from the highly compressed images (0.1 bpp per view in this case). In Figs. 4.11(a) and (c) we show the disparity images estimated from the compressed images encoded at bit rates 0.1 and 0.4 bpp (per view) respectively. The corresponding disparity errors are shown in Figs. 4.11(b) and (d). From Fig. 4.11(b) and Fig. 4.11(d) we see that at low rates the disparity image is poor, with DE=78% and DE=49% respectively. From the above experiments, it is clear that the standard rate allocation schemes are not good at detecting the disparity activity in the regions where the intensity variation is small. We need to preserve sufficient intensity variations (i.e., bimodal intensity distribution), to be able to detect the disparity activity in the low contrast regions.

4.4 Rate allocation for improved disparity estimation

In the previous section, we have seen that the underlying disparity structure of the scene is not precisely estimated in the low contrast regions, when the stereo image pair is encoded using traditional coding schemes. In order to improve the quality of the disparity image estimation at the joint decoder, we propose here a rate allocation scheme that preserves the weak edges (i.e., low contrast regions) in the compressed views $\tilde{I}_1$

(a) $\mathbf{m}^h$ (b) $|\mathbf{M}^h - \mathbf{m}^h| > 1$ (c) $\mathbf{m}^h$ (d) $|\mathbf{M}^h - \mathbf{m}^h| > 1$

Figure 4.11: Comparison of the disparity images estimated from the SPIHT-based compressed images at rates of 0.1 bpp and 0.4 bpp (per view) in the Oval spherical dataset. (a) and (c) disparity images estimated from rates of 0.1 bpp and 0.4 bpp respectively. (b) and (d) respective disparity errors; DE = 78% and 49% respectively.

and $\tilde{I}_2$, at the cost of a marginal penalization in the quality of the background.

4.4.1 General principle

In this section, we outline the general principles of our rate allocation scheme, while the actual coding scheme built on these principles is described in the next section. Let $\check{I}$ denote the transform coefficients computed by applying a linear transform (e.g., wavelet, SLP) to the image $I \in \{I_I, I_2\}$ using k decomposition levels. For the sake of clarity, we assume that $\check{I}^r$ and $\check{I}^b$ denote the coefficients corresponding to the low contrast and the background (i.e., texture) spatial regions respectively, such that $\check{I} = \check{I}^b \bigcup \check{I}^r$. We describe later how the transform coefficients $\check{I}^r$ and $\check{I}^b$ corresponding to the low contrast and the background regions can be identified from the coefficient image $\check{I}$. In the previous section, we have seen that the disparity activity in the low contrast region is not detected as the standard rate allocation schemes fail to encode the corresponding transform coefficients $\check{I}^r$ with sufficient bits. For better understanding, we explain this using an example shown in Fig. 4.12(a); it shows the bitplane representation of the coefficients $\check{I}$ with M bits where $M = \lceil log_2(max(\check{I})) \rceil$ ($M = 8$ in the case). The left shaded region represents the background coefficients $\check{I}^b$ and right shaded region represents the coefficients $\check{I}^r$. The bits in the top row represent the signs of the coefficients. In standard rate allocation schemes [144, 128, 145], the coefficients in Fig. 4.12(a) are encoded progressively starting from MSB to LSB (i.e., from left to right and top to bottom). From Fig. 4.12(a), it is clear that such coding approaches fail to encode the region $\check{I}_r$ at low bit rates as the position of the first 1 appears deep in the bit plane representation (marked in red). Therefore, it is clear that we need to allocate sufficient bits to encode the coefficients $\check{I}^r$ in order to the preserve the visual information in low contrast regions. At the same time, we should not penalize the reconstruction quality in the background region. In essence, we need to allocate sufficient bits to encode the coefficients $\check{I}^r$ and $\check{I}^b$, so that an accurately disparity image can be estimated from the compressed views.

The key idea of our approach is to scale the low contrast coefficients $\check{I}^r$, so that the scaled low contrast coefficients are placed in the higher bitplanes. We propose to calculate the scale as $f = 2^d$, where $d = \lfloor S(M - M_r) \rfloor$, $M_r = \lceil log_2(max(\check{I}^r)) \rceil < M$ and $S > 0$ is a constant. When the coefficients $\check{I}^r$ are scaled by $f = 2^d$, the bits in the bitplane $M_r = \lceil log_2(max(\check{I}^r)) \rceil$ associated to the low contrast coefficients $\check{I}^r$ are shifted up by d positions. For example, in Fig. 4.12(b) we show the effect of scaling the coefficients $\check{I}^r$ when the parameter S is set to 1; therefore $d = 2$ in this case as $M_r = 6$ and $M = 8$. After scaling, we see that the bits in the bitplane $M_r = \lceil log_2(max(\check{I}^r)) \rceil$ (marked in red in Fig. 4.12(a)) associated to $\check{I}^r$ are shifted to the MSB bitplane (marked in green in Fig. 4.12(b)). Also, one can easily check that selecting $S > 1$ shifts the bits in the bitplane $M_r = \lceil log_2(max(\check{I}^r)) \rceil$ associated to $\check{I}^r$ above the MSB bitplane. This can be easily understood from Fig. 4.12 by setting $f = 8$ (i.e., $d = 3$) for example. Therefore, the coefficients belonging to low contrast regions $\check{I}^r$ can be shifted up to desired bitplane by controlling the parameter S.

We now discuss the influence of parameter S for encoding the low contrast $\check{I}^r$ and background $\check{I}^b$

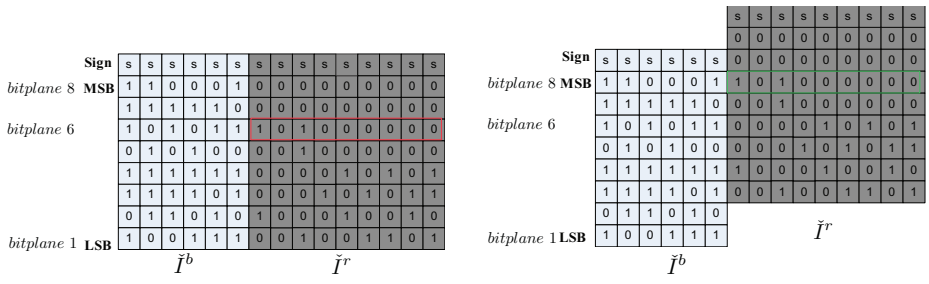


Figure 4.12: General principle of the proposed rate allocation scheme. $\check{I}^b$ and $\check{I}^r$ represents the transform coefficients computed from image I. (a) Bitplane representation of original transform coefficients $\check{I} = \check{I}^b \bigcup \check{I}^r$. The bits in the top row represent the signs of the coefficients, and the number of bitplanes $M = 8$. (b) The coefficients $\check{I}^r$ is scaled by $f = 2$ so that it shifts the bits in the bitplane $M_r = \lceil log_2(max(\check{I}^r)) \rceil = 6$ (marked in red) associated to $\check{I}^r$ are shifted to the MSB bitplane $M = 8$ (marked in green).

coefficients. First of all, it is clear that selecting $S = 0$ (i.e., no scaling since $f = 0$) results in the standard encoding functionality. Then, selecting $S = 1$ results in shifting the bits in the bitplane M_r associated to the coefficients $\check{I}^r$ to the MSB bitplane M. When the coefficients are progressively coded from MSB to LSB, we see that the bits corresponding to $\check{I}^r$ are placed along with the bits corresponding to $\check{I}^b$ (see Fig. 4.12(b)). Therefore, it is clear that $S = 1$ distribute the bits in equal proportions to the low contrast and background coefficients. Then, selecting $S > 1$ shifts the bits in the bitplane M_r associated to $\check{I}^r$ above the MSB bitplane; this eventually allocates more bits to the low contrast coefficients than to the background coefficients. Finally, by following similar ideas, it is easy to check that $S < 1$ allocates more bits to the strong edges than to the smooth regions. Therefore, it is clear that the number of bits allocated to the low contrast and background regions could be controlled using the parameter S.

Based on the ideas discussed here, we build a practical encoding scheme in the following section for improving the quality of the disparity information estimated from the compressed images. In particular, we need to answer following questions:

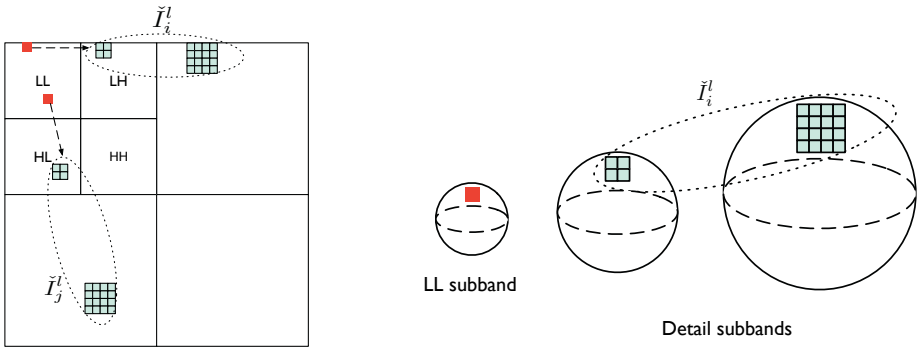
1. How to identify the positions of the low contrast and background (i.e., texture) coefficients from the coefficient image $\check{I}$?

2. How to find the parameter S for an improved disparity estimation ?

4.4.2 SPIHT-based rate allocation algorithm

In this section, we describe the proposed encoding scheme constructed using SPIHT-based coding principles. In particular, our goal is to identify the position of coefficients corresponding to the low contrast regions (i.e., smooth regions) in the image $I \in \{I_1, I_2\}$. Once the coefficient positions are known we scale them prior to encoding with an appropriate S, as discussed in the previous section.

We denote $\check{I}$ as the transform coefficient image and we split the coefficient image into two partitions: (1) LL subband coefficients; (2) detail subband coefficients. It is well known that there exists a spatial relationship [128] among the LL subband and detail subbands coefficients; this is shown in Fig. 4.13. The coefficients in LL subband form the *tree root* (marked in *red*) in a spatial tree (say i). The set of coefficients $\check{I}_i^l$ in the detail subband corresponding to the same spatial location is usually called as *descendants* or *offspring* in the i^{th} spatial tree. In more words, due to the compact support of the decomposition filters (e.g., wavelets, SLP), the coefficients $\check{I}_i^l$ represent a particular spatial region in the image. Furthermore, the magnitude of the transform coefficients $\check{I}_i^l$ depends on the particular image characteristics in the corresponding spatial region



(a) (b)

Figure 4.13: Spatial orientation tree of the transform coefficients in SPIHT [128]. The coefficients in the LL band denote the tree root in a spatial tree i (marked in red). The set of coefficients $\check{I}_i^l$ corresponding to the same spatial location is called as descendants or offsprings in the i^{th} spatial tree. (a) Illustration for the planar images using a wavelet decomposition with two levels. (b) Illustration for the spherical images using a SLP decomposition with two levels.

of the image. In other words, in multi-resolution representations, the coefficients are magnitude ordered based on the spatial characteristics in the image; for strong edges and texture regions $M_i \approx M$, and for low contrast and smooth regions $M_i < M$, where $M_i = \lceil log_2(max(\check{I}_i^l)) \rceil$ and $M = \lceil log_2(max(\check{I})) \rceil$. Therefore, the spatial location of the low contrast regions can be identified directly from the magnitude of the transform coefficients $\check{I}_i^l$. We then calculate the scale $f_i = 2^{d_i}$ for each i^{th} spatial tree, where $d_i = \lfloor S(M - M_i) \rfloor$. It is easy to check that the scale parameter $f_i \approx 1$ for those spatial trees which represent texture spatial regions and close to 2^{SM} for smooth spatial regions. Once the scale f_i is computed, the corresponding coefficients $\check{I}_i^l$ are scaled accordingly, which shifts the bits in the bitplane M_i associated to $\check{I}_i^l$ to the higher bitplane position.

We now discuss the rate allocation between coefficients in the low frequency LL subband and high frequency subbands. Let b_p and b_d denote the bits allocated to encode the coefficients in the LL band and detailed subbands respectively, such that the total bit rate $b = b_p + b_d$. As most of the signal energy is compacted in the highest level of the pyramid (i.e., the LL band) enough bits must be spent to encode the coefficients in the LL band, otherwise it degrades the image quality. Based on experiments, we heuristically select $b_p = 0.2b$ to encode the coefficients in the LL band. However, the total number of bits required to faithfully represent the LL band coefficients are $\frac{N_1}{2^k} \times \frac{N_2}{2^k} \times (M + 1)$, where k denotes the decomposition levels and the one extra bit plane $M + 1$ is accounted to encode the sign of the coefficients. Therefore, the number of bits allocated to the LL band is given as $b_p = min(0.2b, \frac{N_1}{2^k} \times \frac{N_2}{2^k} \times (M + 1))$.

The modified SPIHT-based rate allocation methodology is summarized in Algorithm 3. As shown in Algorithm 3, we first compute the number of bit planes M and the number of bits to encode the LL and detailed subbands, denoted as b_p and b_d respectively. We then compute the shift or scale parameter d_i (see lines 4-7) and it is used to scale the coefficients $\check{I}_i^l$ by $f_i = 2^{Sd_i}$. The set of scale parameters $\{d_i\}$ are entropy coded (e.g., arithmetic encoder) and placed in the header of the final bit stream (denoted as H in line 9). We then progressively encode the low frequency LL band coefficients, that is denoted as Lo in line 14. Then, the scaled detail subband coefficients are progressively encoded (denoted as Ho in line 19) until the final length of the bit stream reaches b bits. Finally, the encoder transmits the header information H and the encoded LL and detail subband coefficients, respectively Lo and Ho. The decoder can reverse the encoding operations, and the image is finally reconstructed after inverting the linear transform (e.g., wavelet, SLP).

Finally, we describe the effect of parameter S on the encoding of low contrast (i.e., smooth) and back-

Algorithm 3 Proposed rate allocation scheme

Input: coefficients $\check{I}$, parameter S and bit rate b.
Output: BS
Calculate $M = \lceil log_2(max(\check{I})) \rceil$, $b_p = min(0.2b, \frac{N_1}{2^k} \times \frac{N_2}{2^k} \times (M+1))$ and $b_d = b - b_p$
Initialize bit counter $s_d = 0$
for each i^{th} spatial tree **do**
    Compute $M_i = \lceil log_2(max(\check{I}_i^l)) \text{ and } d_i = S(M - M_i)$
    Multiply $\check{I}_i^l$ with $f_i = 2^{d_i}$
end for
$H \leftarrow$ arithmetic encoder($\{d_i\}$)
Update bit counter: $s_d \leftarrow s_d + length(H)$
Encoding of the LL band
Initialize bit counter $s_p = 0$
if $s_p < b_p$ **then**
    $Lo \leftarrow$ encode progressively the LL subband coefficients
    Update bit counter: $s_p \leftarrow s_p + length(Lo)$
end if
Encoding of the detail bands
if $s_d < b_d$ **then**
    $Ho \leftarrow$ encode progressively the detail subband coefficients
    Update bit counter: $s_d \leftarrow s_d + length(Ho)$
end if
Transmit BS $\leftarrow [H\ Lo\ Ho]$

ground (i.e., texture) coefficients. Clearly, $S = 0$ corresponds to traditional coding scheme, that allocates more bits to the texture regions compared to the smooth regions. Then, when S is set to one (i.e., $d_i = (M - M_i)$) the bits in the bitplane $M_i = \lceil log_2(max(\check{I}_i^l)) \rceil$ corresponding to $\check{I}_i^l$ are shifted to the MSB bitplane M (see Fig. 4.12). Therefore, $S = 1$ results in distributing the bits b_d equally to all the spatial regions in the image. In other words, it concentrates encoding of both smooth and texture regions. We show by experiments that $S = 1$ is the best choice for an accurate disparity estimation, since it preserves the visual information in smooth regions without significantly penalizing the reconstruction quality in the texture regions. The effect of other choices of S are summarized in Fig. 4.14. By increasing S, the proposed rate allocation scheme gives more importance to the encoding of smooth regions than to the texture regions. As a result, the reconstruction image quality is maximum at $S = 0$ for a given bit rate and it is minimum at $S >> 1$.

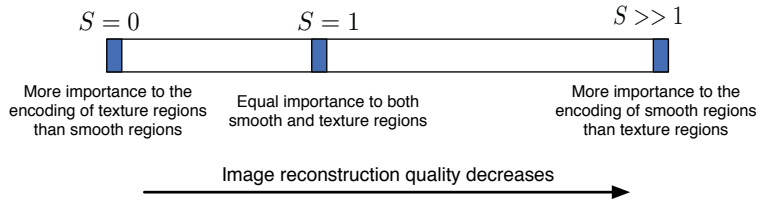


Figure 4.14: Effect of the parameter S in the proposed rate allocation scheme. $S = 0$ corresponds to the traditional SPIHT coding scheme [128]. The quality of reconstructed image is maximized at $S = 0$, and it is penalized with increasing S, due to the allocation of more bits to the smooth regions than texture regions.

4.5 Improved disparity estimation performance

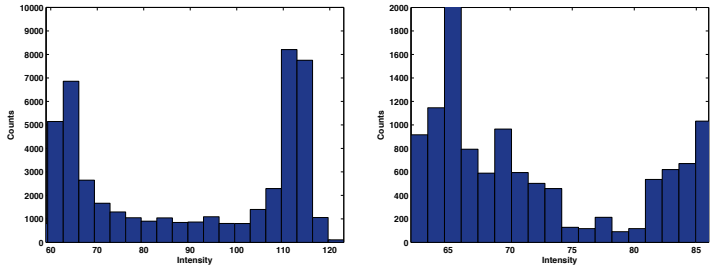
In this section, we demonstrate the benefits of using the proposed rate allocation methodology at encoding for disparity estimation from the compressed images. We study the effect of the parameter S on the image quality and the disparity estimation. Finally, we study the importance of accurate disparity estimation in the view synthesis and joint reconstruction applications. Unless stated differently, we set $S = 1$ in the proposed rate allocation scheme.

4.5.1 Planar images

We first apply our proposed rate allocation methodology to the synthetic Object dataset shown in Fig. 4.2. Fig. 4.15(a) and Fig. 4.15(b) show the compressed left and right images respectively that have been encoded at a bit rate of 0.1 bpp. By comparing these compressed images with the SPIHT-based compressed images available in Fig. 4.4, we observe that the reconstruction quality of the *small* object has been improved at the cost of marginal penalization in the quality of the *big* object. Interestingly, the proposed rate allocation methodology preserves the bimodal intensity distributions for both *big* and *small* objects, as shown in Fig. 4.16(a) and Fig. 4.16(b) respectively. As a result, the disparity values in both objects are estimated correctly, with DE=1.26%, as shown in Fig. 4.15(c).

(a) (b) (c)

Figure 4.15: Disparity estimation from the proposed rate allocation scheme from the encoded images at a bitrate of 0.1 bpp in the Object dataset. (a) Compressed left view $\tilde{I}_1$; (b) compressed right view $\tilde{I}_2$; (c) error in the disparity image with DE = 1.26%.



(a) (b)

Figure 4.16: Effect of proposed rate allocation scheme on the bimodal intensity distribution in the Object dataset. (a) Intensity distribution in the proximity of the *big* object; (b) intensity distribution in the proximity of the *small* object. The intensity distributions are computed from the compressed image $\tilde{I}_1$ shown in Fig. 4.15(a).

We then apply our rate allocation scheme to the natural image datasets and we compare our disparity estimation results with the ones that are estimated from images compressed with the standard encoding

Table 4.2: Comparison of disparity error between the proposed, SPIHT, JPEG 2000 and JPEG coding schemes for the Tsukuba dataset. The influence of parameter S on the disparity estimation accuracy is also compared. Size of the image: $N = 384 \times 288$. Results are tabulated using a wavelet decomposition with $k = 4$ levels.

Bit rate per view (bpp)	Disparity Error (%)					
	Proposed S = 1	Proposed S = 0.5	Proposed S = 2	SPIHT	JPEG 2000	JPEG
0.3	8.6	9.28	22.15	10.24	9.96	11.18
0.4	7.14	8.8	19.8	9.4	9.07	9.26
0.5	5.66	6.3	13.56	7.94	8.42	7.45
0.6	5.58	6.78	10.82	7.52	7.46	6.99
0.7	5.15	5.28	10.1	7.04	7	6.79
0.8	4.8	5	8.9	6.2	6.25	6.11

Table 4.3: Comparison of disparity error between the proposed, SPIHT, JPEG 2000 and JPEG coding schemes for the Venus dataset. The influence of parameter S on the disparity estimation accuracy is also compared. Size of the image: $N = 434 \times 383$. Results are tabulated using a wavelet decomposition with $k = 5$ levels.

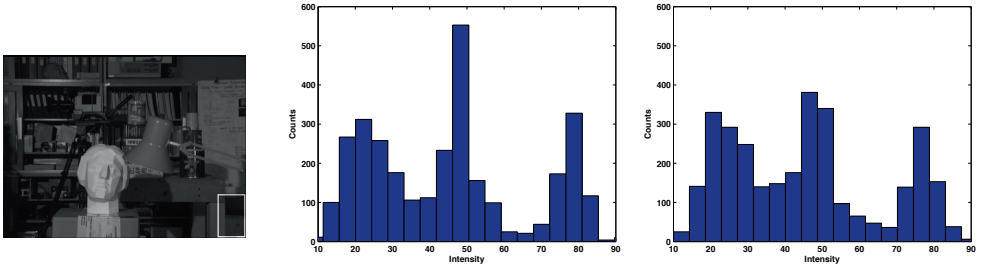
Bit rate per view (bpp)	Disparity Error (%)				
	Proposed S = 1	Proposed S = 2	SPIHT	JPEG 2000	JPEG
0.4	6.56	8.1	10.4	8.92	9.61
0.5	5.67	6.18	9.26	7.54	9.24
0.6	4.89	5.8	8.5	7.64	8.37
0.7	4.14	5.6	8.1	7.22	6.91
0.8	3.91	5.25	7.7	6.93	5.91

solutions based on SPIHT, JPEG 2000 and JPEG. Table 4.2, Table 4.3 and Table 4.4 tabulate the disparity error for the Tsukuba,Venus and Map datasets respectively, computed from the compressed images at various encoded bit rates. From Tables 4.2 and 4.3 we see that for a given bit rate, the proposed rate allocation scheme with $S = 1$ leads to a better disparity estimation than the standard encoding schemes based on SPIHT, JPEG 2000 and JPEG. This is due to the better representation of the visual information in the low contrast regions as described in Section 4.4. In other words, the proposed rate allocation scheme preserves the bimodal intensity distribution in the low contrast regions. For example, in Fig. 4.17(c) we show the intensity distribution of the bottom most right patch (in Tsukuba dataset) generated from the proposed rate allocation scheme with $S = 1$. We see that the proposed rate allocation scheme preserves the bimodal intensity distribution in the low contrast regions, while the standard rate allocation schemes fail to preserve the bimodal distribution in these regions as shown earlier in Fig. 4.7. Finally, for the Map dataset the disparity estimation results (see Table. 4.4) are very similar for the proposed, SPIHT and JPEG 2000 coding schemes due to the textured nature of the scene. For visual comparisons, we present the disparity image $\mathbf{m}^h$ and the absolute disparity error in Fig. 4.18 and Fig. 4.19 for the Tsukuba and Venus datasets respectively. We see that the proposed coding scheme with $S = 1$ permits to accurately compute the disparity value in the regions close to weak edges, without significant penalization in the regions proximity to strong edges. For example, the disparity values in the bottom most right portion of the Tsukuba image set is estimated accurately in the proposed scheme (with $S = 1$), since the bimodal intensity distribution is preserved as shown before. Also, for the same reason the slanted 3D surfaces in the Venus dataset are detected accurately when the images are encoded with the proposed scheme compared to the estimation from standard encoded images. Finally, from Fig. 4.20 we observe that our scheme competes with the standard coding solutions in terms of disparity accuracy for the Map dataset that presents rich textures.

We now study the influence of the coding parameter S on the disparity estimation results. Table 4.2

Table 4.4: Comparison of disparity error between the proposed, SPIHT, JPEG 2000 and JPEG coding schemes for the Map dataset. Size of the image: $N = 216 \times 284$. Results are tabulated using a wavelet decomposition with $k = 4$ levels.

Bit rate per view (bpp)	Disparity Error (%)			
	Proposed S = 1	SPIHT	JPEG 2000	JPEG
0.1	10.66	11	10.43	66
0.15	8.62	9.22	8.83	66
0.2	7.74	7.9	7.75	63
0.3	7.23	7.15	7.24	8.3



(a) (b) (c)

Figure 4.17: Effect of compression on the bimodal intensity distribution in the Tsukuba dataset. (a) Patch (marked in white) on which the histogram is generated. (b) Intensity distribution generated from original image I_1. (c) Intensity distribution generated from the proposed compressed image $\tilde{I}_1$ encoded at a bit rate of 0.4 bpp.

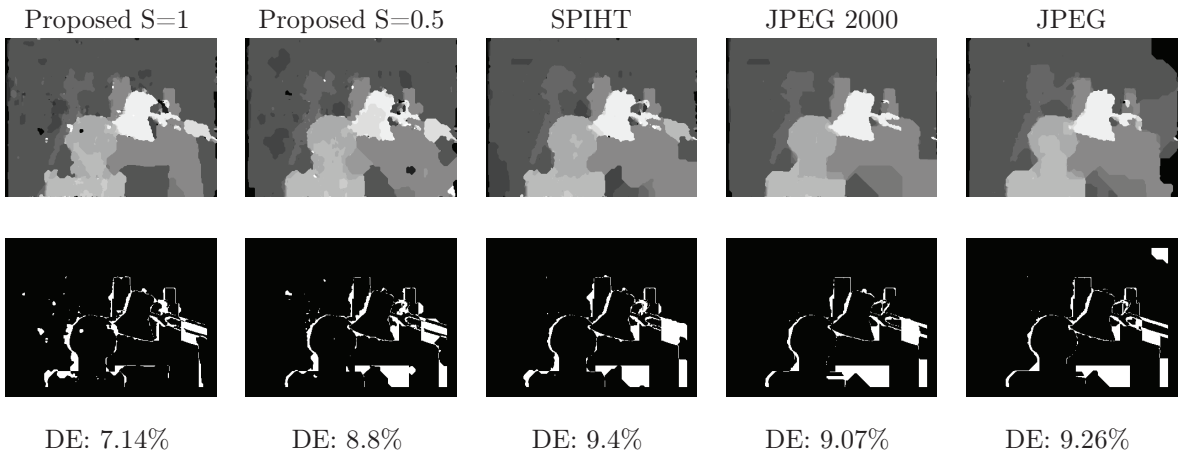


Figure 4.18: Comparison of the disparity image estimation between the proposed and the standard rate allocation schemes for the Tsukuba dataset. The disparity image is estimated from the compressed images encoded at 0.4 bpp per view. Top row: Estimated disparity image results. Bottom row: Corresponding absolute disparity errors. The white pixels denote errors in the disparity value.

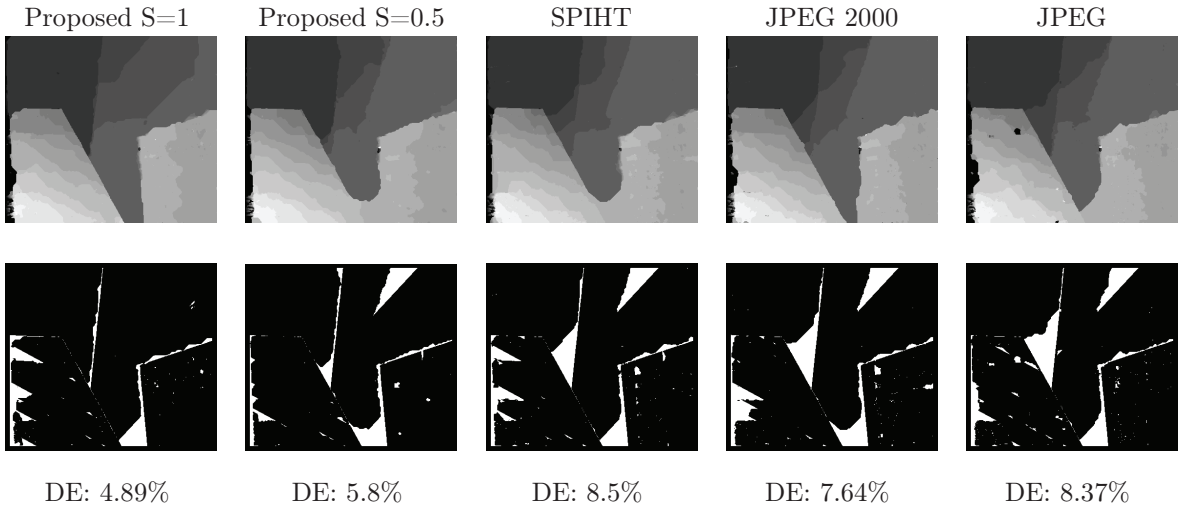


Figure 4.19: Comparison of the disparity image estimation between the proposed and the standard rate allocation schemes for the Venus dataset. The disparity image is estimated from the compressed images encoded at 0.6 bpp per view. Top row: Estimated disparity image results. Bottom row: Corresponding absolute disparity errors. The white pixels denote errors in the disparity value.

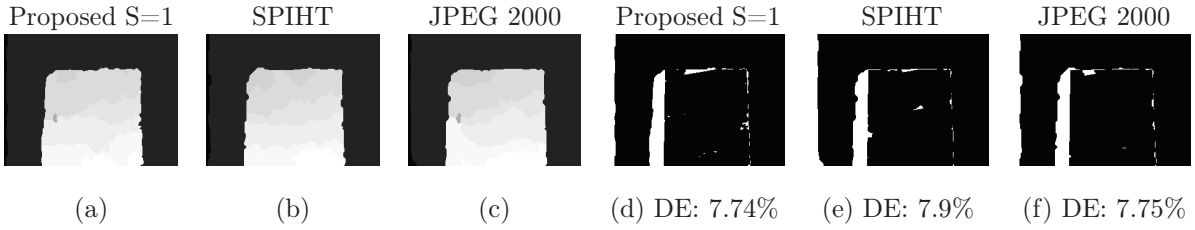


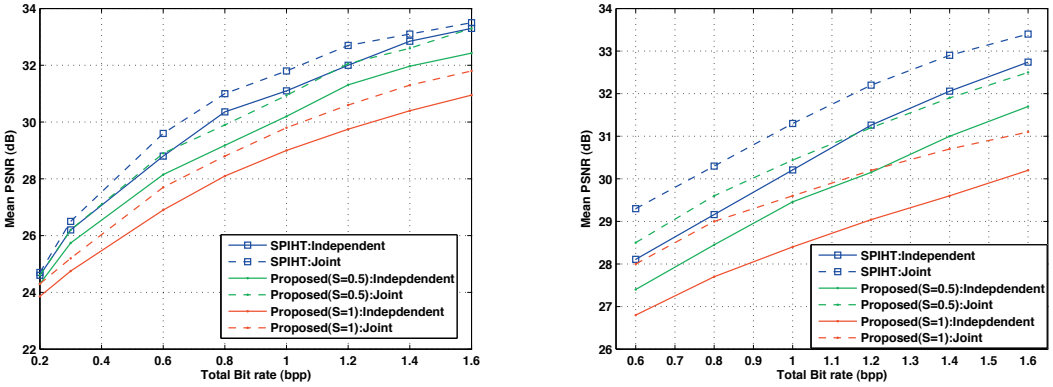
Figure 4.20: Comparison of the disparity image estimation between the proposed and standard rate allocation schemes for the Map dataset. The disparity image is estimated from the compressed images encoded at 0.2 bpp per view. (a), (b) and (c) represent the estimated disparity image results from proposed, SPIHT and JPEG 2000 coding schemes respectively. (d), (e) and (f) represent the corresponding absolute disparity error. The white pixels denote errors in the disparity value.

and Table 4.3 compare the disparity error for various choices of S for the Tsukuba and Venus datasets, respectively. From Table 4.2 and Table 4.3 we see that $S = 1$ gives the best disparity results due to the balanced encoding of low and high frequency components. It can be further noted that the disparity results are not better with increasing S due to the allocation of more bits to the low contrast regions. In order to illustrate this, we show the compressed images in Fig. 4.21(a), Fig. 4.21(b) and Fig. 4.21(c) that are encoded using SPIHT-based coding scheme and the proposed coding scheme with $S = 1$ and $S = 2$ respectively. From Fig. 4.21 we see that when S increases, our coding scheme allocates more bits to the low contrast regions (e.g., bottom most right patch) than to the texture regions (e.g., lamp and head). We now study the impact of the proposed rate allocation scheme on the quality of the compressed images $\tilde{I}_1$ and $\tilde{I}_2$ measured in terms of PSNR. Fig. 4.22(a) and Fig. 4.22(b) compare the RD performances between the proposed and SPIHT-based coding schemes for the Tsukuba and Venus datasets respectively. We see that the proposed coding scheme with $S = 1$ performs on the average 2 dB worse than the SPIHT-based coding scheme mainly due to the encoding of weak edges. It can be further noted that the quality of the

(a) PSNR: 30.36 (b) PSNR: 28.1 (c) PSNR: 22.7

Figure 4.21: Influence of the coding parameter S on the quality of the compressed image $\tilde{I}_1$ in the Tsukuba dataset. (a) SPIHT-based coding scheme; (b) proposed coding scheme with $S = 1$; (c) proposed coding scheme with $S = 2$. The images are encoded using a bit rate of 0.4 bpp.

compressed images $\tilde{I}_1$ and $\tilde{I}_2$ can be improved with $S = 0.5$. However, as demonstrated in Table 4.2 this penalizes the disparity estimation results. Similar behavior is observed on the other datasets. These results show that it is difficult to jointly maximize the quality of the compressed images and the quality of the disparity information estimated from the compressed images.



(a) (b)

Figure 4.22: Comparison of the joint reconstruction performances between the proposed and SPIHT-based rate allocation schemes: (a) Tsukuba dataset; (b) Venus dataset.

In the remaining of this section, we study the effect of using our accurate disparity estimation results in the view synthesis and joint reconstruction applications. We synthesize novel views by warping the approximated reference view $\tilde{I}_1$ using the associated disparity image $\mathbf{m}^h$. Table 4.5 and Fig. 4.23 show the prediction error for the novel views measured in terms of PSNR for the Tsukuba dataset and the Flower garden sequence, respectively. In these plots, the position of the synthesized viewpoint is numbered according to the distance from the reference image (viewpoint 1). We observe that in spite of accurate disparity image estimation, the quality of the rendered view is degraded in our scheme when compared to the SPIHT-based scheme especially for viewpoints closer to the reference image. The degradation is mainly due to the poor image quality of the reference view $\tilde{I}_1$ comparing to the SPIHT-based scheme as described previously. But, as the distances between the reference and synthesized view increases, we see that the performance gap between the two schemes decrease as shown in Fig. 4.23. When views are far apart, the proposed scheme

Table 4.5: Comparison of the prediction image quality between the proposed and SPIHT rate allocation schemes for the Tsukuba dataset. The viewpoints are rendered by warping the reference image $\tilde{I}_1$. Viewpoint 1 is closer to the reference image.

Bit rate per view (bpp)	Viewpoint 1			Viewpoint 2		
	Proposed S=1	Proposed S=0.5	SPIHT	Proposed S=1	Proposed S=0.5	SPIHT
0.3	26.27	27.21	27.7	24.95	25	25.1
0.4	27.11	27.95	28.71	25.4	25.6	25.65
0.5	27.85	28.58	29.2	25.53	25.64	25.76
0.6	28.26	29.47	29.78	25.72	25.88	25.96
0.7	28.56	29.8	30.15	25.94	26.08	26.09
0.8	29.07	30.02	30.5	25.98	26.13	26.09

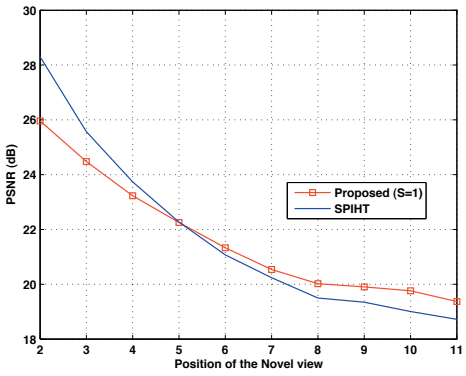
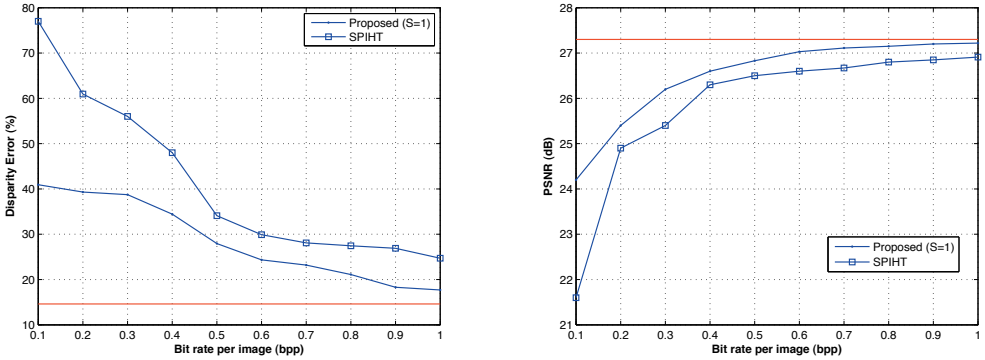


Figure 4.23: Comparison of the quality of the predicted synthesized views for various novel view positions for the proposed and SPIHT rate allocation schemes in the Flower Garden sequence. The view 2 is closer to the reference image and view 12 is the farthest. The quality of the reference view $\tilde{I}_1$ used for image prediction is 27 dB and 29.2 dB for the proposed and SPIHT-based schemes.

performs better than the SPIHT-based scheme due to the accurate estimation of the disparity image, even if the quality becomes pretty low. We finally analyze in Fig. 4.22, the benefit of using accurate disparity estimation for the joint reconstruction as proposed in Section 4.2.3. The joint reconstruction performances based on SPIHT, and proposed coding schemes with $S = 0.5$ and $S = 1$ are denoted in *dotted* lines. The corresponding independent coding performances are denoted in *solid lines*. We use the Bjontegaard metric [146] to calculate the average difference in the RD performances between a particular joint coding scheme and the corresponding independent coding scheme. For the Tsukuba dataset, the average quality improvement due to joint reconstruction is found out to be 0.81 for the SPIHT, and 0.87 and 0.98 dB respectively for the proposed rate allocation scheme with $S = 0.5$ and $S = 1$. Similarly, for the Venus dataset the gain is found to be 1.1, 1.16 and 1.29 dB respectively. We see that the proposed rate allocation methodology improves the average joint reconstruction image quality by 0.2 dB when compared to the joint reconstruction quality improvement in the SPIHT-based coding scheme. This highlights the benefit of using accurate disparity results for the image reconstruction. However, by comparing the *blue* and *red dotted* lines in Fig. 4.22, we infer that the proposed rate allocation scheme could not outperform the SPIHT-based coding scheme due to the compressed image quality penalization, as shown above. It can be noted that the overall joint reconstruction quality can be improved by reducing S (e.g., $S = 0.5$), however this comes at the price of a worse correlation estimation. Therefore, joint optimization of the image reconstruction quality and the correlation estimation would not be possible for a given bit rate, as both demand accurate encoding of

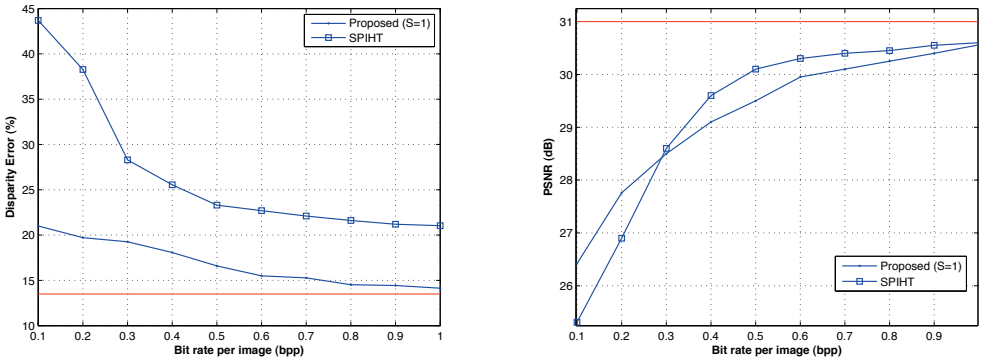


(a) (b)

Figure 4.24: Performance comparison between the proposed and SPIHT rate allocation scheme for the Oval spherical dataset: (a) disparity error; (b) prediction image quality. The red straight line corresponds to the disparity estimated using the original spherical images I_1 and I_2.

different image characteristics. In particular, the former representation requires encoding of strong edges and textures, while the latter requires encoding of smooth regions.

4.5.2 Spherical images



(a) (b)

Figure 4.25: Performance comparison between the proposed and SPIHT rate allocation scheme for the Room spherical dataset: (a) disparity error; (b) prediction image quality. The red straight line corresponds to the disparity estimated using the original spherical images I_1 and I_2.

We also illustrate the benefit of our proposed rate allocation methodology on the correlated Oval and Room spherical datasets. Fig. 4.24(a) and Fig. 4.25(a) compare the disparity error between the proposed rate allocation and SPIHT schemes at various encoding bit rates for the Oval and Room datasets respectively. The *red line* represents the disparity error that is computed from the original images without compression, therefore it represents benchmark performance. From Fig. 4.24(a) and Fig. 4.25(a) we see that the proposed rate allocation scheme with $S = 1$ leads to a better disparity map compared to the SPIHT-based rate

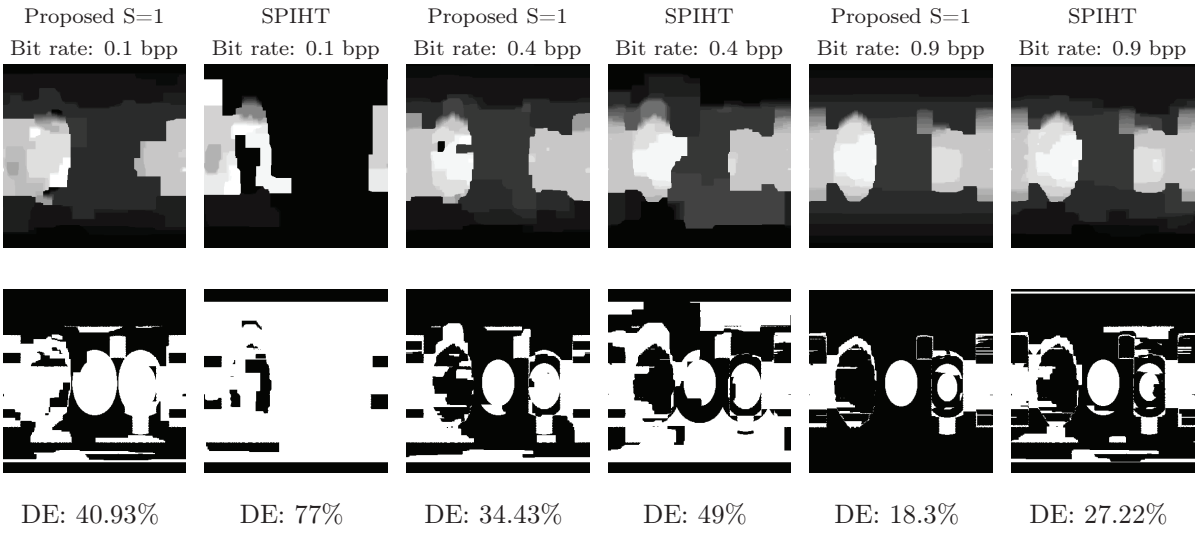
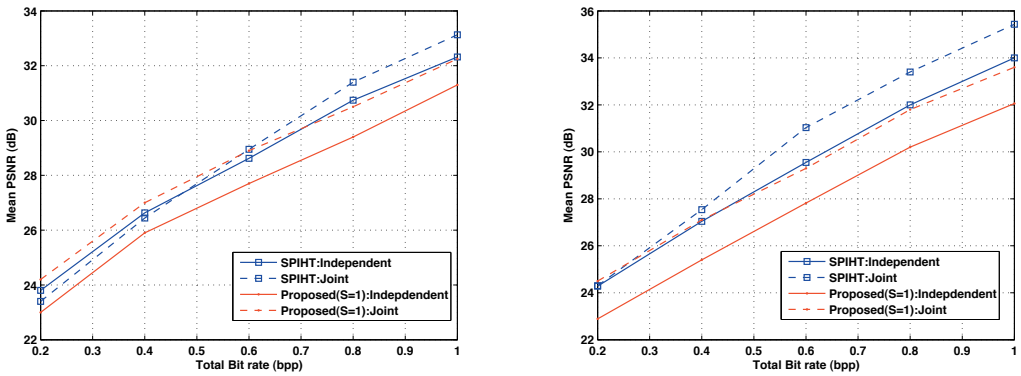


Figure 4.26: Oval spherical dataset: Comparison of the estimated disparity image between the proposed and SPIHT-based rate allocation schemes for the bit rates $0.1, 0.4$ and 0.9 bpp. Top row: Estimated disparity image results. Bottom row: Corresponding absolute disparity errors. The white pixels denote errors in the disparity value.

allocation methodology. That is consistent with our earlier observations in the planar datasets. For visual comparisons, we show in Fig. 4.26 the disparity results from the proposed and SPIHT rate allocation schemes at encoding rates of $0.1, 0.4$ and 0.9 bpp. From Fig. 4.26 we see that at a bit rate of 0.1 bpp, the proposed scheme computes a coarse disparity image, while the SPIHT scheme fails to estimate a meaningful disparity result. At high bit rates corresponding to 0.9 bpp, due to the balanced encoding of weak and strong edges, the proposed scheme estimates an accurate disparity image with DE=18.3%, that is very close to the disparity results estimated from the original images (see Fig. 4.24(a)). Therefore, it is clear that the proposed coding scheme with $S = 1$ identifies and encodes the necessary information in each view to reduce the distortion in the disparity image.



(a) (b)

Figure 4.27: Comparison of the joint reconstruction performances between the proposed and SPIHT-based rate allocation schemes: (a) Oval dataset; (b) Room dataset.

We then evaluate the accuracy of the disparity image based on the quality of image prediction, where we use the compressed view $\tilde{I}_1$ and the respective disparity estimation $\mathbf{m}^h$ to predict the second view. Fig. 4.24(b) and Fig. 4.25(b) compare the quality of the predicted view at various encoded bit rates for both datasets respectively. In spite of poor compressed reference image $\tilde{I}_1$ quality, the proposed scheme outperforms the image prediction quality of the SPIHT scheme due to the accurate disparity estimation. This illustrates the importance of estimating an accurate correlation information in omnidirectional sensor networks. We finally analyze in Fig. 4.27 the benefit of using accurate disparity estimation for the joint reconstruction of the images as proposed in Section 4.2.3. We compute the average RD improvement between the joint reconstruction schemes implemented with SPIHT and proposed rate allocation schemes using the Bjontegaard metric [146]. The average joint reconstruction quality improvement compared to the respective independent reconstruction is found out to be 0.1 and 1.2 dB respectively for the Oval dataset. For the Room dataset, it is found out to be 0.6 and 1.7 dB respectively. Due to the better correlation estimation, we see that the proposed scheme improves the quality of the compressed views by a margin of 1.1 dB when compared to the SPIHT-based joint reconstruction scheme. In spite of 1.1 dB improvement in the joint reconstruction, we see from Fig. 4.27 that the proposed scheme cannot outperform the SPIHT-based joint representation scheme in terms of image quality. This is consistent with our earlier observations in the planar datasets. Therefore, under fixed bit budget we can either optimize image reconstruction or depth estimation as they represent different optimization objectives.

4.6 Conclusions

In this chapter, we have first contributed a novel rate balanced distributed representation scheme for compressing a pair of correlated images captured in perspective or omnidirectional camera networks. We have developed a new joint decoding algorithm based on a constrained optimization problem that permits to improve the reconstruction quality by exchanging information between images. We have shown that our joint reconstruction problem is convex and we have proposed to solve it effectively using proximal methods. We have shown experimentally that the proposed joint decoding algorithm outperforms the independent coding solutions for both planar and spherical datasets. However, at low bit rates the benefit of joint reconstruction is not noticeable due to the poor correlation estimation from the highly compressed images.

We have then contributed a novel rate allocation scheme that preserves the visual information in the low contrast regions in order to improve the quality of disparity estimation. Extensive simulations carried out on planar and omnidirectional datasets confirm that the proposed rate allocation scheme brings a better disparity image estimation compared to one that is estimated from standard coding solutions. We then show that the improved disparity estimation however comes at a price of image quality penalization. For a given bit budget, we have shown that there exists an actual trade-off between the quality of image reconstruction and the accuracy of the correlation estimation. Therefore, our result is highly beneficial to distributively compress the visual information in camera networks that target a particular application such as distributed scene analysis (e.g., object detection) or distributed image coding. In the next chapter, we replace the traditional imaging cameras with the CS-cameras and we propose a distributed representation scheme that estimates a correlation model from the quantized linear measurements.

Chapter 5

Asymmetric Distributed Representation from Linear Measurements

5.1 Introduction

In the previous chapter, we have presented a distributed representation scheme for compressing the correlated images acquired by traditional cameras. In such systems, the entire image is acquired prior to compression and the encoding complexity grows with the resolution of image. In this chapter, we replace the traditional cameras with CS-based cameras that directly acquires the compressed image in the form of quantized linear measurements.

We consider the problem of estimating the correlation model between images where the common objects in different images are displaced due to the viewpoint change or motion of the scene objects. We concentrate on the particular problem where one image is selected as the reference image and it is used as a side information for decoding the other correlated images[1]. These compressed images are built on random measurements that are further quantized and entropy coded. One solution in such settings is to reconstruct all the compressed frames from the respective linear measurements by solving an l_2-TV or l_2-l_1 optimization problem (see Chapter 2), and then estimate the correlation model from the reconstructed images. Unfortunately, reconstructing the compressed images based on solving an l_2-l_1 or l_2-TV optimization problem is computationally expensive. Also, the correlation model estimated from highly compressed images usually fails to capture the actual geometrical relationship between images, as described in Chapter 4. Motivated by these issues, we estimate in this chapter a robust correlation model directly in the compressed domain without explicitly reconstructing all the compressed images.

We propose a geometry-based correlation model to describe the common information in image pairs. We assume that the constitutive components of natural images can be captured by visual features that undergo local geometric transformations in different images. We first estimate prominent visual features by computing a sparse approximation of the reference image with a dictionary of geometric basis functions. Then, we formulate a regularized optimization problem whose objective is to identify the features in the compressed images that correspond to the prominent components in the reference image. Correspondences then define relative transformations between images that form the geometric correlation model. A regularization constraint ensures that the estimated correlation is consistent and correspond to the actual motion of visual objects. We then use the estimated correlation in a new joint decoding algorithm that approximates the multiple images with the help of the estimated correlation. The joint decoding is cast as an optimization problem that warps the reference image according to the transformation described in the correlation infor-

[1]Part of this work has been submitted to: V. Thirumalai and P. Frossard, "Distributed Representation of Geometrically Correlated Images with Compressed Linear Measurements," IEEE Trans. on Image Proc., 2011.

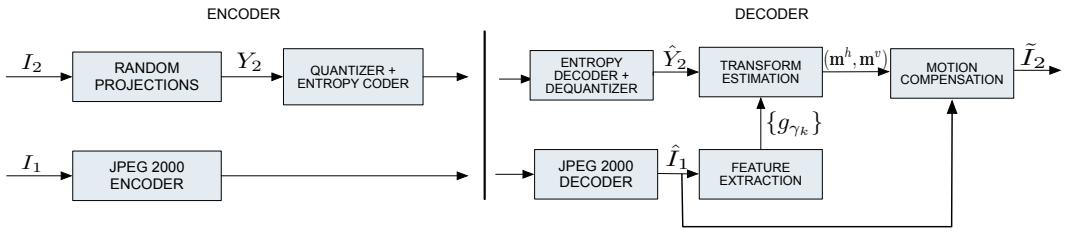


Figure 5.1: Schematic representation of the proposed scheme. The images I_1 and I_2 are correlated through displacement of scene objects, due to viewpoint change or motion of scene objects.

mation, while enforcing the decoded images to be consistent with the quantized measurements. We finally propose an extension of our algorithm to the joint decoding of multi-view images.

We analyze the performance of our novel framework on distributed video coding and multi-view imaging applications. We show by experiments that the proposed algorithm computes a good estimation of the correlation between multiple images captured in multi-view and video imaging scenarios. In particular, the results confirm that dictionaries based on geometric basis functions permit to capture the correlation more efficiently than a dictionary built on patches or blocks from the reference image as proposed in [103]. In addition, we show that the estimated correlation model can be used to decode the compressed image by disparity or motion compensation. Such a decoding strategy permits to outperform independent coding solutions based on JPEG 2000 and state-of-the-art distributed coding schemes based on disparity/motion learning [86, 147] in terms of rate-distortion (RD) performance computed on the compressed images. Finally, the experiments outline that the consistent prediction term proves to be very effective in increasing the decoding quality of the images given by the quantized linear measurements.

The rest of this chapter is organized as follows. The geometric correlation model used in our framework is presented in Section 5.2. Section 5.3 describes the proposed regularized energy model for an image pair. The proposed optimization algorithm is presented in Section 5.4 and the consistent image prediction algorithm is presented in Section 5.5. Section 5.6 describes the extension of our scheme to cases with more than two images. Experimental results in distributed multi-view and video imaging are finally given in Section 5.7.

5.2 Framework

5.2.1 Overview

We first describe our framework for a pair of images, and then the extension to more images is presented in Section 5.6. We consider a pair of images I_1 and I_2 (with dimensions $N = N_1 \times N_2$) that represent a scene at different time instants or from different viewpoints; these images are correlated through the motion of visual objects. These images are encoded independently and transmitted to a joint decoder. The joint decoder estimates the relative motion or disparity between the received signals and jointly decodes the images. The framework is illustrated in Fig. 5.1.

We focus on the particular problem where one of the images serves as a reference for the correlation estimation and decoding of the second image. While the reference image I_1 could be encoded with any coding algorithm (e.g., JPEG, compressed sensing framework [30]), we choose here to encode the reference image I_1 based on JPEG 2000. Next, we concentrate on the independent coding and joint decoding of the second image, where the first image $\hat{I}_1$ serves as the side information. At encoder, the second image I_2 is projected on a random matrix Φ to generate the measurements $Y_2 = \Phi I_2$. The measurements Y_2 are quantized with a uniform quantization algorithm. Finally, the quantized linear measurements are encoded with an entropy coder (e.g., arithmetic coder).

The joint decoder first computes the sparse approximation of the image $\hat{I}_1$ using features in a parametric dictionary of geometric functions. Such an approximation captures the most prominent geometrical components that represent the visual information in the image $\hat{I}_1$. Given the most prominent geometrical features in the image $\hat{I}_1$, we estimate the corresponding features in the second image I_2 from quantized linear measurements $\hat{Y}_2$. In particular, the corresponding visual features in both images are related using a geometry-based correlation model, where the correspondences describe local geometric transformations between images. We generate the horizontal $\mathbf{m}^h$ and vertical $\mathbf{m}^v$ components of the dense motion field that represents the correlation from the geometric transformations of the visual features between images. This motion information $(\mathbf{m}^h, \mathbf{m}^v)$ is further used to predict the compressed image $\tilde{I}_2$ from the reference image $\hat{I}_1$. We further ensure a consistent prediction of $\tilde{I}_2$ by explicitly considering the quantized measurements $\hat{Y}_2$ during the warping process. Before getting into the details of the correlation estimation algorithm, we describe the sparse image approximation algorithm and the geometry-based correlation model built on a parametric dictionary.

5.2.2 Sparse image approximation

We discuss here the sparse approximation of images using geometric basis functions in a structured dictionary. We propose to represent the images by a sparse linear expansion of geometric function g_γ taken from a parametric and overcomplete dictionary $\mathcal{D} = \{g_\gamma\}$. The geometric function g_γ in the dictionary $\mathcal{D}$ is usually called *atom*. The dictionary is constructed by applying a set of geometric transformations to a generating function g. These geometric transformations can be represented by a family of unitary operators $U(\gamma)$, so that the dictionary takes the form $\mathcal{D} = \{g_\gamma = U(\gamma)g, \gamma \in \Gamma\}$ for a given set of transformation indexes Γ. Typically, this transformation set consists of scales s_x, s_y, rotation θ, and translations t_x, t_y operators, defined as

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1/s_x & 0 \\ 0 & 1/s_y \end{bmatrix} \begin{bmatrix} cos\ \theta & sin\ \theta \\ -sin\ \theta & cos\ \theta \end{bmatrix} \begin{bmatrix} m - t_x \\ n - t_y \end{bmatrix},$$

where (m, n) define the image coordinates. Thus, each of the transformations or equivalently each atom in $\mathcal{D}$ is indexed by five parameters denoted by $\gamma = (t_x, t_y, \theta, s_x, s_y)$.

We can write the linear approximation of the reference image $\hat{I}_1$ with functions in $\mathcal{D}$ as

$$\hat{I}_1 \approx \sum_{i=1}^{K} c_i\ g_{\gamma_i}, \tag{5.1}$$

where $\{c_i\}$ represents the coefficient vector. The number of atoms K used in the approximation of $\hat{I}_1$ is usually much smaller than the dimensions of the image. It should be noted that in our framework we are interested in approximating the reconstructed version of the first image $\hat{I}_1$ using the functions in the geometric dictionary $\mathcal{D}$.

Given the dictionary $\mathcal{D}$, finding the sparse approximation of the image using K atoms as given in Eq. (5.1), is usually a NP-hard problem. Several suboptimal algorithms exist in the literature (e.g., Basis Pursuit, Matching Pursuit) in order to find and capture the most salient and prominent visual features in the images, in a reasonable computational time. In this work, we have chosen Matching Pursuit (MP) [148], which greedily estimates the K atoms $\{g_{\gamma_i}\}$ that best match the image $\hat{I}_1$ [38]. We briefly discuss how the MP selects a set of K atoms from the dictionary $\mathcal{D}$. Initially, MP sets the residual signal $r = \hat{I}_1$. Then, an atom g_{γ_i} from the dictionary $\mathcal{D}$ that best matches the residual signal r is chosen. The algorithm then updates the residual as $r = r - \langle r, g_{\gamma_i} \rangle g_{\gamma_i} = r - c_i g_{\gamma_i}$; this removes the component g_{γ_i} from the residual signal r. The atom selection and the residue update steps are repeated K times that eventually results in a set of K atoms $\{g_{\gamma_i}\}$ and coefficients $\{c_i\}$.

5.2.3 Joint correlation model

We describe briefly the geometric correlation model between images I_1 and I_2 in the rest of this section. For more details, we encourage the reader to refer to [70, 92]. We first compute the set of functions $\{g_{\gamma_i}\}$ that form the sparse approximation of the reconstructed version of the reference image $\hat{I}_1$ (see Eq. (5.1)). Under the assumption that the images are correlated based on local geometric transforms of their principal components, the second image I_2 could be approximated with transformed versions of the atoms used in the approximation of $\hat{I}_1$. We can thus approximate I_2 as

$$I_2 \approx \sum_{i=1}^{K} c_i \; F^i(g_{\gamma_i}), \tag{5.2}$$

where $F^i(g_{\gamma_i})$ represents a geometrical transformation of the atom g_{γ_i}. For consistency, we have described the correlation model given in Eq. (5.2) using the functions $\{g_{\gamma_i}\}$ that are picked from the reconstructed image $\hat{I}_1$. However, this correlation model holds well even if the atoms $\{g_{\gamma_i}\}$ are picked from the original image I_1. Due to the parametric form of the dictionary, the effect of F^i generally corresponds to a geometrical transformation of the atom g_{γ_i} that results in another atom in the same dictionary $\mathcal{D}$. Therefore, it is interesting to note that the transformation F^i on g_{γ_i} boils down to a transformation $\delta\gamma$ of the atom parameters, i.e.,

$$F^i(g_{\gamma_i}) = U(\delta\gamma)g_{\gamma_i} = U(\delta\gamma \circ \gamma_i)g = g_{\delta\gamma\circ\gamma_i} = g_{\gamma_i'}. \tag{5.3}$$

This means that the same dictionary can be used to represent the reference image $\hat{I}_1$ and the compressed images.

The main challenge in the joint decoder consists in estimating the local geometrical transformation F^i for each of the atoms g_{γ_i} in $\hat{I}_1$ from the compressed linear measurements $\hat{Y}_2$. We formulate in the next section a regularized optimization problem in order to estimate F^i, or equivalently the relative motion or disparity between images I_1 and I_2.

5.3 Correlation estimation with reference image

Given the set of K atoms $\{g_{\gamma_i}\}$ that approximate the first image $\hat{I}_1$, the disparity or motion estimation problem consists in finding the corresponding visual patterns in the second image I_2, while the latter is given only by the compressed random measurements $\hat{Y}_2$. This is equivalent to finding the correlation between images I_1 and I_2 with the joint sparsity model based on local geometrical transformations described in Section 5.2.3. We describe here the proposed regularized optimization framework that leads to the estimation of the correlation between images I_1 and I_2.

5.3.1 Regularized energy function

We are looking for a set of K atoms in I_2 that corresponds to the K visual features $\{g_{\gamma_i}\}$ selected in the first image. We denote their parameters by Λ, where $\Lambda = (\gamma_1', \gamma_2', \ldots, \gamma_K')$ for some γ_i', $\forall i, 1 \leq i \leq K$. We propose to select this set of atoms in a regularized energy minimization framework as a trade-off between the set that well approximates I_2 and the set that results in smooth local transformations between images. The energy model E proposed in our scheme is expressed as

$$E(\Lambda) = E_d(\Lambda) + \alpha_1 E_s(\Lambda), \tag{OPT-1}$$

where E_d and E_s represent data and smoothness terms respectively; α_1 is the regularization parameter that balances the data and smoothness terms. The solution of the correlation estimation problem is given by the

set of K atom parameters Λ^* that minimizes the energy E, i.e.,

$$\Lambda^* = arg\min_{\Lambda \in S} E(\Lambda), \tag{5.4}$$

where S represents the search space given by

$$S = \{(\gamma_1', \gamma_2', \ldots, \gamma_K') \mid \gamma_i' = \delta\gamma_i \circ \gamma_i, 1 \leq i \leq K, \delta\gamma \in \mathcal{L}\}, \tag{5.5}$$

where $\{\gamma_i\}$ are the parameters of the atoms in the reference image. The $\mathcal{L} \subset \mathbb{R}^5$ is given by $\mathcal{L} = [-\delta t_x,\ \delta t_x] \times [-\delta t_y,\ \delta t_y] \times [-\delta\theta,\ \delta\theta] \times [-\delta s_x,\ \delta s_x] \times [-\delta s_y,\ \delta s_y]$, where $\delta t_x, \delta t_y, \delta\theta, \delta s_x, \delta s_y$ determine the search window size corresponding to the translation parameters t_x, t_y, rotation θ, and scales s_x, s_y respectively. We describe below the two cost functions used in OPT-1.

5.3.2 Data cost function

The data fidelity term computes in the compressed domain, the accuracy of the sparse approximation of the second image with geometric atoms. The decoder receives the measurements $\hat{Y}_2$ that are computed by the quantized projections of I_2 onto a sensing matrix Φ. For each set of K atom parameters $\Lambda = \{\gamma_i'\}$, the data cost function E_d then reports the error between the measurements $\hat{Y}_2$ and the orthogonal projection of $\hat{Y}_2$ onto Ψ_Λ that is formed by the compressed versions of the atoms, i.e., $\Psi_\Lambda = \Phi[g_{\gamma_1'}|g_{\gamma_2'}|\ldots|g_{\gamma_K'}]$. It turns out that the orthogonal projection of $\hat{Y}_2$ is given as $\Psi_\Lambda \Psi_\Lambda^\dagger \hat{Y}_2$, where $\dagger$ represents the pseudo-inverse operator. More formally, the data cost is computed using the following relation:

$$E_d(\Lambda) = \|\hat{Y}_2 - \Psi_\Lambda \Psi_\Lambda^\dagger \hat{Y}_2\|_2^2 = \|\hat{Y}_2 - \Psi_\Lambda c\|_2^2. \tag{5.6}$$

The data cost function given in Eq. (5.6) first calculates the coefficients $c = \Psi_\Lambda^\dagger \hat{Y}_2$ and then measures the distance between the observations $\hat{Y}_2$ and $\Psi_\Lambda c$. In other words, the data cost function E_d accounts for the intensity variations between images by estimating the coefficients c of the warped atoms.

However, when the measurements are quantized, the coefficient vector c fails to properly account for the error introduced by the quantization. The quantized measurements only provide the index of the quantization interval containing the actual measurement value and the actual measurement value could be any point in the quantization interval. Let $Y_{2,p}$ be the p^{th} coordinate of the original measurement, and $\hat{Y}_{2,p}$ be the corresponding quantized value. It should be noted that the joint decoder has only access to the quantized value $\hat{Y}_{2,p}$, but not the original value $Y_{2,p}$. Henceforth, the joint decoder knows that the quantized measurement lies within the quantized interval, i.e., $\hat{Y}_{2,p} \in \mathcal{R}_{\hat{Y}_p} = (r_p,\ r_{p+1}]$, where r_p and r_{p+1} define the lower and upper bounds of the quantizer bin $\mathcal{Q}_p$. We propose to refine the data term by computing a coefficient vector $\tilde{c}$ as the best solution when considering all the valid measurement values in the quantization interval, i.e., $\tilde{Y}_2 \in \mathcal{R}_{\hat{Y}}$, where $\mathcal{R}_{\hat{Y}}$ is the Cartesian product of all the quantized regions $\mathcal{R}_{\hat{Y}_p}$, i.e., $\mathcal{R}_{\hat{Y}} = \prod_p \mathcal{R}_{\hat{Y}_p}$. The coefficients $\tilde{c}$ and the measurements $\tilde{Y}_2$ can be jointly estimated by solving the following optimization problem:

$$(\tilde{c}, \tilde{Y}_2) = arg\min_{\tilde{c}, \tilde{Y}_2} \|\tilde{Y}_2 - \Psi_\Lambda \tilde{c}\|_2^2 \ \text{ s.t. } \ \tilde{Y}_2 \in \mathcal{R}_{\hat{Y}}. \tag{5.7}$$

By following the steps in Proposition 1 described in Chapter 4, it can be easily shown that the Hessian of the objective function $h(\tilde{Y}_2, \tilde{c}) = \|\tilde{Y}_2 - \Psi_\Lambda \tilde{c}\|_2^2$ is positive semidefinite, i.e., $\nabla^2 h \succeq 0$, and hence the objective function h is convex (not strictly). Also the region $\mathcal{R}_{\hat{Y}}$ forms a closed convex set, as each region $\mathcal{R}_{\hat{Y}_p} = (r_p,\ r_{p+1}], \forall p$ forms a convex set. Therefore, the optimization problem given in Eq. (5.7) is convex, but not strictly convex as the Hessian of the objective function is not positive definite. Finally, the data

fidelity term given in Eq. (5.6) can be modified with the estimated coefficients $\tilde{c}$ and the measurements $\tilde{Y}_2$ as

$$\tilde{E}_d(\Lambda) = \|\tilde{Y}_2 - \Psi_\Lambda \tilde{c}\|_2^2, \tag{5.8}$$

where $\tilde{E}_d(\Lambda)$ represents the robust data function that efficiently accounts for the quantization noise.

5.3.3 Smoothness cost function

The goal of the smoothness term E_s is to penalize the atom transformations such that they result in coherent transformations of neighbor atoms. In other words, the atoms in the neighborhood are likely to undergo similar transformations, when the correlation between images is due to object or camera motion. Instead of penalizing directly the transformations $\{F^i\}$ (or equivalently $\{\delta\gamma\}$) such that they tend to be coherent for neighbor atoms, we propose to generate a dense motion field from the atom transformations and to penalize the motion (or disparity) field such that it is coherent for adjacent pixels. This regularization is easier to handle than a regular set of transformations F^i and directly corresponds to the physical constraints that explain the formation of correlated images. For a given transformation field $\mathbf{f} = (\mathbf{f}^h, \mathbf{f}^v, \mathbf{f}^\theta, \mathbf{f}^a, \mathbf{f}^b)$, we compute the horizontal $\mathbf{m}^h$ and vertical $\mathbf{m}^v$ components of the motion field as

$$\begin{bmatrix} \mathbf{m}^h(\mathbf{z}) \\ \mathbf{m}^v(\mathbf{z}) \end{bmatrix} = \begin{bmatrix} m(\mathbf{z}) - t_x(\mathbf{z}) \\ n(\mathbf{z}) - t_y(\mathbf{z}) \end{bmatrix} - \underbrace{\begin{bmatrix} \mathbf{f}^a(\mathbf{z}) & 0 \\ 0 & \mathbf{f}^b(\mathbf{z}) \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} cos(\mathbf{f}^\theta(\mathbf{z})) & sin(\mathbf{f}^\theta(\mathbf{z})) \\ -sin(\mathbf{f}^\theta(\mathbf{z})) & cos(\mathbf{f}^\theta(\mathbf{z})) \end{bmatrix}}_{R} \underbrace{\begin{bmatrix} m(\mathbf{z}) - t_x(\mathbf{z}) - \mathbf{f}^h(\mathbf{z}) \\ n(\mathbf{z}) - t_y(\mathbf{z}) - \mathbf{f}^v(\mathbf{z}) \end{bmatrix}}_{T}, \tag{5.9}$$

where $(m(\mathbf{z}), n(\mathbf{z}))$ represent the Euclidean coordinates, and $t_x(\mathbf{z})$ and $t_y(\mathbf{z})$ represent the translation parameters at location $\mathbf{z}$. The matrices S, R and T represent the grid transformations due to scale, rotation and translation changes respectively. Finally, the smoothness cost E_s in OPT-1 is computed as

$$E_s = \sum_{\mathbf{z},\mathbf{z}' \in \mathcal{N}} V_{\mathbf{z},\mathbf{z}'} = \sum_{\mathbf{z},\mathbf{z}' \in \mathcal{N}} \min\left(|\mathbf{m}^h(\mathbf{z}) - \mathbf{m}^h(\mathbf{z}')| + |\mathbf{m}^v(\mathbf{z}) - \mathbf{m}^v(\mathbf{z}')|, \tau\right), \tag{5.10}$$

where $\mathbf{z}$ and $\mathbf{z}'$ are the adjacent pixel locations, and $\mathcal{N}$ represents a usual 4-pixel neighborhood. The parameter τ sets a maximum limit to the smoothness penalty term, and thus helps to preserve the discontinuities in the motion field [130].

Now, we describe the methodology to estimate the dense transformation field $\mathbf{f}$ from the sets of atom transformations. Given the i^{th} pair of atoms g_{γ_i} and $g_{\gamma_i'}$ with $\gamma_i = (t_x, t_y, \theta, s_x, s_y)$ and $\gamma_i' = (t_x', t_y', \theta', s_x', s_y')$ in the images I_1 and I_2 respectively, we first calculate the local transformation captured by these atoms given by

$$\delta\gamma_i = (t_x - t_x', t_y - t_y', \theta - \theta', s_x'/s_x, s_y'/s_y). \tag{5.11}$$

It should be noted that all the pixels in the support of the atom g_{γ_i} take the transformation value $\delta\gamma_i$, i.e., $\mathbf{f}(\mathbf{z}) = \delta\gamma_i, \forall \mathbf{z} \in \mathcal{Z}_i$, where $\mathcal{Z}_i$ denotes the set of pixels in the support of the atom g_{γ_i} given as

$$\mathcal{Z}_i = \{\mathbf{z} = (m, n) \mid |g_{\gamma_i}(m, n)| > \epsilon\}, \tag{5.12}$$

where $\epsilon > 0$ is a constant. Using a similar process (see Eq. (5.11)) a local transformation is established for all the atom pairs. Then, the transformations $\{\delta\gamma_k\}$ captured by the K pairs of atoms are fused together to estimate a global transformation field $\mathbf{f}$. For a given location $\mathbf{z}$, we first assign relative weights $\{w_{\mathbf{z}}^{(k)}\}$ to each candidate transformation $\delta\gamma_i$ based on the response of the i^{th} atom at the pixel location $\mathbf{z}$. Then, the fusion process is simply implemented by choosing the most confident transformation for each pixel position $\mathbf{z}$. Mathematically, we write the transformation at position $\mathbf{z}$ as

$$\mathbf{f}(\mathbf{z}) = \delta\gamma_{\hat{k}(\mathbf{z})}, \tag{5.13}$$

where $\hat{k}(\mathbf{z}) = arg \max_{i=1,2,\ldots,K} w_{\mathbf{z}}^{(i)}$, and $w_{\mathbf{z}}^{(i)}$ is the response of the i^{th} atom at the location $\mathbf{z}$, i.e., $w_{\mathbf{z}}^{(i)} = g_{\gamma_i}(\mathbf{z}) = g_{\gamma_i}(m,n)$. We denote this methodology as MAX in this chapter. The simple MAX-based scheme however fails to effectively consider all the transformation values $\{\delta\gamma_{k(\mathbf{z})}\}$ for the given pixel $\mathbf{z}$; this usually results in a suboptimal solution. In order to estimate a better solution we have proposed energy minimization problems in Appendix A.1, which effectively consider all the transformation values $\{\delta\gamma\}$ in the given pixel location $\mathbf{z}$. However, in this chapter we use MAX approach to generate a transformation field since it provides a good trade-off between the complexity and the performance (see Appendix A.1 for more details).

5.4 Optimization algorithms

In this section, we describe the optimization methodology that is used to solve OPT-1. Recall that our objective is to assign a transformation F^i to each atom g_{γ_i} in the reference image in order to build a set of smooth local transformations that is consistent with the quantized measurements $\hat{Y}_2$. The candidate transformations are chosen from a finite set of labels $\mathcal{L} = \mathcal{L}_x \times \mathcal{L}_y \times \mathcal{L}_\theta \times \mathcal{L}_a \times \mathcal{L}_b$, where $\mathcal{L}_x$, $\mathcal{L}_y$, $\mathcal{L}_\theta$, $\mathcal{L}_a$ and $\mathcal{L}_b$ refer to the label sets corresponding to translations along x, y direction, rotations and scales, respectively (see Eq. (5.5)). One could use an exhaustive search on the entire label set $\mathcal{L}$ to solve OPT-1. However, the cost for such a solution is high as the size of the label set $\mathcal{L}$ grows exponentially with the size of the search windows for all the transformation parameters. Rather than doing an exhaustive search, we propose two optimization algorithms based on local and global techniques.

5.4.1 Local optimization

The local optimization algorithm estimates the transformation F^i iteratively, by changing successively each of the K atom parameters γ_i by one increment in the parameter space. We focus on the search space that is given by perturbing the transformations of each atom parameter, i.e., $t_x \pm \Delta t_x$, $t_y \pm \Delta t_y$, $\theta \pm \Delta\theta$, $s_x \pm \Delta s_x$ and $s_y \pm \Delta s_y$ for each atom γ_i. The differential parameters $\Delta t_x, \Delta t_y, \Delta\theta, \Delta s_x$ and Δs_y represent a unit change in the translation, rotation and scale components respectively. We first initialize the algorithm with zero motion, i.e., the set of atoms $\{g_{\gamma_i}\}$ generated from $\hat{I}_1$ are used in the first iteration, $\gamma_i' = \gamma_i, \forall i$ where $1 \leq i \leq K$, and the search space S^0 is formed using

$$S^0 = \{(\gamma_1', \gamma_2', \ldots, \hat{\gamma}_i', \ldots, \gamma_K') | \hat{\gamma}_i' = (t_x^i \pm \Delta t_x, t_y^i \pm \Delta t_y, \theta^i \pm \Delta\theta, s_x^i \pm \Delta s_x, s_y^i \pm \Delta s_y), 1 \leq i \leq K\} \subset S. \tag{5.14}$$

We then calculate the energy E in OPT-1 for the set of K atoms in the search space S^0. Once the energy E is computed for atoms in S^0, we select the parameters $\Lambda^0 = (\gamma_1^0, \gamma_2^0, \ldots, \gamma_K^0)$ corresponding to the minimum energy. Then, a new search space S^1 is formed similarly to the definition in Eq. (5.14) with the current solution Λ^0 as reference. Such a procedure is repeated by successively constructing a new search space S^k on the solution Λ^{k-1} from the previous iteration of the algorithm. The algorithm stops when convergence is attained (or till it reaches a maximum number of iterations). The proposed algorithm is guaranteed to converge. Let E_0 be the initial energy, i.e., the energy corresponding to set of parameters $\gamma_i' = \gamma_i, \forall i$ where $1 \leq i \leq K$. If E_k is the minimal energy computed at step k of the algorithm, we clearly have $E_k \leq E_{k-1}$, as the search space S^k includes the best set of parameters Λ^{k-1} from the previous iterations. The energy decreases at every iteration till it reaches a local or global minimum E_{min}. The proposed optimization scheme thus converges and provides a (suboptimal) solution with tractable computational complexity to the estimation of correlation between images. The algorithm that bears some resemblance to a gradient descent solution is summarized in Algorithm 4. Finally, the data cost in the 8^{th} line of Algorithm 4 can be replaced by the robust data cost term $\tilde{E}_d$ as given in Eq. (5.8). We show later that the performance of our scheme can be improved using the robust data cost term $\tilde{E}_d$.

Algorithm 4 Local optimization: Correlation estimation with OPT-1

Input $K, \alpha_1, \delta t_x, \delta t_y, \delta\theta, \delta s_x, \delta s_y$
Generate $\{g_{\gamma_i}\}$ from $\hat{I}_1$ s.t. $\hat{I}_1 \approx \sum_{i=1}^{K} c_i \; g_{\gamma_i}$
Initialize $\Lambda^{-1} = \{\gamma_i\}$, $k = 0$
repeat
Generate index search space S^k based on Λ^{k-1} (with Eq. (5.14))
for all Parameter vectors Λ in S^k **do**
    Compute the motion field
    Compute the data term $E_d(\Lambda)$ with Eq. (5.6)
    Compute the smoothness term $E_s(\Lambda)$ with Eq. (5.10)
    Compute the global energy $E(\Lambda) = E_d(\Lambda) + \alpha_1 E_s(\Lambda)$
end for
$\Lambda^k = arg\min_{\Lambda \in S^k} E(\Lambda)$
$k = k + 1$
until convergence is reached

5.4.2 Global optimization

The local optimization algorithm described above estimates only a suboptimal solution as it does not guarantee to converge to strong local minima or global minima. We therefore propose strong optimization techniques based on Graph Cuts [130, 133] to solve OPT-1. Graph-based minimization techniques usually converge to strong local minima or global minima in a polynomial time with tractable computational complexity.

Recall that $\mathcal{Z}_i$ represents the set of pixels in the support of the atom g_{γ_i}, as defined in Eq. (5.12). Eq. (5.12) allows us to recast the atom-wise labeling problem, $f : g_{\gamma_i} \rightarrow l \in \mathcal{L}, \forall i$ that assigns one transformation label to each atom g_{γ_i}, as a pixel labeling problem where the objective is to assign a label to the set of pixels $\mathcal{Z}_i$, i.e., $f : \mathcal{Z}_i \rightarrow l \in \mathcal{L}, \forall i$. Our problem becomes therefore equivalent to a dense labeling problem that can be solved accurately using strong energy minimization techniques based on Graph Cuts [130, 133]. The cost of assigning a label $l_k \in \mathcal{L}$ to all pixels $\mathbf{z}$ in the support of atom g_{γ_i} (i.e., $\forall \mathbf{z} \in \mathcal{Z}_i$) can be computed using Eq. (5.6), where $\Lambda = (\gamma_1, \gamma_2, \ldots, \gamma_i + l_k, \ldots, \gamma_K)$. However, due to atom overlapping the pixels in the overlapped region could be assigned more than one label. In such cases, for the sake of simplicity we compute the cost corresponding to the atom index $\hat{k}$ (see Eq. (5.13)) that has the maximum atom response (see Section 5.3.3). Finally, the data cost term in OPT-1 can be replaced with the robust data term $\tilde{E}_d$ in order to provide robustness to quantization errors. The resulting optimization problem can be solved accurately using a Graph Cut algorithm.

5.5 Consistent image prediction by warping

Once the transformation between the correlated images has been estimated as described in the previous section, one can simply reconstruct an approximate version of the second image $\tilde{I}_2$ by warping the reference image $\hat{I}_1$ using a set of local transformations that forms the warping operator $\mathcal{W}_\Lambda$ (see Fig. 5.1). The resulting approximation is however not necessarily consistent with the quantized measurements $\hat{Y}_2$; the measurements corresponding to the projection of the image $\tilde{I}_2$ on the sensing matrix Φ are not necessarily equal to $\hat{Y}_2$. The consistency error might be significant, because the atoms used to compute the correlation and the warping operator do not optimally handle the texture information.

We therefore propose to add a consistency term E_t in the energy model of OPT-1 to form a new optimization problem. The consistency term enforces the image predicted through the warping operator $\mathcal{W}_\Lambda$ to be consistent with the quantized measurements. We define the additional term E_t as the l_2 norm error be-

tween the quantized measurements generated from the predicted image $\tilde{I}_2 = \mathcal{W}_\Lambda(\hat{I}_1)$ and the measurements $\hat{Y}_2$. The consistency term E_t is written as

$$E_t(\Lambda) = \|\hat{Y}_2 - \mathcal{Q}[\Phi\tilde{I}_2]\|_2^2 = \|\hat{Y}_2 - \mathcal{Q}[\Phi\mathcal{W}_\Lambda(\hat{I}_1)]\|_2^2, \tag{5.15}$$

where $\mathcal{Q}$ is the quantizer. It should be noted that, when the measurements are not quantized, the consistency term reads

$$E_t(\Lambda) = \|Y_2 - \Phi\mathcal{W}_\Lambda(\hat{I}_1)\|_2^2. \tag{5.16}$$

We then merge the three cost functions E_d, E_s and E_t with regularization constants α_1 and α_2 in order to form a new energy model E_R for consistent image prediction, expressed as

$$E_R(\Lambda) = E_d(\Lambda) + \alpha_1 E_s(\Lambda) + \alpha_2 E_t(\Lambda). \tag{OPT-2}$$

We now highlight the differences between the terms E_d and E_t used in OPT-2. The data cost E_d adapts the coefficient vector to consider the intensity variations between images but fails to properly consider the texture information. On the other hand, the consistency term E_t warps the atoms by considering the texture information in the reconstructed image $\hat{I}_1$ but fails to carefully consider the intensity variations between images. These two terms therefore impose different constraints on the atom selection that effectively reduce the search space. We experimentally show later that the quality of image $\tilde{I}_2$ is maximized, when all the three terms are activated in the optimization problem OPT-2. Finally, it should be noted that the data cost in OPT-2 can be replaced with the robust data cost $\tilde{E}_d$ given in Eq. (5.8) to properly account for the measurement quantization noise. It reads as

$$E_R(\Lambda) = \tilde{E}_d(\Lambda) + \alpha_1 E_s(\Lambda) + \alpha_2 E_t(\Lambda). \tag{OPT-2 (Robust)}$$

In order to solve OPT-2 or equivalently to estimate the correlation model that leads to consistent image prediction, we propose to use the same optimization methods as those described in Section 5.4. In both local and global optimization approaches we modify the objective functions to include the consistency term, since the main difference between the OPT-1 and OPT-2 is the inclusion of the consistent term E_t. The steps carried out to solve OPT-2 problem using local optimization methodology is summarized in Algorithm 5 where in *line 11* we compute the consistent term E_t energy based on warping the reference image (in addition to E_d and E_s given in lines 9 and 10). As described in the previous section, such an approach estimates a suboptimal solution with tractable computational complexity.

On the other hand, the global optimization technique based on Graph Cuts estimate a better solution as they usually converge to a strong local minima. In order to solve OPT-2 problem with Graph Cuts we need to compute the cost of assigning a label $l_k \in \mathcal{L}$ to each pixel $\mathbf{z}$. This follows from the fact that atom-wise labeling problem is recast as a pixel-wise labeling problem (see Section 5.4.2). The cost of assigning a label $l_k \in \mathcal{L}$ to all pixels $\mathbf{z}$ in the support of atom g_{γ_i} (i.e., $\forall \mathbf{z} \in \mathcal{Z}_i$ in Eq. (5.12)) is computed using Eq. (5.6) and Eq. (5.16) where $\Lambda = (\gamma_1, \gamma_2, \ldots, \gamma_i + l_k, \ldots, \gamma_K)$. The former one computes the cost of assigning a label $l_k \in \mathcal{L}$ to pixels $\mathcal{Z}_i$ based on data fidelity and the latter one computes the consistent cost of this label. However, as described earlier the pixels in the overlapped region could be assigned more than one label in the overlapping regions. In such cases, we take the value corresponding the atom index that has maximum response (see Eq. (5.13)). Finally, the data cost E_d in OPT-2 can be further replaced by the robust data cost term $\tilde{E}_d$ given in Eq. (5.8), and it can be solved using local or global optimization techniques described above. We show later that the performance of our scheme can be improved by using the robust data cost term $\tilde{E}_d$.

Algorithm 5 Local optimization: Correlation estimation with OPT-2

Input $K, \alpha_1, \alpha_2, \delta t_x, \delta t_y, \delta\theta, \delta s_x, \delta s_y$
Generate $\{g_{\gamma_i}\}$ from $\hat{I}_1$ s.t. $\hat{I}_1 \approx \sum_{i=1}^{K} c_i \; g_{\gamma_i}$
Initialize $\Lambda^{-1} = \{\gamma_i\}$, $k = 0$
repeat
Generate index search space S^k based on Λ^{k-1} (with Eq. (5.14))
for all Parameter vectors Λ in S^k **do**
    Compute the motion field
    Warp the reference image $\hat{I}_1$ using motion field, $\mathcal{W}_\Lambda(\hat{I}_1)$
    Compute the data term $E_d(\Lambda)$ with Eq. (5.6)
    Compute the smoothness term $E_s(\Lambda)$ with Eq. (5.10)
    Compute the consistency term $E_t(\Lambda)$ with Eq. (5.15)
    Compute the global energy $E_R(\Lambda) = E_d(\Lambda) + \alpha_1 E_s(\Lambda) + \alpha_2 E_t(\Lambda)$
end for
$\Lambda^k = arg\min_{\Lambda \in S^k} E_R(\Lambda)$
$k = k + 1$
until convergence is reached

Complexity Issues

We discuss now briefly the computational complexity of the correlation estimation algorithm, which can basically be divided into two stages. The first stage finds the most prominent features in the reference image using sparse approximations in a structured dictionary. The second stage estimates the transformation for all the features in the reference image by solving a regularized optimization problem such as OPT-1 or OPT-2. Our framework offers a very simple encoding stage with image acquisition based on random linear projections. The computational burden is shifted to the joint decoder, which can still trade-off the complexity and performance. The decoder is able to handle computationally complex tasks in our framework. However, the complexity of our system can be reduced in both stages compared to the generic implementation proposed above. For example, the complexity of the sparse approximation of the reference image can be reduced significantly using a tree-structured dictionary, without significant loss in the approximation performance [149]. In addition, a block-based dictionary can be used in order to reduce the complexity of the transformation estimation problem with block-based computations. Experiments show however that this comes at a price of a performance penalty in the image quality. Overall, it is clear that the decoding scheme proposed above offers high flexibility with an interesting trade-off between the complexity and performance. For example, one might decide to use the simple data cost E_d given in Eq. (5.6) even when the measurements are quantized; it leads to a simpler scheme but to reduced estimation accuracy.

5.6 Extension to multiple image sets

So far, we have focused on the distributed representation of pairs of images. Now, we describe the extension of our framework to datasets with J correlated images $I_1, I_2, \ldots, I_J$. We consider I_1 as the reference image; this image is given in a compressed form $\hat{I}_1$ and its prominent features are extracted at the decoder with a sparse approximation over the dictionary $\mathcal{D}$. The images $I_2, \ldots, I_J$ are sensed independently using the measurement matrix Φ. Their respective measurements $Y_2, \ldots, Y_J$ are quantized and entropy coded. Our framework applies to image sequences and multi-view imaging. For the sake of clarity, we focus on the latter framework where the multiple images captured from different viewpoints.

We are interested in estimating a depth map Z that captures the correlation between J images by assuming that the camera parameters are given a priori. The depth map is constructed using the set of K

features $\{g_{\gamma_i}\}$ in the reference image and the quantized measurements $\hat{Y}_2, \ldots, \hat{Y}_J$. We assume that the depth values Z are discretized such that the inverse depth $1/Z$ is uniformly sampled in the range $[1/Z_{max}, 1/Z_{min}]$, where Z_{min} and Z_{max} are the minimal and maximal depth in the scene, respectively [150]. The problem is equivalent to finding a set of labels $l \in \mathcal{L}$ that effectively captures depth information for each atom g_{γ_i} or pixel $\mathbf{z}$ in the reference image, where $\mathcal{L}$ is a discrete set of labels corresponding to different depth values. We propose to estimate the depth information with an energy minimization problem OPT-3, which includes three cost functions as follows:

$$E_u(\Lambda) = E_{d,u}(\Lambda) + \alpha_1 E_{s,u}(\Lambda) + \alpha_2 E_{t,u}(\Lambda), \tag{OPT-3}$$

where $E_{d,u}, E_{s,u}$ and $E_{t,u}$ represent the data, smoothness and consistency terms respectively. The three terms are balanced with regularization constants α_1 and α_2.

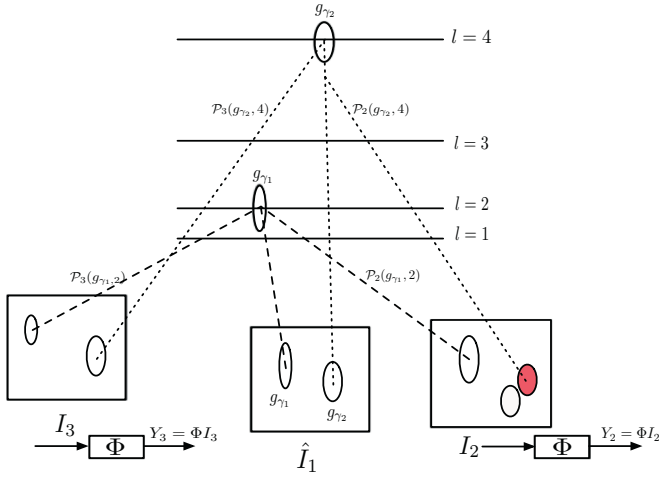


Figure 5.2: Illustration of the atom interactions in the multi-view imaging scenario. The original position of the features in all the images is marked in black color. Projection of the first feature g_{γ_1} at $l = 2$ in the views I_2 and I_3 corresponds to the actual position of the feature in the respective views, and thus forms a valid 3D region at $l = 2$. Meanwhile, the projection of the second feature g_{γ_2} at $l = 4$ corresponds to the actual position only in view I_3, but not in view I_2 (highlighted in red color). Hence, the second feature does not intersect at $l = 4$ which results in suboptimal depth estimation at $l = 4$.

The data term $E_{d,u}$ assigns a set of labels $l_1, l_2, \ldots, l_K$ respectively to the K atoms $g_{\gamma_1}, g_{\gamma_2}, \ldots, g_{\gamma_K}$ while respecting consistency with the quantized measurements. The data cost function reads as

$$E_{d,u}(\{g_{\gamma_i}\}, \{l_i\}) = \sum_{j=2}^{J} \|\hat{Y}_j - \Psi_\Lambda^j {\Psi_\Lambda^j}^\dagger \hat{Y}_j\|_2^2, \tag{5.17}$$

where $\Psi_\Lambda^j = \Phi[\mathcal{P}_j(g_{\gamma_1}, l_1),\ \mathcal{P}_j(g_{\gamma_2}, l_2), \ldots, \mathcal{P}_j(g_{\gamma_i}, l_i), \ldots, \mathcal{P}_j(g_{\gamma_K}, l_K)]$. The operator $\mathcal{P}_j(g_{\gamma_i}, l)$ represents the projection of the atom g_{γ_i} to the j^{th} view when the local transformation is given by the depth label l (see Fig. 5.2). The data term in Eq. (5.17) is similar to the data term described earlier for image pairs (see Eq. (5.6)) except that the sum is computed over all the views. Depending on the relative position of the j^{th} camera with respect to the reference camera, the projection $\mathcal{P}_j(g_{\gamma_i}, l)$ can involve changes in the translation, rotation or scaling parameter, or combinations of them. The projection $\mathcal{P}_j(g_{\gamma_i}, l)$ of the atom g_{γ_i} to the j^{th} view approximately corresponds to another atom in the dictionary $\mathcal{D}$. It is interesting to note

that the data cost is minimal if the projection of the feature g_{γ_i} onto another view corresponds to its actual position in this view[2]. This happens when the depth label l corresponds to the true distance to the visual object represented by the atom g_{γ_i}. For example, the projection of the feature g_{γ_1} in Fig. 5.2 corresponds to the actual position of the features in views I_2 and I_3. Therefore, the data cost for this feature g_{γ_1} at label $l = 2$ is minimal. On the other hand, the projection of the feature g_{γ_2} is far from the actual position of the corresponding feature in the view I_2. The corresponding data cost $\|Y_2 - \Psi^2_\Lambda {\Psi^2_\Lambda}^\dagger Y_2\|^2_2$ is high, which indicates a suboptimal estimation of the depth label l in this case.

The smoothness cost $E_{s,u}$ enforces consistency in the depth label for the adjacent pixels $\mathbf{z}$ and $\mathbf{z}'$. It is given as

$$E_{s,u} = \sum_{\mathbf{z},\mathbf{z}' \in \mathcal{N}} \min\left(|l(\mathbf{z}) - l(\mathbf{z}')|, \tau\right), \tag{5.18}$$

where τ is a constant and $\mathcal{N}$ represents the usual 4-pixel neighborhood. Finally, the consistency term $E_{t,u}$ favors depth labels that lead to image predictions that are consistent with the quantized measurements. We compute the consistency for all the views as the sum of terms E_t given in Eq. (5.15). More formally, the consistency term $E_{t,u}$ in the multi-view scenario is given as

$$E_{t,u}(\{g_{\gamma_i}\}, \{l_i\}) = \sum_{j=2}^{J} \|\hat{Y}_j - \mathcal{Q}[\Phi \tilde{I}_j]\|^2_2 = \sum_{j=2}^{J} \|\hat{Y}_j - \mathcal{Q}[\Phi \mathcal{W}^j(\hat{I}_1, \{l_i\})]\|^2_2, \tag{5.19}$$

where $\mathcal{W}^j(\hat{I}_1, \{l_i\})$ predicts the j^{th} view with the label set $\{l_i\}$ corresponding to the set of K atoms $\{g_{\gamma_i}\}$. Finally, it is interesting to note that the optimization problem OPT-3 is similar to the OPT-2 problem except that the sum is carried out for all the J views in the data and consistency terms. Therefore, the optimization problem OPT-3 can be solved in polynomial time using the optimization methodologies described in Section 5.4.

5.7 Experimental results

5.7.1 Setup

We analyze the performance of the correlation estimation algorithms in multi-view imaging and distributed video coding applications. In order to compute a sparse approximation of the reference image at the decoder, we use a dictionary that is constructed using two generating functions [38]. The first one consists of 2D Gaussian functions to capture the low frequency components. The second function represents Gaussian in one direction and the second derivative of 2D Gaussian in the orthogonal direction to capture the edges. The discrete parameters of the functions in the dictionary are chosen as follows. The translation parameters t_x and t_y take any positive value and cover the full height N_1 and width N_2 of the image. Ten rotation values are used between 0 and π, with increments of $\pi/18$. Five scaling parameters are equi-distributed in the logarithmic scale from 1 to $N_1/8$ vertically and 1 to $N_2/9.77$ horizontally, where $N = N_1 \times N_2$ is the resolution of the image. Unless stated differently, we use global optimization methods based on Graph Cuts to solve OPT-1 and OPT-2. The regularization parameters α_1 and α_2 can be selected in two ways: (1) the motion field minimizes the error with respect to groundtruth transformation field, i.e., flow field should be as close as possible to the groundtruth (2) the flow field maximizes the quality of the predicted image $\tilde{I}_2$. While this could be done in both ways we generally select the parameter set α_1 and α_2 (based on trial and error experiments) such that the estimated flow field maximizes the quality of the predicted image $\tilde{I}_2$. Note that in practice, the groundtruth correlation model and the original images are not available a priori to estimate an optimal regularization parameters α_1 and α_2. In such cases, the parameters α_1 and α_2 can

[2] we assume here that we have no occlusions.

be estimated based on learning from a set of training images or using the automated method proposed in [143].

The image I_2 is captured by random linear projections using a scrambled block Hadamard transform with a block size of 8 [30]. The measurements Y_2 are quantized using a uniform quantizer, and the bit rate is computed by encoding the quantized measurements using an arithmetic coder. We report in this section the performance of our correlation estimation algorithm applied to the multiple images captured in multi-view and video imaging scenarios. We then analyze the influence of the measurement consistency term in the energy minimization constraint and analyze the influence of the quantization of measurements for the second image. We then compare the performance between the local and global optimization schemes. Then, we study the image prediction results as a function of the measurement rate or the coding rate of the second image. Finally, we compare the rate-distortion performance of the second image to state-of-the-art solutions for independent and distributed image coding.

5.7.2 Generic transformation

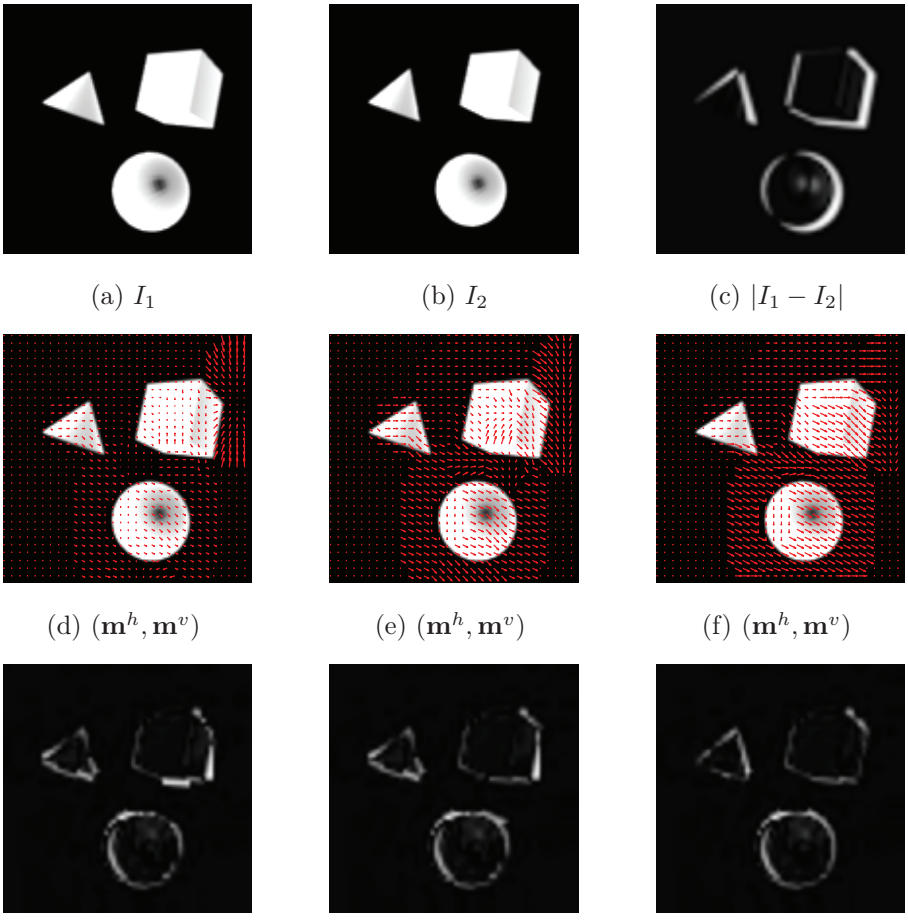
We first study the performance of our scheme with a pair of synthetic images that contains three objects. The original images I_1 and I_2 are given in Fig. 5.3(a) and Fig. 5.3(b) respectively. It is clear that the common objects in the images have different positions and scales. The absolute error between the original images is given in Fig. 5.3(c), where the PSNR between I_1 and I_2 is 15.6 dB.

We encode the reference image I_1 to a quality of 35 dB and the number of features used for the approximation of $\hat{I}_1$ is set to $N = 15$. The transformation field is estimated with $\delta t_x = \delta t_y = 3$ pixels, $\delta s_x = \delta s_y = 2$ samples and $\delta\theta = 0$. We first estimate the transformation field with problem OPT-1, by setting $\alpha_1 = 0$, i.e., the smoothness term E_s is not activated. The resulting motion field is shown in Fig. 5.3(d). From Fig. 5.3(d) we observe that the proposed scheme gives a good estimation of the transformation field even with a 5% measurement rate that are quantized with 2 bits. We further see that the image $\tilde{I}_2$ predicted with help of the estimated correlation is closer to the original image I_2 than to I_1 (see Fig. 5.3(g)). We then include the consistency term in addition to the data cost and we solve the problem OPT-2 without activating the smoothness term, i.e., $\alpha_1 = 0$. The estimated transformation field and the prediction error are shown in Fig. 5.3(e) and Fig. 5.3(h), respectively. We observe that the consistency term improves the quality of the motion field and the prediction quality. Finally, we highlight the benefit of enforcing smoothness constraint in our OPT-2 problem. The estimated transformation field with the OPT-2 problem including the smoothness term is shown in Fig. 5.3(f). By comparing the correlation estimation in Fig. 5.3(e) and Fig. 5.3(f) we see that the dense transformation map is smoother and more coherent when the smoothness term is activated. Quantitatively, the smoothness energy E_s of the motion field shown in Fig. 5.3(f) is 1479, which is clearly smaller comparing to the solutions given Fig. 5.3(d) and Fig. 5.3(e) (resp. 4309 and 4851). Also, the smoothness term effectively improves the quality of the predicted image and the predicted image $\tilde{I}_2$ gets closer to the original image I_2, as shown in Fig. 5.3(i).

5.7.3 Stereo image coding

We then study the performance of our distributed image representation algorithms in stereo imaging frameworks. We use two image datasets, namely *Plastic* and *Sawtooth*[3] [7]. These datasets have been captured by a camera array where the different viewpoints are uniformly arranged on a line. As this corresponds to translating the camera along one of the image coordinate axes, the disparity estimation problem becomes a one-dimensional search problem, and the smoothness term in Eq. (5.10) is simplified accordingly. The images are downsampled to a resolution $N = 144 \times 176$. We carry out experiments using the views 1 and 5 of the Sawtooth image set, and views 1 and 2 of the Plastic image set. The viewpoint 1 is selected as the reference image I_1 and it is encoded such that the quality of $\hat{I}_1$ is approximately 33 dB with respect to

[3]These image sets are available at http://vision.middlebury.edu/stereo/data/



(a) I_1 (b) I_2 (c) $|I_1 - I_2|$

(d) $(\mathbf{m}^h, \mathbf{m}^v)$ (e) $(\mathbf{m}^h, \mathbf{m}^v)$ (f) $(\mathbf{m}^h, \mathbf{m}^v)$

(g) $|I_2 - \tilde{I}_2|$ (h) $|I_2 - \tilde{I}_2|$ (i) $|I_2 - \tilde{I}_2|$

Figure 5.3: Comparison of the estimated motion fields and the predicted images with the OPT-1 and OPT-2 problems in the synthetic scene: (a) original image I_1; (b) original image I_2; (c) absolute error between I_1 and I_2; (d) motion field $(\mathbf{m}^h, \mathbf{m}^v)$ estimated with OPT-1 without activating E_s, i.e., $\alpha_1 = 0$; (e) motion field $(\mathbf{m}^h, \mathbf{m}^v)$ estimated with OPT-2 without activating E_s, i.e., $\alpha_1 = 0$; (f) motion field $(\mathbf{m}^h, \mathbf{m}^v)$ with OPT-2. The smoothness energy E_s of the motion fields are (d) 4309 (e) 4851 and (f) 1479. (g), (h) and (i) represent the quality of the predicted image using the motion field in (d), (e) and (f) respectively. The PSNR of the predicted image $\tilde{I}_2$ with respect to I_2 is found out to be 20 dB, 20.4 dB and 21.53 dB respectively. Quality of the reference image used in this experiment is 35 dB. The motion field is estimated using a measurement rate of 5% with a 2-bit quantization.

I_1. Matching pursuit is then carried out on $\hat{I}_1$, and we select $K = 30$ and $K = 60$ features for the Plastic and Sawtooth datasets respectively. The measurements are generally quantized using a 2-bit uniform quantizer. At the decoder, the search for the geometric transformations F^i is carried out along the translational component t_x with window size $\delta t_x = 4$ pixels, and no search is considered along the vertical direction, i.e., $\delta t_y = 0$. Unless stated explicitly, we solve the OPT-1 and OPT-2 optimization problems using the data cost E_d given in Eq. (5.6).

We first study the performance of the estimated disparity information. We show in Fig. 5.4 and Fig. 5.5 the estimated disparity field $\mathbf{m}^h$ from 8870 quantized measurements (i.e., a measurement rate of 35%) for both image sets respectively. The groundtruth $\mathbf{M}^h$ is given in Fig. 5.4(a) and Fig. 5.5(a) respectively. The

(a) $\mathbf{M}^h$ (b) $\mathbf{m}^h$ (c) $|\mathbf{M}^h - \mathbf{m}^h| > 1$ (d) $\mathbf{m}^h$ (e) $|\mathbf{M}^h - \mathbf{m}^h| > 1$

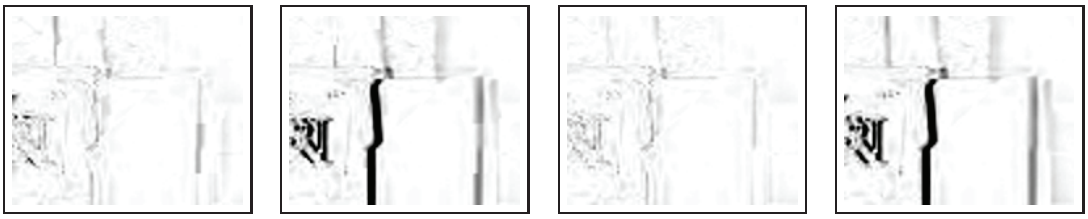
Figure 5.4: Comparison of the estimated disparity fields with OPT-1 and OPT-2 for the Plastic dataset: (a) groundtruth disparity field $\mathbf{M}^h$ between views 1 and 2; (b) estimated disparity field with OPT-1; (c) error in the disparity map with OPT-1 (DE = 10.8%); (d) estimated disparity field with OPT-2; (d) error in the disparity map with OPT-2 (DE = 4.1%). The disparity field is estimated using a measurement rate of 35% with a 2-bit quantization.

(a) $\mathbf{M}^h$ (b) $\mathbf{m}^h$ (c) $|\mathbf{M}^h - \mathbf{m}^h| > 1$ (d) $\mathbf{m}^h$ (e) $|\mathbf{M}^h - \mathbf{m}^h| > 1$

Figure 5.5: Comparison of the estimated disparity fields with OPT-1 and OPT-2 for the Sawtooth dataset: (a) groundtruth disparity field $\mathbf{M}^h$ between views 1 and 5; (b) estimated disparity field with OPT-1; (c) error in the disparity map with OPT-1 (DE = 8.1%); (d) estimated disparity field with OPT-2; (d) error in the disparity map with OPT-2 (DE = 4.1%). The disparity field is estimated using a measurement rate of 35% with a 2-bit quantization.

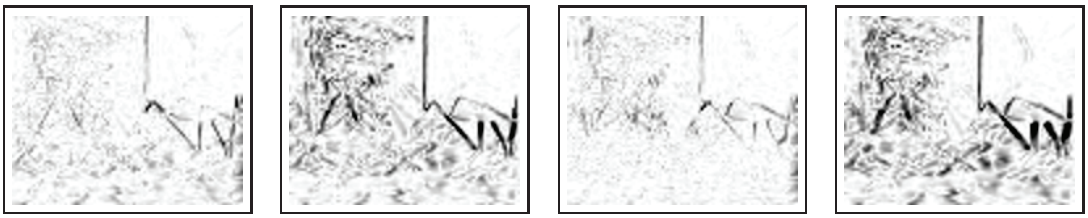
transformation F^i is estimated by solving OPT-1 and the resulting dense disparity fields are illustrated in Fig. 5.4(b) and Fig. 5.5(b) respectively for the datasets. In this particular experiment, the parameters α_1 and α_2 are selected such that the error in the disparity map is minimized. The disparity error DE is computed between the estimated disparity field $\mathbf{m}^h$ and groundtruth $\mathbf{M}^h$ as $DE = \frac{1}{N}\sum_{\mathbf{z}=(m,n)} \left\{|\mathbf{M}^h(\mathbf{z}) - \mathbf{m}^h(\mathbf{z})| > 1\right\}$ where N represents the pixel resolution of the image [7]. From Fig. 5.4(b) and Fig. 5.5(b) we observe that OPT-1 gives a good estimation of the disparity map; in particular the disparity value is correctly estimated in the regions with texture or depth discontinuity. We can also observe that the estimation of the disparity field is however less precise in smooth regions as expected from feature-based methods. Fortunately, the wrong estimation of the disparity value in the smooth regions does not significantly affect the warped or predicted image quality. Fig. 5.4(c) and Fig. 5.5(c) confirm such a distribution of the disparity estimation error, where the white pixels denote an estimation error larger than one. We can see that the error in the disparity field is highly concentrated along the edges, since crisp discontinuities cannot be accurately captured due to the scale and smoothness of the atoms in the chosen dictionary. The disparity information estimated by OPT-2 is presented in Fig. 5.4(d) and Fig. 5.5(d) and the corresponding errors in Fig. 5.4(e) and Fig. 5.5(e). We clearly see that the addition of the consistency term E_t in the correlation estimation algorithm improves the performance.

We propose a different illustration of the disparity estimation performance in Fig. 5.6 and Fig. 5.7. The dense disparity field $\mathbf{m}^h$ is used to warp the reference image $\hat{I}_1$ and the predicted image is represented by $\tilde{I}_2$. We estimate the correlation between images in both datasets with OPT-1 using 2534 quantized measurements (i.e., a measurement rate of 10%). We then warp the reference image to reconstruct an approximation of the second image, denoted as $\tilde{I}_2$. We show in Figs. 5.6(a) and (b) (resp. Figs. 5.7(a) and (b)) the comparison between $\tilde{I}_2$ and respectively I_2 and I_1, where white pixels represent a correct



(a) MSE: 59.39 (b) MSE: 631.5 (c) MSE: 38.38 (d) MSE: 652.11

Figure 5.6: Comparison of the warped image $\tilde{I}_2$ with respect to I_2 and I_1 in the Plastic dataset: (a) $1-|\tilde{I}_2-I_2|$ with OPT-1; (b) $1-|\tilde{I}_2-I_1|$ with OPT-1; (c) $1-|\tilde{I}_2-I_2|$ with OPT-2; (d) $1-|\tilde{I}_2-I_1|$ with OPT-2. The image $\tilde{I}_2$ is predicted using a measurement rate of 10% with a 2-bit quantization.

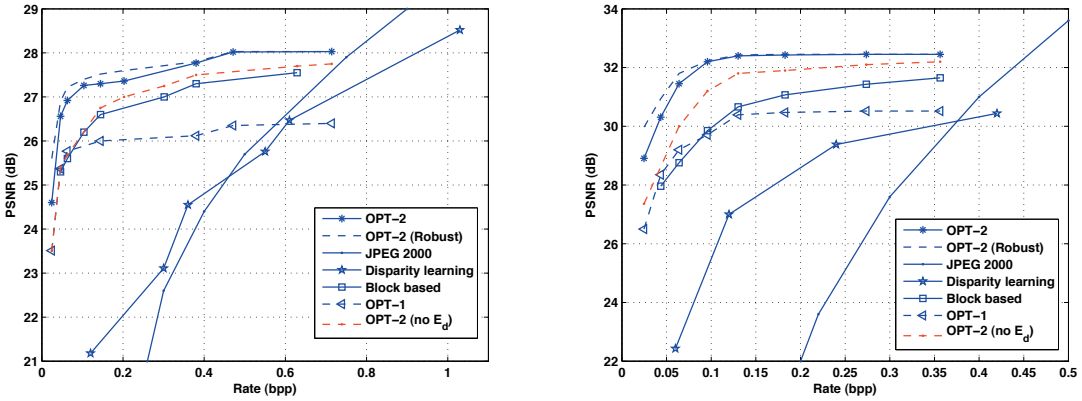


(a) MSE: 161.74 (b) MSE: 583.30 (c) MSE: 125.41 (d) MSE: 747.52

Figure 5.7: Comparison of the warped image $\tilde{I}_2$ with respect to I_2 and I_1 in the Sawtooth dataset: (a) $1-|\tilde{I}_2-I_2|$ with OPT-1; (b) $1-|\tilde{I}_2-I_1|$ with OPT-1; (c) $1-|\tilde{I}_2-I_2|$ with OPT-2; (d) $1-|\tilde{I}_2-I_1|$ with OPT-2. The image $\tilde{I}_2$ is predicted using a measurement rate of 10% with a 2-bit quantization.

reconstruction. It is clear that $\tilde{I}_2$ is closer to I_2 than I_1, which confirms that the proposed scheme efficiently captures the correlation between images. The same comparisons are given in Figs. 5.6(c) and (d) (resp. Figs. 5.7(c) and (d)) when OPT-2 is used for correlation estimation. The results confirm that the addition of the consistency term again provides a more accurate disparity field since the warped image $\tilde{I}_2$ gets quite closer to the target image I_2.

We study now the rate-distortion performance of the proposed algorithms for the prediction of the image $\tilde{I}_2$ in Fig. 5.8 for both datasets. We show the performance of the reconstruction by warping the reference image according to the correlation computed by OPT-1 and OPT-2. We compare the rate-distortion performance to a distributed coding solution (DSC) based on the LDPC encoding of DCT coefficients, where the disparity field is estimated at the decoder using Expected Maximization (EM) principles [86]. The scheme is denoted as *Disparity learning* in the figures. Then, in order to demonstrate the benefit of geometric dictionaries, we also propose a scheme denoted as *block-based* that adaptively constructs the dictionary using blocks or patches in the reference image [103]. As described in [103], we construct a dictionary in the joint decoder from the reference image $\hat{I}_1$ segmented into 8×8 blocks. The search window size is $\delta t_x = 4$ pixels along the horizontal direction. We then use the optimization scheme described in OPT-2 to select the best block from the adaptive dictionary. In order to have a fair comparison, we encode the reference image I_1 similarly for both schemes (*Disparity learning* and *block-based*) with a quality of 33 dB (see Section 5.2). Finally, we also provide the performance of a standard JPEG 2000 independent encoding of the image I_2. From Fig. 5.8, we first see that the measurement consistency term E_t greatly improves the prediction quality, as OPT-2 gives better performance than OPT-1. We then solve the OPT-2 without the data cost function E_d, i.e., using only the smoothness E_s and consistency E_t terms (marked as *OPT-2 (no E_d)*). From Fig. 5.8, as expected, we see that the RD performance is maximized only when all the three terms in OPT-2



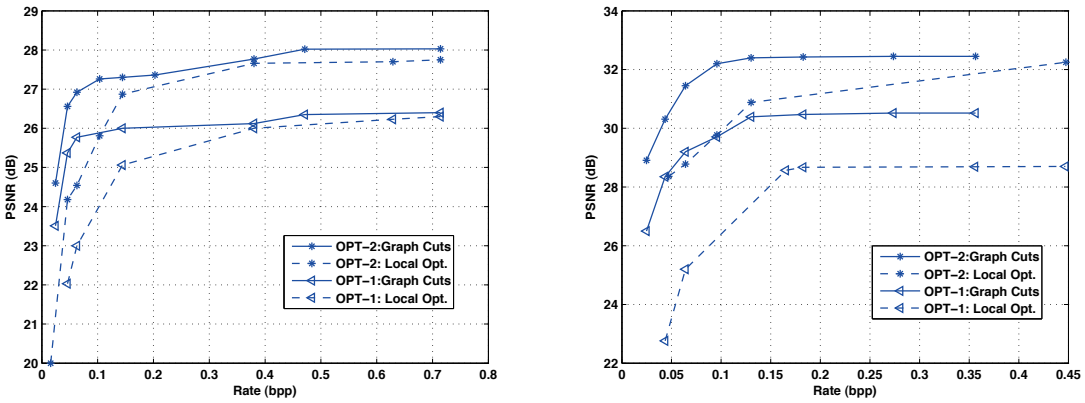
(a) (b)

Figure 5.8: Comparison of the RD performances between the proposed scheme, DSC scheme [86], block-based scheme [103] and independent coding solutions based on JPEG 2000: (a) Sawtooth image set; (b) Plastic image set.

are activated. The results confirm that the proposed correlation estimation algorithms unsurprisingly outperform independent coding based on JPEG 2000, which outlines the benefits of the correlation information in the decoding of compressed correlated images. It is worth mentioning that the JPEG 2000 compression standard fails to provide a good RD performance for low resolution images, typically of image resolution $N = 144 \times 176$ used in our experiments. Though this is the case, JPEG 2000 is still state-of-the-art coding solutions for independent compression. From Fig. 5.8 we see that at high bit rate the performance of the proposed algorithms however tend to saturate as our model mostly handles the geometry and the correlation between images. But our scheme is not able to efficiently handle the fine details or texture in the scene, due to the image reconstruction $\hat{I}_2$ based on warping. In Chapter 7, we propose a joint reconstruction methodology to correct this saturation behavior by estimating the details and texture information from quantized linear measurements. From Fig. 5.8, it is then clear that the reconstruction based on OPT-1 and OPT-2 outperforms the DSC coding scheme based on EM principles for both datasets. Finally, the experimental results also show that our schemes outperform the scheme based on block-based dictionary mainly because of the richer representation of the geometry and local transformations with the structured dictionary.

We now compare the RD performances between the local and global optimization methodologies with the OPT-1 and OPT-2 schemes. It should be noted that our local optimization scheme implements constructive parameter search strategy summarized in the Algo. 4 and Algo. 5 respectively. The comparison is available in Fig. 5.9 where the RD performances are represented in the *dashed* and *dotted lines* respectively. From Fig. 5.9 we observe that the local optimization scheme estimates a suboptimal solution, where it has a lower RD performance comparing to the one estimated based on Graph Cuts. However, comparing Fig. 5.8 and Fig. 5.9 it is interesting to note that the suboptimal solution estimated based on the constructive search space still outperforms state-of-the-art independent and distributed coding solutions.

We now study the performance of OPT-2 in different settings in terms of camera distances, quality of the reference image and quantization of the measurements. We first illustrate the rate-distortion performance of the proposed scheme for images captured at various distances from the reference camera. In particular, we study the quality of the images at viewpoints 3 and 5 in the Sawtooth dataset that are decoded by warping the reference image at viewpoint 1. Fig. 5.10 confirms that the performance is better when the correlation between images is stronger (the reference image is closer to viewpoint 3 than viewpoint 5). We further see that the coding performance is better than state-of-the-art independent coding with JPEG 2000 and DSC based on disparity learning [86], when the correlation between images is high. We further study the influence



(a) (b)

Figure 5.9: Comparison of the RD performances between the global and local optimization methodologies: (a) Sawtooth image set; (b) Plastic image set.

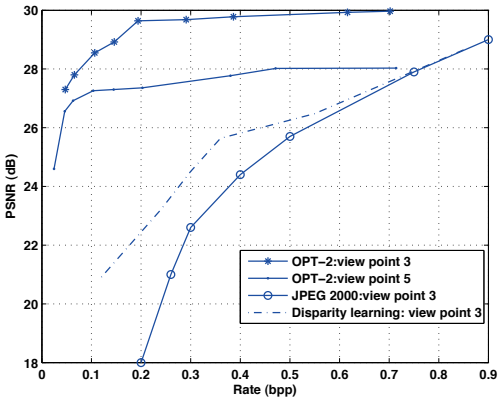


Figure 5.10: Rate-distortion performance as a function of the correlation between images. OPT-2 is used to predict images at different viewpoints in the Sawtooth dataset, while the image at viewpoint 1 is selected as a reference image.

of the quality of the reference image $\hat{I}_1$ on the decoding performance. We use OPT-2 to decode $\tilde{I}_2$ (viewpoint 5) by warping, when the reference image is encoded at different qualities (i.e., different bit rates). Fig. 5.11 shows that the prediction image quality $\tilde{I}_2$ improves with the quality of the reference image $\hat{I}_1$, as expected. While the error in the disparity estimation is not drastically reduced by improved reference image $\hat{I}_1$ quality, the warping stage permits to provide more details in the representation of $\tilde{I}_2$, when the reference is of better quality. Now, we study the cumulative RD performance of views 1 and 5 for the Sawtooth dataset, i.e., we include the bit rate and quality of the reference image I_1 (viewpoint 1), in addition to the rate and quality of image I_2 (viewpoint 5). Fig. 5.12 shows the cumulative RD performance for various reference image bit rates 0.2, 0.3, 0.4, 0.5, 0.75 and 1.5 bpp. In our experiments, for a given reference image quality we estimate the correlation model using OPT-2 (with 2-bit quantized measurements), and we compute the joint RD performance at that specific reference image bit rate. As shown before, the RD performance improves with increasing reference image quality. When we take the convex hull of the RD performances

(which corresponds to implementing a proper rate allocation strategy), we outperform independent coding solutions based on JPEG 2000.

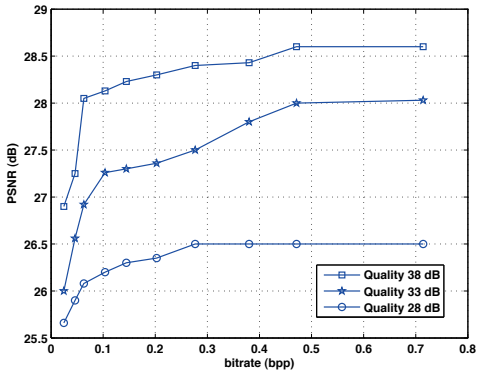


Figure 5.11: Rate-distortion performance with OPT-2 for decoding $\tilde{I}_2$ (viewpoint 5) as a function of the quality of the reference image $\hat{I}_1$ (resp. 28 dB, 33 dB and 38 dB) in the Sawtooth dataset.

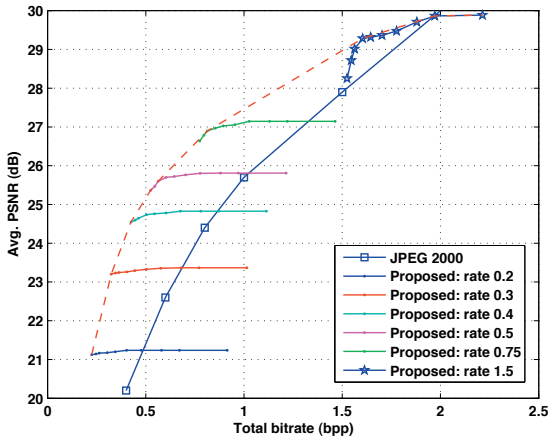
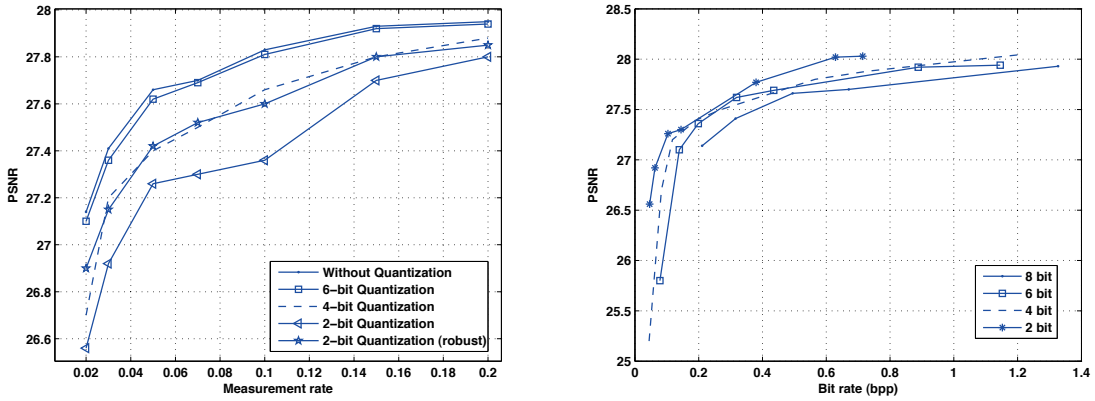


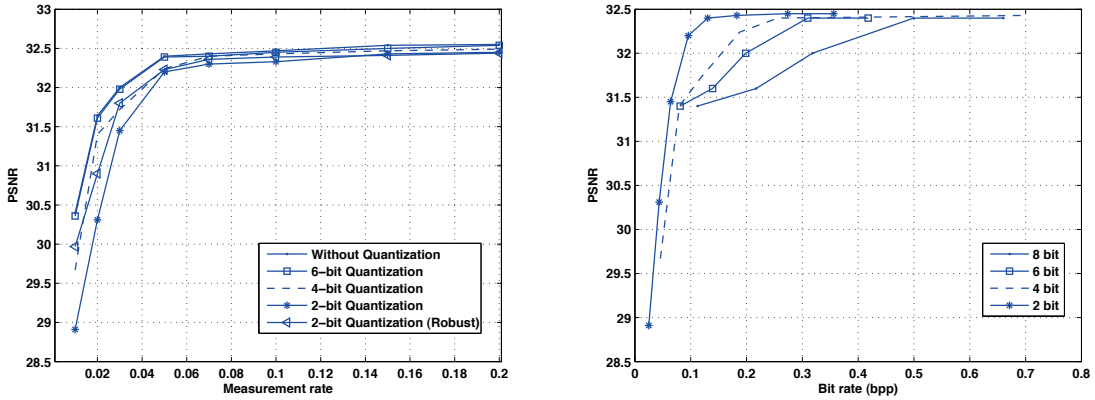
Figure 5.12: Cumulative RD performance of the views 1 and 5 for the Sawtooth dataset. OPT-2 is used to predict the image $\tilde{I}_2$ (view 5) using the image $\hat{I}_1$ (view 1) as the reference image. The image at view 5 is predicted with varying reference image bit rates 0.1, 0.2, 0.3, 0.4, 0.5, 0.75 and 1.5.

We finally study the influence of the quantization bit rate on the RD performance of $\tilde{I}_2$ with the OPT-2. We quantize the measurements $\hat{Y}_2$ with a number of bits between 2 to 6. While the quality of the correlation estimation degrades when the number of bits reduces (see Fig. 5.13(a) and Fig. 5.14(a)), it is largely compensated by the reduction in bit rate in the RD performance, as confirmed by Fig. 5.13(b) and Fig. 5.14(b). This means that the proposed correlation estimation is relatively robust to quantization so that it is possible to attain good rate-distortion performance by drastic quantization of the measurements. Then, we study the improvement offered by the robust data cost $\tilde{E}_d$ (see Eq. (5.8)) in OPT-2 (i.e., OPT-2 (Robust) problem), when the measurements have been compressed with a 2-bit uniform quantizer and an arithmetic coder. We use the optimization toolbox based on CVX [151] in order to solve the optimization problem described in Eq. (5.7). From Fig. 5.13(a) and Fig. 5.14(a) it is clear that the performance of the scheme can be improved by activating the robust data term. Also, Fig. 5.8 compares the modified OPT-2



(a) (b)

Figure 5.13: Effect of measurement quantization on the quality of the image $\tilde{I}_2$ decoded with OPT-2 scheme in the Sawtooth dataset. The quality of the predicted image $\tilde{I}_2$ is given in terms of (a) measurement rate and (b) bit rate. The benefit of using robust data cost is illustrated using a 2-bit uniform quantizer.



(a) (b)

Figure 5.14: Effect of measurement quantization on the quality of the image $\tilde{I}_2$ decoded with OPT-2 scheme in the Plastic dataset. The quality of the predicted image $\tilde{I}_2$ is given in terms of (a) measurement rate and (b) bit rate. The benefit of using robust data cost is illustrated using a 2-bit uniform quantizer.

with DSC based on disparity learning [86], a joint reconstruction with a block-based dictionary [103] and independent coding with JPEG 2000. The relative performance of the different schemes are maintained with the robust data cost $\tilde{E}_d$. However, the robust data cost permits to improve the quality of the predicted image $\tilde{I}_2$ by 0.5-1 dB, especially at low measurement rate.

5.7.4 Distributed video coding

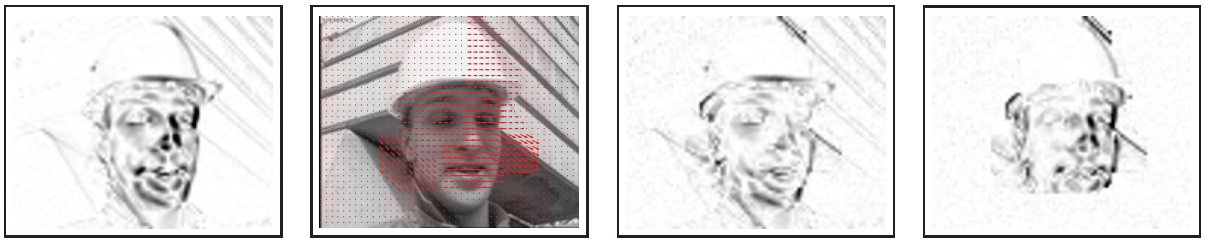
We study now the performance of the proposed algorithms in distributed video coding applications. The experimental setup is similar to the stereo imaging framework described in the previous section, except

that the correlation estimation relates to the motion estimation instead of disparity computation. We have tested our algorithm on three datasets. The first dataset is built using the frames 2 and 3 of the *Foreman* sequence, and the second dataset is built using the frames 65 and 66 of the *Tennis* sequence. The third dataset *Mequon* is selected from the Middlebury optical flow database[4]. Frames 2 and 66 are selected as the reference image I_1 in the first and second datasets respectively. The quality of the reference image is approximately 45 dB for the first dataset, and 33 dB for the second and third datasets. We use the same dictionary described in the previous section for approximating the image $\hat{I}_1$. For the first dataset we approximate $\hat{I}_1$ using $K = 60$ atoms, and for the second and third datasets we approximate $\hat{I}_1$ using $K = 90$ atoms. The measurements Y_2 are compressed using a 2-bit uniform quantizer and an arithmetic coder. The search window size is $\delta t_x = \delta t_y = 4$ pixels for the translational components t_x and t_y for the first dataset, and $\delta t_x = \delta t_y = 6$ for the second and third datasets.

Fig. 5.15 illustrates the accuracy of the motion information computed in the OPT-2 scheme with 1267 and 3801 quantized measurements (i.e., 5% and 15% measurement rate respectively). It compares the image $\tilde{I}_2$ reconstructed by warping the reference image, to respectively the original images I_2 and I_1. We see that the warped image is closer to I_2 than I_1, which confirms the benefit of the motion estimation in the joint decoder. We further observe that the error denoted by black pixels is reduced significantly in the face region due to a good estimation of the motion field in smooth areas. Similarly to the stereo experiments the motion around sharp edges is however not perfectly captured due to the choice of a dictionary that does not include very thin geometrical patterns. Similar experimental findings are observed in the Tennis and Mequon datasets shown in Fig. 5.16 and Fig. 5.17 respectively, where the proposed algorithm can capture efficiently the complex and large motion fields in the Tennis and Mequon datasets from 1267 (5%) quantized measurements.

We further study the RD performance of the proposed algorithms in the decoding of the image $\tilde{I}_2$. From Fig. 5.18(a) and Fig. 5.18(b) it is clear that our proposed solutions outperform independent coding since it exploits the correlation between images. We then compare the performance to state-of-the-art solutions in joint and distributed video coding. First, we provide the performance of a DSC scheme based on motion learning [147], using the experimental setup similar to the one demonstrated in the previous section and a reference image $\hat{I}_1$ of 45 dB for a fair comparison (denoted as *Motion learning* in the figures). In addition, we implement OPT-2 with a different dictionary that is built on blocks of the reference image, similarly to [103] (denoted as *Block scheme* in the figures). For the sake of completeness, we further provide results of a joint video encoding solution based on H.264 with an IP encoding structure (i.e., a GOP size of 2). We again encode the reference I-frame (I_1) at a quality of 45 dB for the Foreman dataset (33 dB for the Tennis dataset), and we vary the quantization parameter for the P-frame (I_2) to build the rate-distortion characteristics. We consider two different settings in the H.264 motion estimation, which is performed with variable and fixed macro block size. From Fig. 5.18(a) we first observe that the measurement consistency term E_t in OPT-2 greatly improves the performance of our motion estimation algorithm. It also outperforms the DSC solution based on motion learning due to better model of the geometric correlation between images. The correlation estimation with block-based dictionary is less efficient than the estimation with a dictionary of geometric atoms. Finally, from Fig. 5.18(a) and Fig. 5.18(b) we see that the joint encoding based on H.264 is better than the distributed coding solutions for both Foreman and Tennis datasets. However, our algorithm is able to compete at low bit rates with H.264 based on a fixed block-size motion estimation, which is certainly an interesting and promising result. It should be noted that in our scheme we predict the second image based on motion compensation (i.e., warping); this certainly fails to estimate accurately the visual information along the edges and in texture regions. On the other hand, state-of-the-art schemes such as H.264 and DSC-based on motion learning compensate also for the prediction error in addition to correlation estimation. Though this is the case, we show by experiments that the proposed scheme outperforms H.264 (at low rate) and DSC-based on motion learning due to an accurate motion field estimation. In Chapter 7, we propose a joint reconstruction algorithm that improves the quality of the image $\tilde{I}_2$ by estimating the

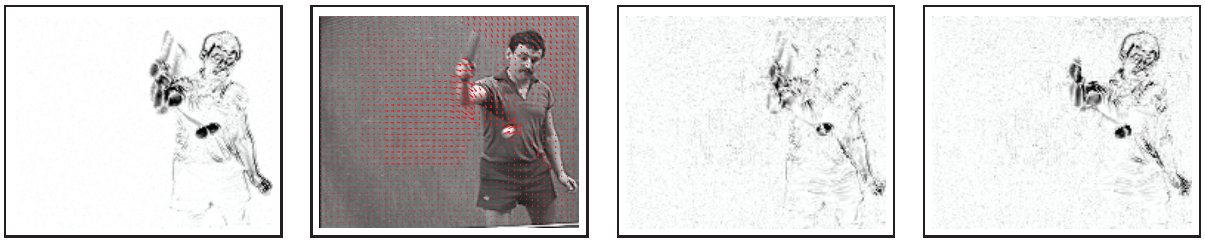
[4]available at http://vision.middlebury.edu/flow



(a) MSE: 78.49 (b) $(\mathbf{m}^h, \mathbf{m}^v)$ (c) MSE: 47.22 (d) MSE: 60.1

(e) $(\mathbf{m}^h, \mathbf{m}^v)$ (f) MSE: 38.1 (g) MSE: 62.11

Figure 5.15: Comparison of the warped image $\tilde{I}_2$ with respect to I_2 and I_1 with the OPT-2 scheme in the Foreman dataset. (a) Inverted absolute error $1 - |I_1 - I_2|$ between original images. Top row: Results estimated from a measurement rate of 5% with a 2-bit quantized measurements: (b) motion map $(\mathbf{m}^h, \mathbf{m}^v)$; (c) inverted prediction error $1 - |\tilde{I}_2 - I_2|$ with respect to I_2; (d) inverted prediction error $1 - |\tilde{I}_2 - I_1|$ with respect to I_1. Bottom row: Results estimated from a measurement rate of 15% with a 2-bit quantized linear measurements: (e) motion map $(\mathbf{m}^h, \mathbf{m}^v)$; (f) inverted prediction error $1 - |\tilde{I}_2 - I_2|$ with respect to I_2; (g) inverted prediction error $1 - |\tilde{I}_2 - I_1|$ with respect to I_1.



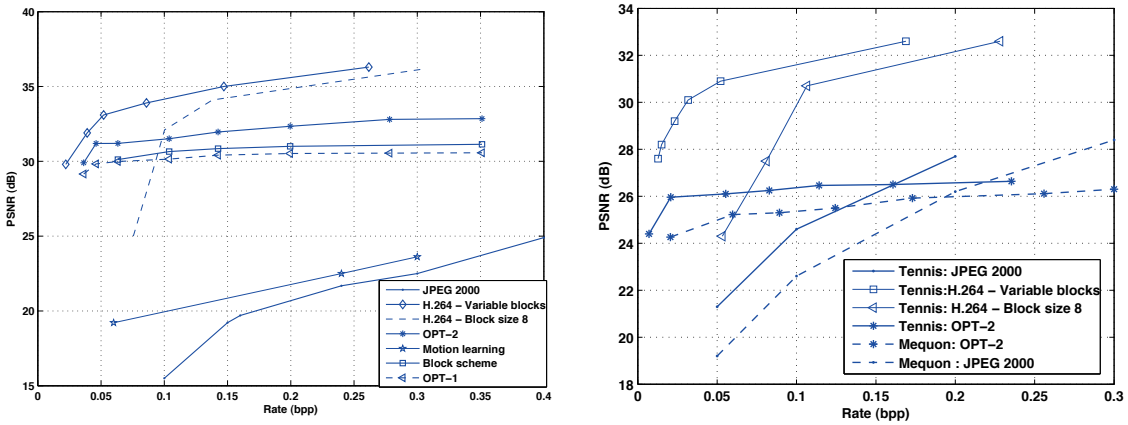
(a) MSE: 282.77 (b) $(\mathbf{m}^h, \mathbf{m}^v)$ (c) MSE: 154.25 (d) MSE: 271.12

Figure 5.16: Comparison of the warped image $\tilde{I}_2$ with respect to I_2 and I_1 with the OPT-2 scheme in the Tennis dataset: (a) inverted absolute error $1 - |I_1 - I_2|$ between original images; (b) motion map $(\mathbf{m}^h, \mathbf{m}^v)$ estimated with OPT-2; (c) inverted prediction error $1 - |\tilde{I}_2 - I_2|$ with respect to I_2; (d) inverted prediction error $1 - |\tilde{I}_2 - I_1|$ with respect to I_1. The motion field is estimated using a measurement rate of 5% with a 2-bit quantized linear measurements.

missing visual information from quantized linear measurements. We then compare the RD performances for the predicted image $\tilde{I}_2$ between the graph-based (global) and the constructive search parameter (local) optimization methodologies. The comparison is available in Fig. 5.19 for the Foreman dataset, where the RD performances for global and local optimization schemes are represented in the *dashed* and *dotted lines* respectively. From Fig. 5.9, we see that the RD performance is significantly improved when the OPT-1 and OPT-2 optimizations are solved using strong optimization techniques based on Graph Cuts, which is

(a) MSE: 481.01 (b) $(\mathbf{m}^h, \mathbf{m}^v)$ (c) MSE: 195.77 (d) MSE: 414.17

Figure 5.17: Comparison of the warped image $\tilde{I}_2$ with respect to I_2 and I_1 with the OPT-2 scheme in the Mequon dataset: (a) inverted absolute error $1 - |I_1 - I_2|$ between original images; (b) motion map $(\mathbf{m}^h, \mathbf{m}^v)$ estimated with OPT-2; (c) inverted prediction error $1 - |\tilde{I}_2 - I_2|$ with respect to I_2; (d) inverted prediction error $1 - |\tilde{I}_2 - I_1|$ with respect to I_1. The motion field is estimated using a measurement rate of 5% with a 2-bit quantized linear measurements.



(a) **(b)**

Figure 5.18: (a) Rate-distortion performance with OPT-1 and OPT-2 schemes for decoding $\tilde{I}_2$ in the Foreman dataset. Comparisons with state-of-the-art coding solutions in independent, joint and distributed video coding schemes. (b) Rate-distortion performance of the proposed OPT-2 scheme with state-of-the-art independent coding solutions based on JPEG 2000 for Tennis and Mequon datasets.

consistent with our earlier observations in the disparity estimation experiments.

5.7.5 Multi-view image coding

We finally evaluate the performance of our correlation estimation algorithms in scenarios with more than 2 correlated images. We use five images from the Tsukuba dataset (center, left, right, bottom and top views), and five frames (frames 3-7) from the Flower Garden sequence [150]. These datasets are down-sampled by a factor 2 and the resolution of these datasets used in our experiments is of 144×192 and 120×180 respectively. In both datasets, the reference image I_1 (center view and frame 5 resp.) is encoded with a quality of approx. 33 dB. The measurements $Y_i, \forall i \in \{1, 2, 3, 4\}$ computed from the remaining four images are quantized using a 2-bit quantizer. We first compare our results to a stereo setup where a depth field

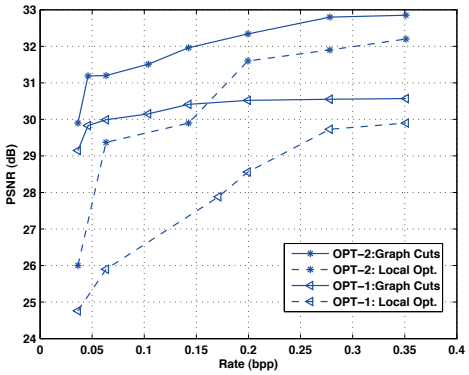


Figure 5.19: Comparison of the RD performances between the global and local optimization methodologies for the Foreman dataset.

is estimated with OPT-2 between the center and left images (in Tsukuba), and frames 5 and 6 (in Flower Garden) respectively. Fig. 5.20 compares the inverse depth error (sum of the labels with error larger than one with respect to groundtruth) between the multi-view and disparity scenarios. In this particular experiment, the parameters α_1 and α_2 are selected such that they minimize the error in the depth image with respect to the groundtruth. It is clear from the plot that the depth error is small for a given measurement rate when all the views are available. It should be noted that the x-axis in Fig. 5.20 represents the measurement rate per view, and hence the total number of measurements used in the multi-view scenario is higher than for the stereo case. However, these experiments show that the proposed multi-view scheme gives a better depth image when more images are available. Fig. 5.21(b) and Fig. 5.21(d) show the depth map estimated using the OPT-3 (multi-view) and OPT-2 (stereo) algorithms respectively from 3% quantized measurements (per view) for the Flower Garden sequence. As the groundtruth for this dataset is not available, we estimate the depth using the original images and it is available in Fig. 5.21(a) for visual comparison. From Fig. 5.21(b) and Fig. 5.21(d) we see that the proposed multi-view scheme estimates a better depth map than the stereo setup, as it is more accurate in the background and tree regions. When this coarse depth map is used for predicting frame 6, we see from Fig. 5.21(b) and Fig. 5.21(e) respectively that the prediction is better when five views are used for depth estimation. We then study the RD performance of the proposed multi-view scheme in the decoding of four images (top, left, right, bottom images in Tsukuba, and frames 3, 4, 6, 7 in Flower Garden). The images are decoded by warping the reference image $\hat{I}_1$ using the estimated depth map. Fig. 5.22 compares the overall RD performance (for 4 images) of our multi-view scheme with respect to independent coding performance based on JPEG 2000. As expected, the proposed multi-view scheme outperforms independent coding solutions based on JPEG 2000, as it benefits from the correlation between images. Note that the JPEG 2000 codec is state-of-the-art coding solutions for independent compression though it has not been optimized for low resolution images. Furthermore, as observed in distributed stereo and video coding the proposed multi-view coding scheme saturates at high rates, as the warping operator captures only the geometry and correlation between images and not the detail and texture information.

Finally, we compare our results with the joint encoding approach where the depth image is estimated from the original images and the computed depth image is transmitted to the joint decoder. At the decoder, the views are predicted from the reconstructed reference image $\hat{I}_1$ and the compressed depth image based on view prediction. The results are presented in Fig. 5.22 (marked as *Joint encoding*), where the bit rate is computed only on the depth image encoded using the JPEG 2000 coding solution. The main difference between our scheme and the joint encoding is that the quantized linear measurements are transmitted for a depth estimation in the former scheme, while the depth information is directly transmitted in the latter scheme. Therefore, by comparing these two approaches we can fairly judge the accuracy of the estimated

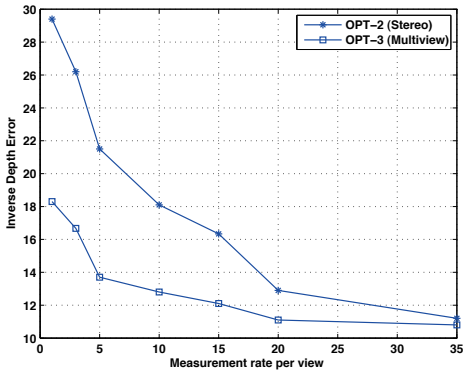


Figure 5.20: Inverse depth error at various measurement rates for the Tsukuba multi-view dataset. OPT-2 and OPT-3 are used to estimate the depth in stereo and multi-view scenarios respectively. The measurements are quantized using a 2-bit quantizer.

(a) (b) (c) MSE: 136.99 (d) (e) MSE: 167.85

Figure 5.21: Comparison of the qualities of the depth image and the warped image $\tilde{I}_2$ estimated in multi-view and stereo scenarios in the Garden dataset: (a) depth image $\mathbf{M}^h$ estimated using the original images; (b) depth image $\mathbf{m}^h$ estimated in the multi-view scenarios using OPT-3; (c) inverted prediction error $1 - |\tilde{I}_2 - I_2|$ when depth image in (b) is used for prediction; (d) depth image $\mathbf{m}^h$ estimated in the stereo scenario using OPT-2; (e) inverted prediction error $1 - |\tilde{I}_2 - I_2|$ when depth image in (c) is used for prediction. The depth image is estimated using a measurement rate of 3% (per image) with 2-bit quantized measurements.

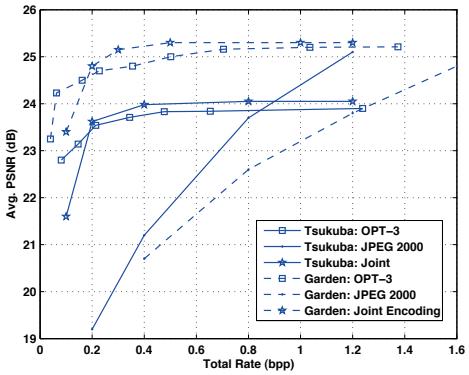


Figure 5.22: Comparison of the overall RD performances between the proposed scheme and the independent coding scheme based on JPEG 2000. Five images are used in both datasets to estimate a depth image using OPT-3. The bit rate of the reference image I_1 is not included in the total bit budget.

correlation model or equivalently the quality of the predicted view at a given bit rate. From Fig. 5.22 we see that at low rates < 0.2, the proposed scheme estimates a better structural information compared to the joint encoding scheme, thanks to the geometry-based correlation representation. However at rates above 0.2, we see that the proposed scheme becomes comparable with the joint encoding approach, leading to the conclusion that the proposed scheme effectively estimates the depth information from the highly compressed quantized measurements. It should be noted that in joint encoding approach the depth images are estimated at a central encoder. In contrary to this, we estimate the depth images at the central decoder from the independently compressed visual information and therefore our encoder is very simple.

5.8 Conclusions

In this chapter we have contributed a novel framework for the distributed representation of correlated images with quantized linear measurements along with novel algorithms for estimating the geometric correlation between images at the joint receiver. We have proposed a correlation model based on local transformations of geometric patterns that are present in the sparse representation of images. The motion or depth information in distributed video coding and respectively multi-view imaging can then be estimated by matching the corresponding geometric features in different compressed images. We have proposed a new regularized optimization problem in order to identify the geometrical transformations that result in smooth motion or depth fields between a reference image and one or more predicted image(s). Furthermore, we have proposed an improved image warping solution that leads to image prediction that is consistent with the compressed measurements.

Experimental results demonstrate that the proposed methodology provides a good estimation of dense disparity/depth or motion fields in different natural image datasets. We also show that our geometry-based correlation model is more efficient than block-based correlation models. Then, the consistency term proves to offer improved image prediction quality, such that the proposed algorithm outperforms JPEG 2000 and DSC schemes in terms of rate-distortion performance computed on the quantized linear measurements. Finally, we show experimentally that our low complex distributive correlation estimation scheme competes with the solution of the scheme that estimates a correlation model in centralized settings. Therefore, our scheme certainly provides an interesting alternative to distributed image processing applications due to a framework based on effective handling of the geometry that is one of the main component of correlated natural images. Note that the experiments are carried out using the datasets with QCIF resolution images. We left to extend our scheme to large image resolution, which is one of the interesting future perspectives

of this thesis. In Chapter 7, we propose a joint reconstruction methodology that estimates the missing high frequency texture information of the reconstructed image from the quantized linear measurements. In the next chapter, we extend the scenario to the symmetric case, where we propose to estimate the correlation model directly from the quantized linear measurements.

Chapter 6

Correlation Estimation from Compressed Linear Measurements

6.1 Introduction

In the previous chapter, we have proposed an algorithm to estimate the geometrical relationship between correlated images from a reference image and compressed linear measurements. In this chapter, we propose to estimate the correlation between images in a symmetric framework, where all the sensors transmit compressed images that have been obtained by a small number of linear projections of the original images, as illustrated in Fig. 6.1. Such linear projections typically represent simple measurements in low complexity sensing systems [15, 16].

We propose a novel solution for the correlation estimation at the joint decoder[1], where the analysis is performed directly in the compressed domain to avoid expensive image reconstruction tasks. Recall that state-of-the-art distributed schemes estimate the correlation model from the reconstructed images, which are typically obtained by solving an l_2-l_1 or l_2-TV optimization problem in the compressed sensing framework. Unfortunately, reconstructing the reference images based on solving optimization problems are highly complex. Also, in image analysis applications (e.g., object detection) it is sufficient to estimate only the correlation model and the explicit reconstruction of images is often not required. Hence, we propose to estimate the correlation model directly from the quantized linear measurements by building a pixel-wise geometrical model between images.

We assume that the correlation between images corresponds to camera or object motion, which can be efficiently represented by a dense disparity or motion field model. We show that such a correlation model can be described by a linear operator and we further analyze in detail the effect of such an operator in the compressed domain. Later, we cast the correlation estimation problem as a regularized energy minimization problem with constraints on data consistency and on consistency of the motion field. In particular, we regularize the correlation model such that the motion values in the neighboring pixels are similar except at image discontinuities. Such an optimization problem can be solved using a Graph Cut algorithm. Finally, we propose robust solutions to the correlation estimation by effectively considering the non-linearities due to quantization of measurements.

We then analyze in details the performance of our novel correlation estimation framework. In particular, we study the penalty in the correlation estimation that is due to working in the compressed domain as opposed to the original image domain as in traditional correlation estimation problems. We show that the penalty decreases when the number of measurements increases and that our algorithm tends to find the

[1] Part of this work has been accepted to: V. Thirumalai and P. Frossard, "Correlation Estimation from Compressed Images," Journal of Visual Communication and Image Representation, 2011.

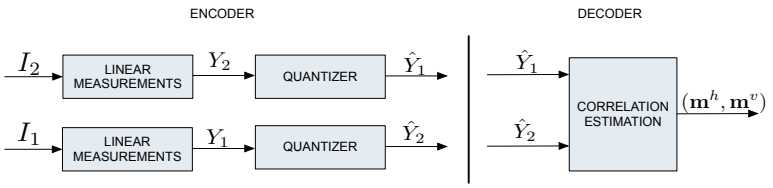


Figure 6.1: Schematic representation of the proposed scheme. The images I_1 and I_2 are correlated through displacement of the scene objects due to view point change or motion of scene objects. The correlation model is directly estimated in the compressed domain without any intermediate image reconstruction steps.

optimal correlation solution at high measurement rates. Extensive simulations in distributed multi-view and video imaging applications confirm the effective estimation performance of our algorithm. Finally, we show experimentally that our correlation estimation algorithm properly handles the quantization noise in the measurements.

The rest of this chapter is organized as follows. In Section 6.2, we describe the proposed framework and show how the correlation estimation problem carries out to the compressed domain. Section 6.3 describes the proposed correlation estimation algorithm; its performance are analyzed in details in Section 6.4. Section 6.5 extends our correlation framework to quantization scenarios. Finally, in Section 6.6 we describe extensions to multi-view imaging frameworks.

6.2 Distributed representation of correlated Images

6.2.1 Framework

We consider a framework where the images represent a scene at different time instants or from different viewpoints. For the sake of clarity, we consider a pair of images I_1 and I_2 (with resolution $N = N_1 \times N_2$), but the framework extends to larger number of images, as described in Section 6.6. These images are represented by linear measurements that correspond to the projection of the image pixel values on a set of coding vectors. Typically, the coding vectors can be constructed from Gaussian or Bernoulli distributions [15] or with a block structured matrix [30, 31] for easier handling and fast sampling of large images. The computed measurements are quantized and transmitted to a joint decoder that estimates the correlation between compressed images. The framework is illustrated in Fig. 6.1.

In more details, the sensors process images row by row. Let $I_{1,m}$ and $I_{2,m}$ represent the m^{th} row of the images I_1 and I_2 respectively, and $Y_{1,m}$ and $Y_{2,m}$ represent the linear measurements computed from $I_{1,m}$ and $I_{2,m}$ using the measurement matrices ϕ_1^m and ϕ_2^m respectively. The measurements $Y_{1,m}$ and $Y_{2,m}$ are computed as

$$\begin{aligned} Y_{1,m} &= \phi_1^m \; I_{1,m}^T, \quad \forall m = 1, 2, \ldots, N_1, \\ Y_{2,m} &= \phi_2^m \; I_{2,m}^T, \quad \forall m = 1, 2, \ldots, N_1, \end{aligned} \tag{6.1}$$

where $(.)^T$ denotes the transpose operator. It should be noted that ϕ_1^m and ϕ_2^m are of dimensions $M \times N_2$, where $M << N_2$ is the number of measurements for each row in the image. From Eq. (6.1) it is easy to check that the measurements $Y_1 = [Y_{1,1}, Y_{1,2}, \ldots, Y_{1,N_1}]^T$ and $Y_2 = [Y_{2,1}, Y_{2,2}, \ldots, Y_{2,N_1}]^T$ can be computed

as

$$\begin{bmatrix} Y_{j,1} \\ Y_{j,2} \\ \vdots \\ Y_{j,N_1} \end{bmatrix} = \Phi_j \underbrace{\begin{bmatrix} I_{j,1}^T \\ I_{j,2}^T \\ \vdots \\ I_{j,N_1}^T \end{bmatrix}}_{I_j}, \quad \forall j \in \{1,2\}, \tag{6.2}$$

where Φ_j is the measurement matrix used to sample the j^{th} image $\forall j = \{1,2\}$. It can be represented as

$$\Phi_j = \begin{bmatrix} \phi_j^1 & 0 & \dots & 0 \\ 0 & \phi_j^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \phi_j^{N_1} \end{bmatrix}_{K \times N}, \forall j \in \{1,2\}, \tag{6.3}$$

where $K = MN_1$ and $N = N_1N_2$.

6.2.2 Correlation model

In the above settings, the correlation between images is mainly explained by the relative displacement of objects in the scene. This can be modeled effectively by the optical flow that determines the amount of displacement of objects or pixels in different images. Such a correlation model can be described by a linear operator. In Chapter 4, we have described that the disparity field between two images can be represented using a linear operator. In this section, we go one step further to show that the motion compensation can also be described linearly. Let $\mathbf{m}^h$ and $\mathbf{m}^v$ represent the horizontal and vertical motion at each pixel in the image. As the visual objects in the images I_1 and I_2 are displaced, the pixel at position $\mathbf{z} = (m,n)$ in one image moves to $\mathbf{z}' = (m + \mathbf{m}^h(m,n), l + \mathbf{m}^v(m,n))$ in the second image. The images I_1 and I_2 can thus simply be related by a linear operator $\mathcal{T}$ that changes the coordinate system from (m,n) in the first image to $(m + \mathbf{m}^h(m,n), l + \mathbf{m}^v(m,n))$ in the second image, i.e.,

$$\begin{aligned} I_2 &= \mathcal{T}\{I_1\}, \\ I_{2,m}(n) &= I_{1,(m+\mathbf{m}^h(m,n)}(n + \mathbf{m}^v(m,n)). \end{aligned} \tag{6.4}$$

For mathematical convenience we use an equivalent representation of Eq. (6.4) in the form of matrix multiplication:

$$I_{2,m}^T = A^m \underbrace{\begin{bmatrix} I_{1,1}^T \\ I_{1,2}^T \\ \vdots \\ I_{1,N_1}^T \end{bmatrix}}_{I_1}, \quad \forall m = 1,2,\dots,N_1, \tag{6.5}$$

where A^m is a matrix of dimensions $N_2 \times N_1N_2$ whose entries are determined by the horizontal and vertical components of the motion field in the m^{th} row, i.e., $\mathbf{m}^h(m,.)$ and $\mathbf{m}^v(m,.)$. The elements of the matrix A^m are given by

$$A^m(n, n + \beta_1 + \beta_2 N_2) = \begin{cases} 1 & \text{if } \mathbf{m}^h(m,n) = \beta_1, \mathbf{m}^v(m,n) = \beta_2, \\ 0 & \text{otherwise.} \end{cases} \tag{6.6}$$

If $n + \beta_1 + \beta_2 N_2 > N_1N_2$ (e.g., at image boundaries), we set $n + \beta_1 + \beta_2 N_2 = N_1N_2$ so that the dimensions of the matrix A^m stays $N_2 \times N_1N_2$. It is easy to check that the matrix A^m in Eq. (6.6) contains only one

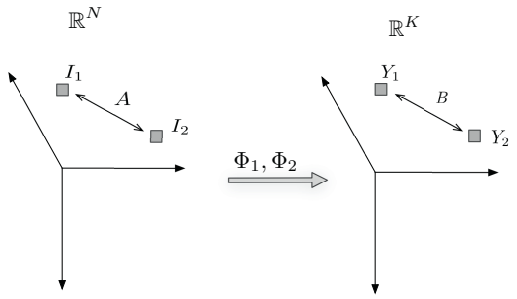


Figure 6.2: Illustration of the relation between the matrices A and B. On the left, the images $I_1, I_2 \in \mathbb{R}^N$ are related using the matrix A with $N = N_1 N_2$. In the compressed domain, the measurement vectors Y_1, Y_2 are related using the matrix B, where $K = MN_1$. The matrices A and B can be related by $B \approx \Phi_2 A \Phi_1^\dagger$, where Φ_1 and Φ_2 are the sensing matrices.

entry '1' in each row; this implies $I_{2,m}(n) = I_{1,m+\beta_1}(n+\beta_2)$ if $A^m(n, n+\beta_1+\beta_2 N_2) = 1$. The action of the matrix A^m shifts the pixels in I_1 according to the motion given as $\mathbf{m}^h(m,.)$ and $\mathbf{m}^v(m,.)$, and forms an estimate of the image $I_{2,m}$. It should be noted that the matrix A^m is completely determined by the m^{th} row of the motion vectors $\mathbf{m}^h(m,.)$ and $\mathbf{m}^v(m,.)$.

The relation given in Eq. (6.5) can be extended to all rows of I_2. The images I_1 and I_2 are finally related by a linear operator A such that $I_2 = A\, I_1$, which can be written as

$$\underbrace{\begin{bmatrix} I_{2,1}^T \\ I_{2,2}^T \\ \vdots \\ I_{2,N_1}^T \end{bmatrix}}_{I_2} = \underbrace{\begin{bmatrix} A^1 \\ A^2 \\ \vdots \\ A^{N_1} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} I_{1,1}^T \\ I_{1,2}^T \\ \vdots \\ I_{1,N_1}^T \end{bmatrix}}_{I_1}. \tag{6.7}$$

This relation is illustrated on the lefthand side of Fig. 6.2.

6.2.3 Relation between measurements

We now extend the above correlation model to the compressed domain. Without loss of generality we first assume that the measurements Y_1 and Y_2 can be related by a linear transformation B, i.e.,

$$\underbrace{\begin{bmatrix} Y_{2,1} \\ Y_{2,2} \\ \vdots \\ Y_{2,N_1} \end{bmatrix}}_{Y_2} = \underbrace{\begin{bmatrix} B^1 \\ B^2 \\ \vdots \\ B^{N_1} \end{bmatrix}}_{B} \underbrace{\begin{bmatrix} Y_{1,1} \\ Y_{1,2} \\ \vdots \\ Y_{1,N_1} \end{bmatrix}}_{Y_1}, \tag{6.8}$$

where B^m, $\forall m = 1, 2, \ldots, N_1$ is a matrix with dimensions $M \times MN_1$. In other words, the measurements $Y_{2,m}$ can be related to Y_1 as

$$Y_{2,m} = B^m\, Y_1,\ \forall m = 1, 2, \ldots, N_1. \tag{6.9}$$

Any two vectors $Y_1, Y_2 \in \mathbb{R}^{MN_1}$ can be related by a linear transformation B as long as $Y_1 \neq \mathbf{0}$, which is the case in our framework. The proof is given in the following proposition.

Proposition 2. *Let x and y be two vectors in $\mathbb{R}^K$. We show that two vectors can be related by $Bx = y$ if*

$x \neq \mathbf{0}$.

Proof: The system of linear equations $Bx = y$ can be written as

$$\begin{bmatrix} b_{11} & b_{12} & ... & ... & b_{1K} \\ b_{21} & b_{22} & ... & ... & b_{2K} \\ b_{31} & b_{32} & ... & ... & b_{3K} \\ ... & & & & \\ ... & & & & \\ b_{K1} & b_{K2} & ... & ... & b_{KK} \end{bmatrix}_{K\times K} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ ... \\ ... \\ x_K \end{bmatrix}_{K\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ ... \\ ... \\ y_K \end{bmatrix}_{K\times 1}$$

The above equation can be rewritten as $Xb = y$, where X and b are defined as

$$X = \begin{bmatrix} x_1 & x_2 & ... & x_K & 0 & ... & 0 & ... & 0 & ... 0 \\ 0 & 0 & ... & 0 & x_1 & ... & x_K & ... & 0 & ... 0 \\ .... & & & & & & & & & \\ 0 & 0 & ... & 0 & 0 & ... & 0 & ... & x_1 & ... x_K \end{bmatrix}_{K\times K^2}, b = \begin{bmatrix} b_{11} \\ b_{12} \\ ... \\ b_{1K} \\ b_{21} \\ ... \\ b_{2K} \\ ... \\ ... \\ b_{K1} \\ ... \\ b_{KK} \end{bmatrix}_{K^2\times 1}$$

The solution to the above equation $Xb = y$ exists if $y \in span(X)$. In other words, b can be found if $y \in span(X)$, where $span(X)$ refers to the subspace spanned by the columns of X. It is clear that the columns of matrix X forms an orthogonal basis in $\mathbb{R}^K$ if the vector x contains at least one non-zero value, i.e., any of $x_i \neq 0$ for $i = \{1, 2, \ldots, K\}$. The matrix X therefore spans the space $\mathbb{R}^K$, and hence $y \in \mathbb{R}^K$. □

The linear transformation B that relates the measurements Y_1 and Y_2 a priori does not have any special form. We are interested in understanding the relation between this matrix and the matrix A that shifts the pixels between images I_1 and I_2. Pre-multiplying Eq. (6.7) by Φ_2 on both sides, one can write

$$Y_2 = \Phi_2 I_2 = \Phi_2 A I_1. \tag{6.10}$$

In addition, by replacing $Y_1 = \Phi_1 I_1$ in Eq. (6.8), one obtains

$$Y_2 = BY_1 = B\Phi_1 I_1. \tag{6.11}$$

From Eqs. (6.10) and (6.11), the relation between B and A can finally be given as

$$B\Phi_1 = \Phi_2 A. \tag{6.12}$$

This forms an over-determined system of linear equations, as the number of unknown in matrix B is smaller than the number of equations in $\Phi_2 A$. In this case, the optimal matrix $\hat{B}$ that minimizes $\|B\Phi_1 - \Phi_2 A\|_2$ is given by

$$\hat{B} = \Phi_2 A {\Phi_1}^{\dagger}, \tag{6.13}$$

where $\dagger$ denotes the pseudo-inverse operator. As the rows in Φ_1 are generally orthonormal, the pseudo-inverse of Φ_1 can be computed using the transpose operator, i.e., ${\Phi_1}^{\dagger} = {\Phi_1}^T$. Substituting $\hat{B} = \Phi_2 A {\Phi_1}^T$ in

Eq. (6.8), the relation between the measurements becomes

$$Y_2 \approx \Phi_2 A {\Phi_1}^{\dagger} Y_1 = \Phi_2 A {\Phi_1}^{T} Y_1, \tag{6.14}$$

$$Y_{2,m} \approx \phi_2^m A^m {\Phi_1}^{T} Y_1, \ \forall m = 1, 2, \ldots, N_1, \tag{6.15}$$

where Eq. (6.15) comes from the fact that measurements are computed across rows of pixels in our framework. The relationship between the matrices A and B is illustrated in Fig. 6.2, where the matrices A and B are used to relate the points in the original and compressed domains, respectively. In the next section, we propose an algorithm for estimating the correlation model directly from the linear measurements Y_1 and Y_2.

6.3 Correlation estimation from linear measurements

6.3.1 Regularized energy minimization problem

We propose in this section a method for estimating the correlation between images from the compressed measurements without any explicit image reconstruction step. In particular, we concentrate on estimating the correlation model in the unquantized scenarios. Later, in Section 6.5, we describe our extension to the quantized scenarios. Our objective is to compute an optical flow or motion field that represents the motion between images I_1 and I_2. We denote this field as $\mathcal{M} = (\mathbf{m}^h, \mathbf{m}^v)$, where $\mathbf{m}^h$ and $\mathbf{m}^v$ are horizontal and vertical components of the motion respectively. The problem is then to find the value of the motion field $\mathcal{M}$ at each pixel position $\mathbf{z} = (m, n)$ such that the estimated correlation is consistent with the measurement vectors Y_1 and Y_2. At the same time, the motion field has to be piecewise smooth in order to model consistent motion of visual objects. We propose to cast the correlation estimation problem as a regularized energy minimization problem, where the energy $E(\mathcal{M})$ is composed of a data fidelity term $E_d(\mathcal{M})$ and a smoothness term $E_s(\mathcal{M})$. The optimal mapping $\mathcal{M}^*$ is obtained by minimizing the energy function $E(\mathcal{M})$ as

$$\mathcal{M}^* = \arg\min_{\mathcal{M}} \ E(\mathcal{M}) = \arg\min_{\mathcal{M}} [E_d(\mathcal{M}) + \lambda E_s(\mathcal{M})], \tag{6.16}$$

where λ balances the trade-off between the data and smoothness terms.

We now discuss in more details the components of the energy function. The smoothness term measures the penalty of assigning different motion values to adjacent pixels. We write it as

$$E_s(\mathcal{M}) = \sum_{\mathbf{z}, \mathbf{z}' \in \mathcal{N}} min(|\mathbf{m}^h(\mathbf{z}) - \mathbf{m}^h(\mathbf{z}')| + |\mathbf{m}^v(\mathbf{z}) - \mathbf{m}^v(\mathbf{z}')|, \tau), \tag{6.17}$$

where τ sets an upper bound on the smoothness penalty that helps preserving the discontinuities [130], and $\mathbf{z}$ and $\mathbf{z}'$ represent adjacent pixels in the 4-pixel neighborhood $\mathcal{N}$.

Next, the data term measures the consistency of a particular motion value for pixel $\mathbf{z}$ with the vectors Y_1 and Y_2. Classically, the accuracy of the motion value is evaluated with the original images [8], and the data cost is typically given by,

$$\tilde{E}_d(\mathcal{M}) = \sum_{m=1}^{N_1} \sum_{n=1}^{N_2} \Delta_{\mathcal{M}}(m, n), \tag{6.18}$$

where $\Delta_{\mathcal{M}}(m, n) = \|I_{2,m}(n) - I_{1, m+\mathbf{m}^h(m,n)}(n + \mathbf{m}^v(m, n))\|_2^2$ represents the error of matching the pixel at position (m, n) in the second image with a pixel in the first image that is selected according to the motion information. As discussed in Section 6.2.2, the effect of motion between images can be captured by a linear

operator A that is a composition of sub-matrices $\{A^m\}$. We can therefore rewrite Eq. (6.18) as

$$\tilde{E}_d(\mathcal{M}) = \sum_{m=1}^{N_1} \|I_{2,m}^T - A_{\mathcal{M}}^m I_1\|_2^2, \tag{6.19}$$

where the sub-matrix A^m depends on the motion field $\mathcal{M}$ according to Eq. (6.6). In the rest of the development, we drop the index $\mathcal{M}$, as the dependency on the motion field is clear from the context.

In our framework however, we do not have access to the original images but only to the measurement vectors Y_1 and Y_2. We thus approximate the data cost $\tilde{E}_d(\mathcal{M})$ by another term $E_d(\mathcal{M})$ that is computed directly from the measurement vectors. It can be written as

$$\begin{aligned}
\tilde{E}_d(\mathcal{M}) &= \sum_{m=1}^{N_1} \|I_{2,m}^T - A^m I_1\|_2^2 && (6.20)\\
&\approx \sum_{m=1}^{N_1} \|Y_{2,m} - B^m Y_1\|_2^2 && (6.21)\\
&\approx \sum_{m=1}^{N_1} \|Y_{2,m} - \phi_2^m A^m {\Phi_1}^T Y_1\|_2^2 && (6.22)\\
&= E_d(\mathcal{M}). && (6.23)
\end{aligned}$$

Note that the data cost approximation due to working in the compressed domain comes from the approximate relationship between the matrices A and B that is given as $B \approx \Phi_2 A \Phi_1^T$. We study in details the effect of this approximation in the next section.

We can finally rewrite the regularized energy objective function for the correlation estimation problem. It reads

$$E(\mathcal{M}) = \sum_{m=1}^{N_1} \|Y_{2,m} - \phi_2^m A_{\mathcal{M}}^m {\Phi_1}^T Y_1\|_2^2 + \lambda \sum_{\mathbf{z},\mathbf{z}' \in \mathcal{N}} min(|\mathbf{m}^h(\mathbf{z}) - \mathbf{m}^h(\mathbf{z}')| + |\mathbf{m}^v(\mathbf{z}) - \mathbf{m}^v(\mathbf{z}')|, \tau). \tag{6.24}$$

This cost function is used in the optimization problem of Eq. (6.16), which usually becomes a non-convex problem. The search space is discrete and usually constrained by limits on each of the motion values, which typically define a motion search window. The solution to this problem can be determined with strong optimization techniques based on Graph Cuts [130, 133] or Belief propagation [132]. A comprehensive overview of various energy minimization techniques are summarized in [152]. In this chapter, we use an optimization algorithm based on the α-expansion mode in Graph Cuts [130, 133].

Finally, it should be noted that the correlation estimation can also be performed by block of pixels. In this case, each block of pixels is assumed to move in a coherent way, and the objective of the correlation estimation problem is to compute one motion vector per block. The data cost function can then be modified in a straightforward way by imposing the same motion vectors for all the pixels in a block. The smoothness function is also modified in this case such that it penalizes the difference between the motion values of adjacent blocks rather than neighboring pixels. The optimization problem keeps the same form in block-based motion estimation but the search space is dramatically reduced as the number of motion vectors is smaller.

6.3.2 Compressed domain penalty

We now discuss the penalty of estimating the correlation from measurements instead of the original images. When the correlation $\mathcal{M}$ is given, this penalty corresponds to the difference between the values of the

regularized energy function of Eq. (6.16) that is evaluated from the original images and the respective linear measurements. Recall that the smoothness cost function $E_s(\mathcal{M})$ depends only on the correlation (see Eq. (6.17)). Therefore, the estimation penalty is identical to the error between the data cost functions $\tilde{E}_d(\mathcal{M})$ and $E_d(\mathcal{M})$ that are computed in the original and compressed domains respectively. In what follows, we first show that this penalty is bounded. Then, we show that the penalty decreases monotonically when the number of measurements increases.

Proposition 3. *The penalty of estimating the correlation from measurements is bounded. In particular we have* $|(1-\delta)^2\tilde{E}_d(\mathcal{M}) - C_l| \leq E_d(\mathcal{M}) \leq (1+\delta)^2\tilde{E}_d(\mathcal{M}) + C_u$, *where* $\delta > 0$, $C_l = \eta^2 + 2(1-\delta)\alpha\eta$, $C_u = \eta^2 + 2(1+\delta)\alpha\eta$, $\alpha = \sum_{m=1}^{N_1} \|I_{2,m}^T - A^m I_1\|_2$, $\eta = \sum_{m=1}^{N_1} \sigma_{max}(A^m) \sum_{q=m-w_y}^{m+w_y} \|\grave{I}_{1,q} - I_{1,q}\|_2$ *with* $\grave{I}_1 = \Phi_1^T Y_1$.

Proof: Let us assume that $\mathcal{M}$ or equivalently A is given. Then, the points $\mathcal{P} = \{I_{2,m}^T, A^m I_1 : m = 1, 2, \ldots, N_1\}$ forms a finite set. According to the Johnson-Lindenstrauss (JL) lemma, the distances between points in $\mathcal{P}$ are preserved in the measurement domain $\mathbb{R}^M$ when $M = \mathcal{O}(\delta^{-2}\, log|\mathcal{P}|)$ measurements are computed with a measurement matrix ϕ_2^m [153, 154], where $|\mathcal{P}|$ denotes the number of points in $\mathcal{P}$. Mathematically, the JL-embedding is given as

$$(1-\delta)\|I_{2,m}^T - A^m I_1\|_2 \leq \|\phi_2^m I_{2,m}^T - \phi_2^m A^m I_1\|_2 \leq (1+\delta)\|I_{2,m}^T - A^m I_1\|_2, \tag{6.25}$$

for a positive constant δ. It should be noted that, when the measurement matrix ϕ_2^m satisfies Eq. (6.25), then with high probability it satisfies the restricted isometry property (RIP). For more details related to the connection between the JL-lemma and the RIP we refer the reader to [153, 154]. Eq. (6.25) holds with high probability not only for measurement matrices constructed using Gaussian and Bernoulli distributions but also for structured measurement matrices constructed using orthonormal bases, e.g., DCT, FFT [154]. In our experiments we construct measurement matrices using structured FFT.

For a given row index m, the term $Y_{2,m} - \phi_2^m A^m \Phi_1^T Y_1$ in Eq. (6.22) can be written as

$$\begin{aligned} Y_{2,m} - \phi_2^m A^m \Phi_1^T Y_1 &= \phi_2^m I_{2,m}^T - \phi_2^m A^m \Phi_1^T \Phi_1 I_1 \\ &= \phi_2^m I_{2,m}^T - \phi_2^m A^m I_1 + \phi_2^m A^m I_1 - \phi_2^m A^m \Phi_1^T \Phi_1 I_1 \\ &= \phi_2^m I_{2,m}^T - \phi_2^m A^m I_1 + E I_1, \end{aligned} \tag{6.26}$$

where $E = \phi_2^m A^m - \phi_2^m A^m \Phi_1^T \Phi_1$. The term $\|Y_{2,m} - \phi_2^m A^m \Phi_1^T Y_1\|_2$ can be upper bounded as

$$\begin{aligned} \|Y_{2,m} - \phi_2^m A^m \Phi_1^T Y_1\|_2 &= \|\phi_2^m I_{2,m}^T - \phi_2^m A^m I_1 + E I_1\|_2 \\ &\leq \|\phi_2^m I_{2,m}^T - \phi_2^m A^m I_1\|_2 + \|E I_1\|_2 \\ &\leq (1+\delta)\|I_{2,m}^T - A^m I_1\|_2 + \|E I_1\|_2, \end{aligned} \tag{6.27}$$

where the last inequality is derived from Eq. (6.25). Similarly the term $\|Y_{2,m} - \phi_2^m A^m \Phi_1^T Y_1\|_2$ can be lower bounded as

$$\begin{aligned} \|Y_{2,m} - \phi_2^m A^m \Phi_1^T Y_1\|_2 &= \|\phi_2^m I_{2,m}^T - \phi_2^m A^m I_1 + E I_1\|_2 \\ &= \|\phi_2^m I_{2,m}^T - \phi_2^m A^m I_1 - (-E I_1)\|_2 \\ &\geq |\|\phi_2^m I_{2,m}^T - \phi_2^m A^m I_1\|_2 - \|E I_1\|_2| \qquad (6.28) \\ &\geq |(1-\delta)\|I_{2,m}^T - A^m I_1\|_2 - \|E I_1\|_2|, \qquad (6.29) \end{aligned}$$

where Eq. (6.28) follows from $\|x-y\|_2 \geq |\|x\|_2 - \|y\|_2|$, and Eq. (6.29) is derived from Eq. (6.25). The

term $\|EI_1\|_2$ in Eqs. (6.27) and (6.29) can also be bounded as

$$\begin{aligned}
\|EI_1\|_2 &= \|\phi_2^m A^m I_1 - \phi_2^m A^m \Phi_1^T \Phi_1 I_1\|_2 \\
&= \|\phi_2^m A^m (\Phi_1^T \Phi_1 I_1 - I_1)\|_2 \\
&\leq \|A^m (\grave{I}_1 - I_1)\|_2 \qquad (6.30) \\
&\leq \|A^m\|_2 \sum_{q=m-w_y}^{m+w_y} \|\grave{I}_{1,q} - I_{1,q}\|_2 \\
&= \sigma_{max}(A^m) \sum_{q=m-w_y}^{m+w_y} \|\grave{I}_{1,q} - I_{1,q}\|_2 = \eta_m, \qquad (6.31)
\end{aligned}$$

where Eq. (6.30) follows from $\|\Phi x\|_2 \leq \|x\|_2$, as Φ is a non expanding operator [153] and $\grave{I}_1 = \Phi_1^T \Phi_1 I_1$ is the pre-image of I_1. The parameter $\sigma_{max}(A^m)$ in Eq. (6.31) denotes the largest singular value of A^m. The summation in Eq. (6.31) is carried out from rows $m - w_y$ to $m + w_y$ as the search window is usually bounded, where w_y is the admissible search size along the vertical direction. Combining Eq. (6.27), Eq. (6.29) and Eq. (6.31) and by taking squares we get for each row of pixels

$$\left(|(1-\delta)\|I_{2,m}^T - A^m I_1\|_2 - \eta_m|\right)^2 \leq \|Y_{2,m} - \phi_2^m A^m \Phi_1^T Y_1\|_2^2 \leq \left((1+\delta)\|I_{2,m}^T - A^m I_1\|_2 + \eta_m\right)^2. \qquad (6.32)$$

Adding the second and third inequality terms of Eq. (6.32) for all rows $m = 1, 2, \ldots, N_1$ results in

$$\begin{aligned}
E_d(\mathcal{M}) &= \sum_{m=1}^{N_1} \|Y_{2,m} - \phi_2^m A^m \Phi_1^T Y_1\|_2^2 \\
&\leq (1+\delta)^2 \sum_{m=1}^{N_1} \|I_{2,m}^T - A^m I_1\|_2^2 + \sum_{m=1}^{N_1} \eta_m^2 + \sum_{m=1}^{N_1} 2(1+\delta)\eta_m \|I_{2,m}^T - A^m I_1\|_2 \\
&\leq (1+\delta)^2 \tilde{E}_d(\mathcal{M}) + \left(\sum_{m=1}^{N_1} \eta_m\right)^2 + 2(1+\delta) \underbrace{\sum_{m=1}^{N_1} \eta_m}_{\eta} \underbrace{\sum_{m=1}^{N_1} \|I_{2,m}^T - A^m I_1\|_2}_{\alpha} \\
&= (1+\delta)^2 \tilde{E}_d(\mathcal{M}) + \eta^2 + 2(1+\delta)\eta\alpha \\
&= (1+\delta)^2 \tilde{E}_d(\mathcal{M}) + C_u, \qquad (6.33)
\end{aligned}$$

where

$$C_u = \eta^2 + 2(1+\delta)\eta\alpha, \qquad (6.34)$$

and

$$\eta = \sum_{m=1}^{N_1} \eta_m = \sum_{m=1}^{N_1} \sigma_{max}(A^m) \sum_{q=m-w_y}^{m+w_y} \|\grave{I}_{1,q} - I_{1,q}\|_2. \qquad (6.35)$$

In a similar way, from the first and second inequality terms of Eq. (6.32) we get

$$\begin{aligned}
E_d(\mathcal{M}) &= \sum_{m=1}^{N_1} \|Y_{2,m} - \phi_2^m A^m \Phi_1^T Y_1\|_2^2 \\
&\geq \sum_{m=1}^{N_1} \left\{ |(1-\delta)^2 \|I_{2,m}^T - A^m I_1\|_2^2 + \eta_m^2 - 2(1-\delta)\eta_m \|I_{2,m}^T - A^m I_1\|_2 | \right\} \\
&\geq |\sum_{m=1}^{N_1} (1-\delta)^2 \|I_{2,m}^T - A^m I_1\|_2^2 + \sum_{m=1}^{N_1} \eta_m^2 - \sum_{m=1}^{N_1} 2(1-\delta)\eta_m \|I_{2,m}^T - A^m I_1\|_2| \\
&\geq |(1-\delta)^2 \sum_{m=1}^{N_1} \|I_{2,m}^T - A^m I_1\|_2^2 - \left(\sum_{m=1}^{N_1} \eta_m \right)^2 - 2(1-\delta) \sum_{m=1}^{N_1} \eta_m \sum_{m=1}^{N_1} \|I_{2,m}^T - A^m I_1\|_2| \\
&= |(1-\delta)^2 \tilde{E}_d(\mathcal{M}) - (\eta^2 + 2(1-\delta)\eta\alpha)| \\
&= |(1-\delta)^2 \tilde{E}_d(\mathcal{M}) - C_l| \qquad (6.36)
\end{aligned}$$

where

$$C_l = \eta^2 + 2(1-\delta)\eta\alpha, \qquad (6.37)$$

and η is given in Eq. (6.35). □

Proposition 4. *The penalty of estimating the correlation from measurements monotonically decreases when the measurement rate K/N increases. It further becomes negligible at high measurement rate.*

Proof: In Proposition 3 we have shown that the difference between the data cost functions estimated from compressed measurements $E_d(\mathcal{M})$ and images $\tilde{E}_d(\mathcal{M})$ is lower and upper bounded by errors C_l and C_u, respectively given in Eq. (6.37) and Eq. (6.34). The error $\eta = \sum_{m=1}^{N_1} \eta_m \propto \sum_{m=1}^{N_1} \|\grave{I}_{1,m} - I_{1,m}\|_2$ (see Eq. (6.35)) decreases with increasing measurement rate because $(\phi_1^m)^T \phi_1^m$ becomes an orthogonal projection operator, and $\grave{I}_1 = \Phi_1^T \Phi_1 I_1$ becomes arbitrarily close to I_1 when the number of measurements increases. Therefore, the errors C_l and C_u decrease as the measurement rate increases and when sufficient number of measurements are taken the errors C_l and C_u become negligible, i.e., $E_d(\mathcal{M}) \approx \tilde{E}_d(\mathcal{M})$. □

Due to the error between the cost functions $E_d(\mathcal{M})$ and $\tilde{E}_d(\mathcal{M})$, the solution $\mathcal{M}$ estimated from the linear measurements is not accurate especially at low measurement rates. The solution of the correlation estimation problem in the compressed domain might thus be quite far from the actual correlation between the images. However, as the number of measurements increases the approximation in the compressed domain becomes more accurate and the solution of the correlation estimation tends to the actual correlation between original images.

6.4 Experimental results

We analyze the performance of the proposed correlation estimation algorithm in both stereo and video imaging applications. The random projections are computed using a scrambled Fourier measurement matrix, where the scrambled operator is a diagonal matrix with entries ± 1 taken from an i.i.d. Bernoulli random variable with equal probability [154]. In all experiments, we sample both images using the same measurement rate. The correlation is estimated by minimizing the objective function in Eq. (6.24). The accuracy of the correlation estimation is evaluated in three ways: (1) comparing the estimated correlation with respect to the groundtruth information; (2) comparing the estimated correlation with the solution computed from the reconstructed images; (3) evaluating the quality of the second view that is reconstructed by prediction

(denoted as $\tilde{I}_2$) using the estimated correlation. In practice, the groundtruth correlation model and the original images are not available a priori to estimate an optimal regularization parameter λ. In such cases, the regularization parameter λ can be estimated based on learning from a set of training images or using the automated method proposed in [143]. In our experiments, we however select the parameter λ based on trial and error experiments.

6.4.1 Disparity estimation

For the stereo imaging case, we evaluate the disparity estimation performance in two natural image sets namely *Tsukuba* and *Venus*[2] [7]. These datasets have been captured by a camera array where the different viewpoints correspond to translating the camera along one of the image coordinate axis. In such a scenario, the motion of objects due to the viewpoint change is restricted to the horizontal direction with no motion along the vertical direction. Thus, the disparity estimation is a one-dimensional search problem, and the smoothness and data cost functions are modified accordingly by assuming that $\mathbf{m}^v = 0$. The size of the search windows used in our experiments are 16 pixels for Tsukuba and 20 pixels for Venus [7]. In our experiments we estimate disparity in both dense (per pixel) and block settings, where the block size is fixed to 4×4 pixels.

(a) $\mathbf{M}^h$ (b) $\mathbf{m}^h$ (c) $|\mathbf{M}^h - \mathbf{m}^h| > 1$ (d) $\mathbf{m}^h$ (e) $|\mathbf{M}^h - \mathbf{m}^h| > 1$

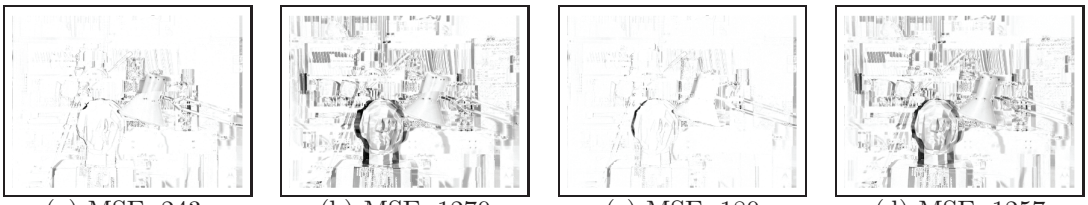
Figure 6.3: Comparison of the estimated disparity image with respect to groundtruth information at measurement rates 0.2 and 0.7 in the Tsukuba dataset. (a) Groundtruth disparity image $\mathbf{M}^h$; (b) computed dense disparity image $\mathbf{m}^h$ at measurement rate 0.2; (c) disparity error at rate 0.2. The pixels with absolute error greater than one is marked in white. The percentage of white pixels is 32%. (d) Computed dense disparity image $\mathbf{m}^h$ at measurement rate 0.7; (e) disparity error at rate 0.7. The percentage of white pixels is 10.3%.

(a) $\mathbf{M}^h$ (b) $\mathbf{m}^h$ (c) $|\mathbf{M}^h - \mathbf{m}^h| > 1$ (d) $\mathbf{m}^h$ (e) $|\mathbf{M}^h - \mathbf{m}^h| > 1$

Figure 6.4: Comparison of the estimated disparity image with respect to groundtruth information at measurement rates 0.2 and 0.7 in the Venus dataset. (a) Groundtruth disparity image $\mathbf{M}^h$; (b) computed dense disparity image $\mathbf{m}^h$ at measurement rate 0.2; (c) disparity error at rate 0.2. The pixels with absolute error greater than one is marked in white. The percentage of white pixels is 41%. (d) Computed dense disparity image $\mathbf{m}^h$ at measurement rate 0.7; (e) disparity error at rate 0.7. The percentage of white pixels is 10.7%.

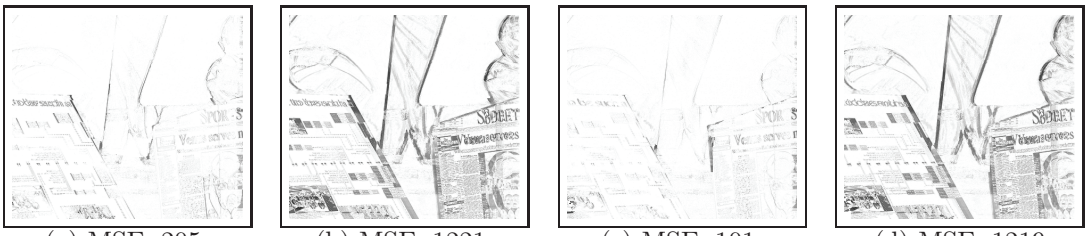
We first illustrate the disparity maps for the Tsukuba and Venus datasets, where the compressed data is obtained with a different measurement matrix for each image, i.e., $\Phi_1 \neq \Phi_2$. Fig. 6.3(b) and Fig. 6.4(b)

[2] Available at http://vision.middlebury.edu/stereo/data/scenes2001/



(a) MSE: 243 (b) MSE: 1270 (c) MSE: 180 (d) MSE: 1257

Figure 6.5: Evaluating the accuracy of disparity image in Fig. 6.3(b) and Fig. 6.3(d) in terms of image prediction quality for the Tsukuba dataset. The disparity images in Fig. 6.3(b) and Fig. 6.3(d) are used to predict the image $\tilde{I}_2$ at measurement rates 0.2 and 0.7 respectively. (a) Inverse prediction $1 - |\tilde{I}_2 - I_2|$ at a measurement rate of 0.2; (b) inverse prediction $1 - |\tilde{I}_2 - I_1|$ at a measurement rate of 0.2; (c) inverse prediction $1 - |\tilde{I}_2 - I_2|$ at a measurement rate of 0.7; (d) inverse prediction $1 - |\tilde{I}_2 - I_1|$ at a measurement rate of 0.7. The error is inverted, so that the white pixels correspond to no error.



(a) MSE: 205 (b) MSE: 1221 (c) MSE: 101 (d) MSE: 1219

Figure 6.6: Evaluating the accuracy of disparity image in Fig. 6.4(b) and Fig. 6.4(d) in terms of image prediction quality for the Venus dataset. The disparity images in Fig. 6.4(b) and Fig. 6.4(d) are used to predict the image $\tilde{I}_2$ at measurement rates 0.2 and 0.7 respectively. (a) Inverse prediction $1 - |\tilde{I}_2 - I_2|$ at a measurement rate of 0.2; (b) inverse prediction $1 - |\tilde{I}_2 - I_1|$ at a measurement rate of 0.2; (c) inverse prediction $1 - |\tilde{I}_2 - I_2|$ at a measurement rate of 0.7; (d) inverse prediction $1 - |\tilde{I}_2 - I_1|$ at a measurement rate of 0.7. The error is inverted, so that the white pixels correspond to no error.

show the disparity maps estimated from a measurement rate $K/N = 0.2$ for each dataset. Fig. 6.3(c) and Fig. 6.4(c) represent the corresponding disparity errors. Comparing these results with the respective groundtruth given in Fig. 6.3(a) and Fig. 6.4(a), we see that at low measurement rates (0.2 in this case) we estimate a coarse version of the disparity map. Quantitatively, the disparity errors with respect to the groundtruth are found out to be 41% and 32% respectively for the Tsukuba and Venus datasets, when it is measured as the percentage of pixels with absolute error greater than one [7]. We then estimate the disparity image at a measurement rate of 0.7. The results are shown in Fig. 6.3(d) and Fig. 6.4(d), and the corresponding disparity errors are available in Fig. 6.3(e) and Fig. 6.4(e). At higher rate, we see that the disparity map is more accurate and that the disparity error drops below 11% for both datasets.

We then evaluate in Fig. 6.5 and Fig. 6.6 the accuracy of the correlation estimation in terms of image prediction quality. In our experiments, we predict the second view from the first view using the estimated correlation information. When a coarse disparity map $\mathbf{m}^h$ (i.e., estimated at a low measurement rate) is used for image prediction, the resulting predicted image $\tilde{I}_2$ is closer to I_2 than I_1, as shown in Figs. 6.5(a) and (b) (resp. Figs. 6.6(a) and (b)). We observe that the mean square error (MSE) between the predicted image $\tilde{I}_2$ and I_2 is smaller than the error between $\tilde{I}_2$ and I_1, which confirms the benefit of the disparity estimation in the image prediction step. When the measurement rate increases, the quality of the disparity map improves. Meanwhile, the quality of the predicted image $\tilde{I}_2$ also improves substantially as observed in Figs. 6.5(c) and (d) (resp. Figs. 6.6(c) and (d)), where the measurement rate is set to 0.7.

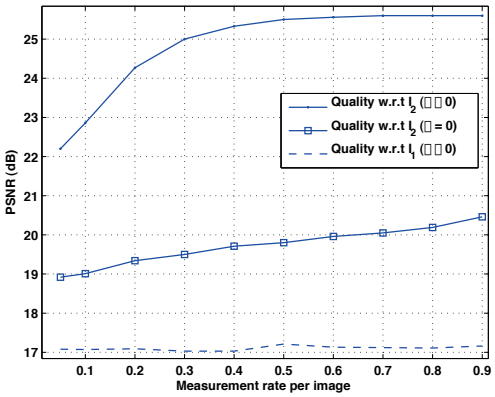
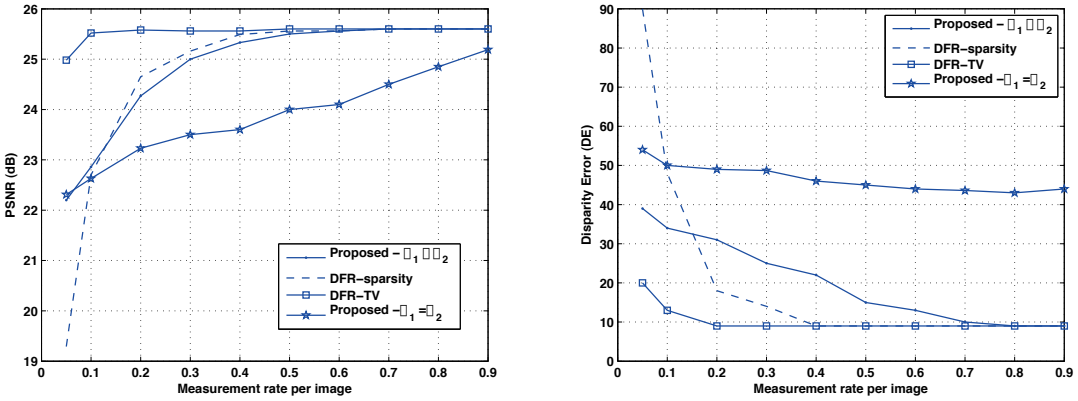


Figure 6.7: Comparison of the quality of predicted image $\tilde{I}_2$ with respect to I_2 and I_1 with and without regularization, i.e., $\lambda \neq 0$ and $\lambda = 0$ in Eq. (6.24) respectively in the Tsukuba dataset. The image prediction is carried out using dense disparity image.

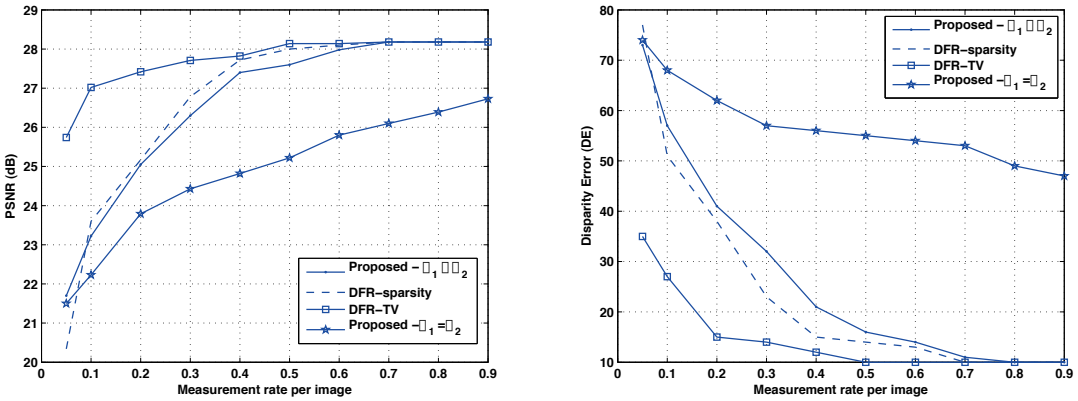
We now illustrate the benefit of regularization in the disparity estimation problem. Fig. 6.7 plots the quality of the predicted image $\tilde{I}_2$ with and without the smoothness cost in the correlation estimation problem (i.e., $\lambda \neq 0$ and $\lambda = 0$ respectively in Eq. (6.24)) for the Tsukuba dataset. It is clear that the quality of $\tilde{I}_2$ is improved by enabling the regularization term in our optimization framework. Similar behavior is also observed for the Venus dataset. From Fig. 6.7, we further notice that the quality of the predicted image $\tilde{I}_2$ at a measurement rate of 0.05 is measured as 22.2 dB (with corresponding disparity error 39%), which is approximately 3.5 dB away from the saturation point (or from the global minima solution) due to the influence of the penalty terms C_l and C_u discussed in Section 6.3.2. As the measurement rate increases the influence of the penalty terms C_l and C_u decreases. As a result, the quality of the predicted image $\tilde{I}_2$ increases and saturates with measurement rates above 0.5. In other words, our scheme gives accurate disparity estimation at high measurement rates. We also carry out experiments using the same measurement matrix for both images, i.e., $\Phi_1 = \Phi_2$. Fig. 6.8(a) and Fig. 6.8(b) compare the quality of the predicted image $\tilde{I}_2$ in terms of PSNR and the disparity error respectively, with the results obtained using different measurement matrices. It is clear that the prediction image quality and the disparity accuracy improve when different measurement matrices are used, since this brings more information from both images to solve the correspondence problem. Similar conclusions can be derived for the Venus dataset in Fig. 6.9.

We finally compare our disparity estimation results to a scheme that first reconstructs the images before estimating the disparity map. The images are reconstructed independently from the corresponding measurements by solving a convex optimization problem. We denote this methodology as *disparity from reconstructed images* (DFR). We have implemented two different reconstruction methodologies: (1) DFR-sparsity that minimizes the l_1 norm of the sparse coefficients assuming that the image is sparse in a particular orthonormal basis (e.g., a wavelet basis). This problem is solved using the GPSR [155]; (2) DFR-TV that minimizes the TV norm of the reconstructed image. This problem is solved using the BPDQ toolbox [156]. In both approaches, the disparity map is then estimated using the α-expansion mode in Graph Cuts applied on the reconstructed images. Fig. 6.8 and Fig. 6.9 show the comparison of the proposed scheme with DFR-sparsity and DFR-TV schemes, respectively for the Tsukuba and Venus datasets. From Fig. 6.8 and Fig. 6.9 we observe that the performance of our low complexity correlation solution competes with the DFR-sparsity scheme. At rates smaller than 0.1, our scheme performs even better than the DFR-sparsity scheme, as the poor quality of the reconstructed image in DFR-sparsity scheme leads to inaccurate estimation of the disparity map. On the other hand, the DFR-TV scheme significantly outperforms our scheme at low rates, due to the good quality of the reconstructed image. However, our scheme estimates an accurate correlation



(a) (b)

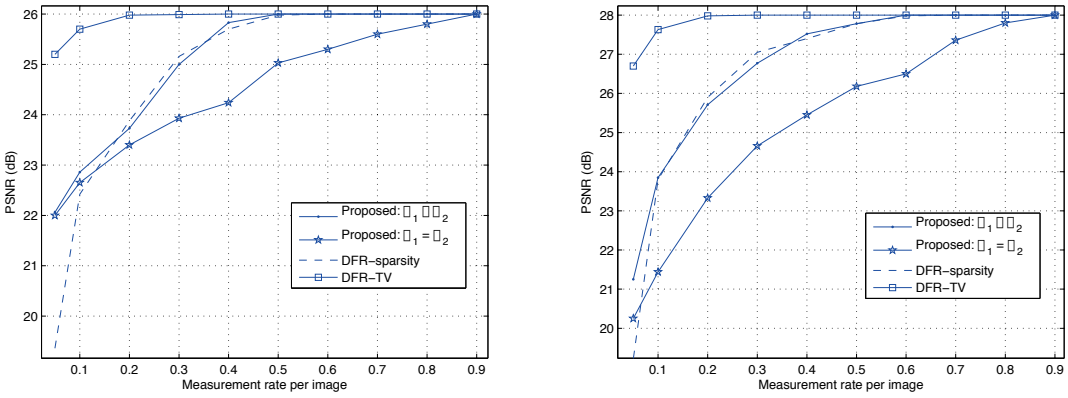
Figure 6.8: Comparison of the proposed scheme with the DFR schemes in the Tsukuba dataset: (a) comparison in terms of image prediction quality; (b) comparison in terms of disparity error. The performance of the proposed scheme is evaluated using both the same and different sets of measurement matrices.



(a) (b)

Figure 6.9: Comparison of the proposed scheme with the DFR schemes in the Venus dataset: (a) comparison in terms of image prediction quality; (b) comparison in terms of disparity error. The performance of the proposed scheme is evaluated using both the same and different sets of measurement matrices.

for rates above 0.5 and performs similarly to the DFR-TV scheme at high measurement rates, but with a complexity that is dramatically smaller due to the absence of image reconstruction. In particular, in our experiments we have observed that the running time of Graph Cuts algorithms that estimate the correlation information from linear measurements is approximately same as the one that estimates the correlation information from reconstructed images. The complexity of our correlation estimation scheme stays reasonable due to the efficiency of Graph Cuts algorithms whose complexity is bounded by a low order polynomial [130, 133]. Comparing to the DFR-sparsity and DFR-TV schemes, we save on the complexity corresponding to solving the l_2-l_1 and l_2-TV optimization problems respectively. It is however hard to precisely give the order of complexity of solving the l_2-l_1 and l_2-TV optimization problems, as it is highly depend on the



(a) (b)

Figure 6.10: Comparison of the proposed scheme with the DFR schemes: (a) Tsukuba dataset; (b) Venus dataset. The disparity is estimated per block with a block size 4×4. The performance of the proposed scheme is evaluated using both the same and different sets of measurement matrices.

type of solvers. For some of the popular solvers like GPSR [155] and NESTA [157], the order of complexity is given as $T\mathcal{O}(NlogN)$, where T is the number of iterations and N is the resolution of the image [158]. Therefore, comparing to the DFR schemes we save a complexity of $2T\mathcal{O}(NlogN)$.

The complexity of the proposed scheme can be further reduced when a disparity value is estimated per block instead of per pixel. Fig. 6.10(a) and Fig. 6.10(b) compare the performance of our scheme with the DFR schemes, when disparity is estimated per block with a block size 4×4. For the Tsukuba dataset, comparing Fig. 6.10(a) and Fig. 6.8(a) we see that the relative performances between the schemes remain approximately the same when the disparity image is estimated per block or per pixel. Similar conclusions can be derived for the Venus dataset by comparing Fig. 6.10(b) and Fig. 6.9(a). This confirms that the proposed scheme can easily adapt the granularity of disparity estimation without big penalty in order to meet the complexity requirements at the decoder.

6.4.2 Motion estimation

In the video scenario, we analyze the motion estimation accuracy in two synthetic scenes, namely *Yosemite* and *Grove*, and one natural scene, *Mequon*[3] [8]. The Grove and Mequon datasets are resampled to a resolution of 160×120 pixels using bilinear filters. The size of the search windows is ± 3 pixels in both horizontal and vertical directions [8]. For the sake of simplicity, we estimate motion field for blocks of pixels with size 4×4.

We first estimate motion vectors with different sensing matrices $\Phi_1 \neq \Phi_2$ for each image. These vectors are then used to predict the second image from the first image. Fig. 6.11 compares the predicted image $\tilde{I}_2$ with the original images I_2 and I_1 for the Yosemite and Grove datasets, respectively. It is clear that for a given measurement rate the predicted image $\tilde{I}_2$ is closer to I_2 than I_1, which indicates that the motion between images is efficiently captured by our correlation estimation algorithm. Similar experimental results are observed for the Mequon dataset. We then highlight the benefit of sampling the images with different sets of measurement matrices in Fig. 6.12. From Fig. 6.12, we see that the quality of the predicted image $\tilde{I}_2$ is better when the images are sampled with different measurement matrices, compared to the case where the same measurement matrix is used for all images. This is consistent with the results shown for the disparity

[3] Available at http://vision.middlebury.edu/flow

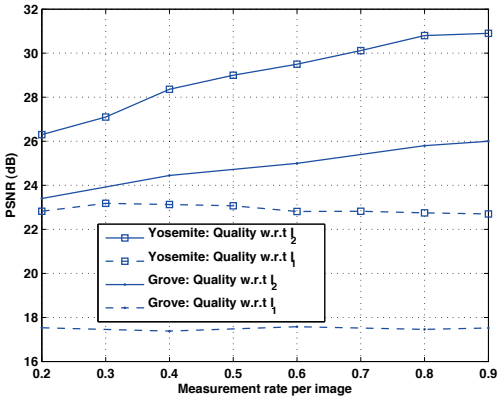
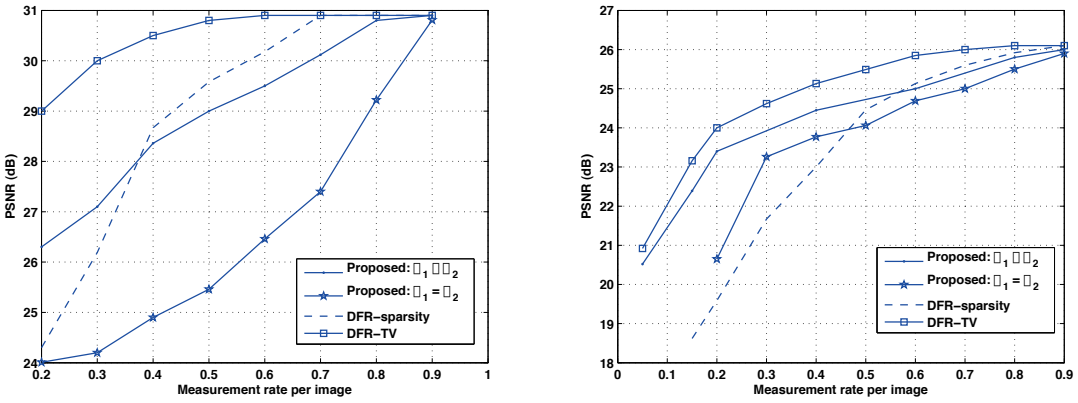


Figure 6.11: Illustration of the accuracy of motion field estimation in terms of image prediction quality for the Yosemite and Grove datasets. The quality of the predicted image $\tilde{I}_2$ is compared with respect to I_2 and I_1. The prediction is carried out using the motion field estimated with block of pixels 4×4.



(a) (b)

Figure 6.12: Comparison of the quality of the predicted image $\tilde{I}_2$ between the proposed, DFR-sparsity and DFR-TV schemes: (a) Yosemite dataset; (b) Grove dataset. The image prediction is carried out using the motion field that is estimated using block of pixels 4×4.

estimation performance. We finally compare our results to DFR-sparsity and DFR-TV schemes that build the correlation model from (independently) reconstructed images based on optimizations with the sparsity and TV priors respectively. Fig. 6.12(a) and Fig. 6.12(b) show the comparison for the Yosemite and Grove datasets respectively. From Fig. 6.12 we see that for both datasets the proposed scheme performs better than DFR-sparsity scheme at low rates, and is competitive at high rates, as already observed in the disparity estimation study. Similar experimental findings are observed for the Mequon dataset. Furthermore, we see that the proposed scheme is also competitive with the performance of DFR-TV scheme at low rate for the Grove dataset (see Fig. 6.12(b)). The reason is that the Grove scene is textured with limited low frequency components, therefore the TV prior results in poor reconstruction quality, especially in the textured areas. Overall, the proposed scheme provides effective motion estimation results, while avoiding an order of computational complexity $2T\mathcal{O}(NlogN)$ involved in the image reconstruction steps with the

DFR-TV and DFR-sparsity schemes.

6.5 Handling non-linearities of quantization

In this section, we extend our framework to efficiently handle the measurement noise, where the noise is introduced by uniform quantization of the measurements (see Fig. 6.1) denoted as $\hat{Y}_1$ and $\hat{Y}_2$. In our framework described in Section 6.3, the smoothness cost is independent of the observation model. Therefore, we modify the data cost function to consider efficiently the error introduced by measurement quantization.

6.5.1 Robust data cost

We propose to handle the non-linearities of quantization on Y_1 and Y_2 in a different way at the joint decoder because of their different impacts on the correlation estimation solution. We first describe how we account for the quantization non-linearities on the measurement vector Y_1 and then on Y_2. Let $Y_{1,m}(p)$ be the p^{th} coordinate of the original measurement in the m^{th} row, and $\hat{Y}_{1,m}(p)$ be the corresponding quantized value. Since the joint decoder has the quantized value $\hat{Y}_{1,m}(p)$, but not the original value $Y_{1,m}(p)$, it only knows that the quantized measurement lies within the quantized interval, i.e., $\hat{Y}_{1,m}(p) \in \mathcal{R}_{\hat{Y}_{1,m}(p)} = (r_p,\ r_{p+1}]$, where r_p and r_{p+1} define the lower and upper bounds of quantizer bin $\mathcal{Q}_p$. Given the quantized measurements $\hat{Y}_1$, we need to pick a measurement $\tilde{Y}_1$ such that $\Phi^T\tilde{Y}_1$ is closer to the original image I_1 than $\Phi^T\hat{Y}_1$, i.e.,

$$\|I_1 - \Phi^T\tilde{Y}_1\|_2 \leq \|I_1 - \Phi^T\hat{Y}_1\|_2. \tag{6.38}$$

By following the analysis described in Section 6.3.2, it is clear that an estimated measurement vector $\tilde{Y}_1$ that satisfies Eq. (6.38) is guaranteed to improve the quality of the correlation estimation. In practice, it is not possible to find $\tilde{Y}_1$ satisfying Eq. (6.38) at the decoder, due to the absence of the original image I_1. We therefore propose to jointly estimate a measurement $\tilde{Y}_1$ and an image $\bar{I}_1$ as a solution to the following optimization problem:

$$(\bar{I}_1, \tilde{Y}_1) = \min_{(\tilde{Y}_1, \bar{I}_1)} \|\Phi_1\bar{I}_1 - \tilde{Y}_1\|_2 \quad \text{s.t.} \quad \tilde{Y}_1 \in \mathcal{R}_{\hat{Y}_1}, \tag{6.39}$$

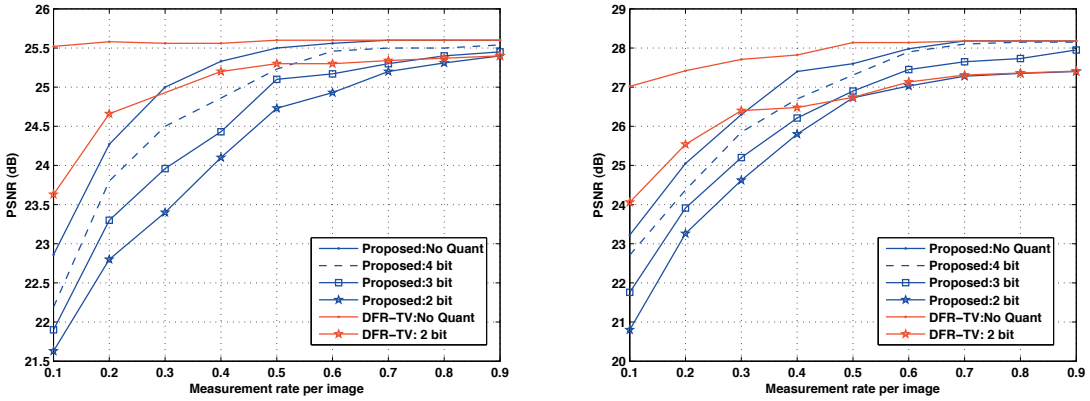
where $\mathcal{R}_{\hat{Y}_1}$ is the cartesian product of all the quantized regions $\mathcal{R}_{\hat{Y}_{1,m}(p)}$, i.e., $\mathcal{R}_{\hat{Y}_1} = \prod_{m,p} \mathcal{R}_{\hat{Y}_{1,m}(p)}$. The above optimization problem considers all the measurements $\tilde{Y}_1$ in the quantized interval $\mathcal{R}_{\hat{Y}_1}$, and jointly estimates $\bar{I}_1$ and $\tilde{Y}_1$ that minimize the objective function $\|\Phi_1\tilde{Y}_1 - \bar{I}_1\|_2$. The solution $\tilde{Y}_1$ estimated from Eq. (6.39) is guaranteed to satisfy Eq. (6.38), since the quantized measurements $\hat{Y}_1$ are included in the search space $\mathcal{R}_{\hat{Y}_1}$, i.e., $\hat{Y}_1 \in \mathcal{R}_{\hat{Y}_1}$. Using the estimated solution $\tilde{Y}_1$ we modify the data cost in Eq. (6.22) as

$$\hat{E}_{d,1} = \sum_m \|\hat{Y}_{2,m} - \phi_2^m A^m {\Phi_1}^T \tilde{Y}_1\|_2^2. \tag{6.40}$$

Now, we describe the proposed methodology to handle the quantization effect on the measurement vector Y_2. We propose to estimate a vector $\tilde{Y}_2$ that minimizes the distance $\|\tilde{Y}_{2,m} - \phi_2^m A^m {\Phi_1}^T \hat{Y}_1\|_2$ by considering all the valid measurement values in the quantization interval $\mathcal{R}_{\hat{Y}_{2,m}}$, where $\mathcal{R}_{\hat{Y}_{2,m}} = \prod_p \mathcal{R}_{\hat{Y}_{2,m}(p)}$. A measurement $\tilde{Y}_2$ is estimated by solving the following problem:

$$\tilde{Y}_2 = \min_{\tilde{Y}_2} \sum_m \|\tilde{Y}_{2,m} - \phi_2^m A^m {\Phi_1}^T \hat{Y}_1\|_2 \quad \text{s.t.} \quad \tilde{Y}_{2,m} \in \mathcal{R}_{\hat{Y}_{2,m}}. \tag{6.41}$$

Using the estimated vector $\tilde{Y}_2$ in Eq. (6.22), we get the robust data term for proper handling of the quanti-



(a) (b)

Figure 6.13: Effect of quantization on the quality of the predicted image $\tilde{I}_2$: (a) Tsukuba dataset; (b) Venus dataset. When the measurements are quantized we used $BPDN_p$ [48] to reconstruct the images in DFR-TV scheme. The image prediction is carried out using the dense disparity image estimated by solving Eq. (6.24).

zation effect on measurements Y_2. It is given as

$$\hat{E}_{d,2} = \sum_m \|\tilde{Y}_{2,m} - \phi_2^m A^m {\Phi_1}^T \hat{Y}_1\|_2^2. \tag{6.42}$$

The objective functions in Eq. (6.39) and Eq. (6.41) are convex, since they can be easily expressed as l_2 norm of a linear or affine function. Also, the regions $\mathcal{R}_{\hat{Y}_1}$ and $\mathcal{R}_{\hat{Y}_2}$ form a closed convex set as each region $\mathcal{R}_{\hat{Y}_{j,m}(p)} = (r_p,\ r_{p+1}], \forall j \in \{1, 2\}$ is a convex set. Therefore, the optimization problems given in Eq. (6.39) and Eq. (6.41) are (not strictly) convex.

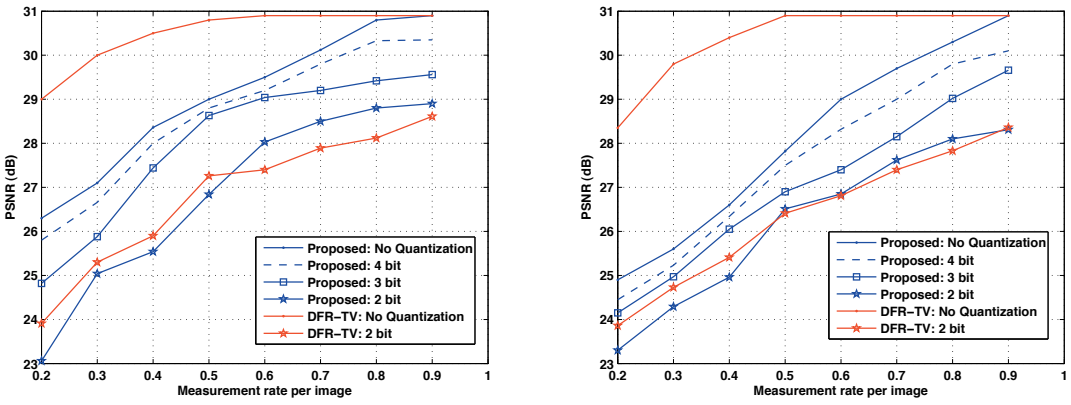
So far, we handle the non-linearities of the quantization independently on Y_1 and Y_2, and the respective modified cost functions are given in Eq. (6.40) and Eq. (6.42). We now describe the methodology to handle the effect simultaneously on both Y_1 and Y_2. One solution is to jointly estimate a pair of measurement vectors $\tilde{Y}_1$ and $\tilde{Y}_2$, by combining the optimization problems in Eq. (6.39) and Eq. (6.41). Such a joint estimation approach is however complex. Instead, we first estimate a vector $\tilde{Y}_1$ by solving Eq. (6.40) and it is used to solve Eq. (6.41), i.e., we solve Eq. (6.41) with the objective function $\|\tilde{Y}_{2,m} - \phi_2^m A^m {\Phi_1}^T \tilde{Y}_1\|_2$. By combining the results in Eq. (6.40) and Eq. (6.42) the robust data cost is given as

$$\hat{E}_d = \sum_m \|\tilde{Y}_{2,m} - \phi_2^m A^m {\Phi_1}^T \tilde{Y}_1\|_2^2, \tag{6.43}$$

where $\tilde{Y}_1$ is estimated by solving Eq. (6.39) and $\tilde{Y}_2$ is estimated by solving Eq. (6.41) with the objective function $\|\tilde{Y}_{2,m} - \phi_2^m A^m {\Phi_1}^T \tilde{Y}_1\|_2$. Finally, the robust correlation solution is estimated by solving the optimization problem in Eq. (6.16) using the modified data function $\hat{E}_d$ given in Eq. (6.43). At last, it should be noted that the roles of I_1 and I_2 (equivalently Y_1 and Y_2) could be reversed if we change the reference grid for the correlation estimation.

6.5.2 Performance analysis

We first estimate the correlation model by minimizing the energy objective function in Eq. (6.24) using the quantized measurements $\hat{Y}_1$ and $\hat{Y}_2$ without using robust data cost. Fig. 6.13 shows the effect of quantization on the disparity estimation performance when measurements are uniformly quantized using 2, 3 and 4 bits. We use the disparity map to predict the second image in the Tsukuba and Venus dataset. Interestingly, we see that the 4-bit quantizer does not significantly affect the quality of the disparity image, as the degradation stays below 0.5 dB in the quality of $\tilde{I}_2$ at low to medium rates. As expected, however, the quality of the predicted image $\tilde{I}_2$ decreases with increasing quantizer coarseness level for a fixed measurement rate. We then compare our results with DFR-TV scheme that first reconstructs the images by solving an optimization problem based on $BPDN_p$ in order to efficiently handle the quantization noise [48], and then use the reconstructed images for disparity estimation. From Fig. 6.13, we see that the performance gap between DFR-TV and compressed domain estimation is approximately the same in both the unquantized and quantized (i.e., 2-bit) scenarios. We show later in this section that the performance of our solution can be improved by effectively considering the non-linearities of the quantization.

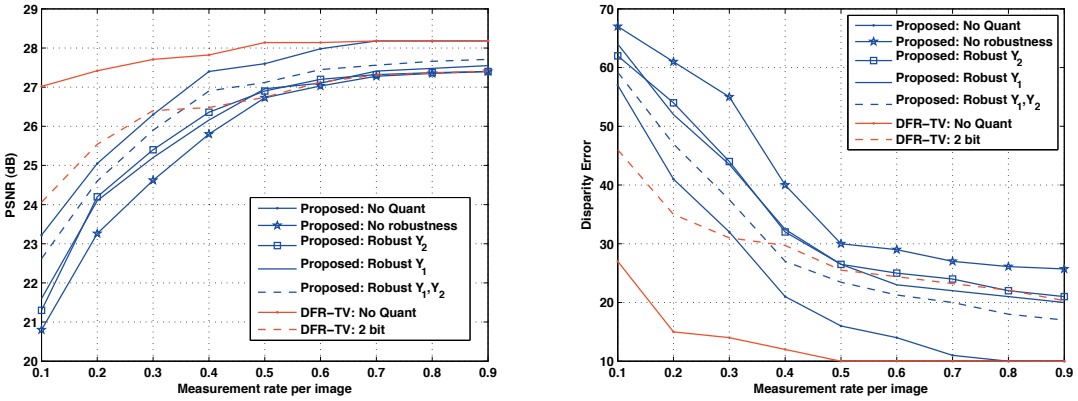


(a) (b)

Figure 6.14: Effect of using 4-, 3- and 2-bits quantizers on the image prediction quality for (a) Yosemite dataset and (b) Mequon dataset. When the measurements are quantized we used $BPDN_p$ [48] to reconstruct the images in DFR-TV scheme. The image prediction is carried out using the motion field estimated by solving Eq. (6.24).

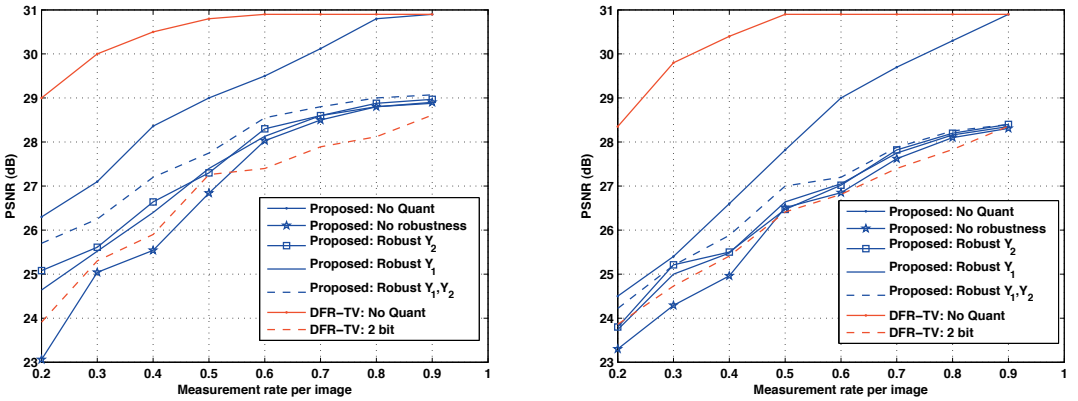
We then carry out similar experiments on the Yosemite and Mequon datasets. Fig. 6.14(a) and Fig. 6.14(b) show the quality of the predicted image, where the prediction is performed with the motion vectors that are directly estimated from quantized measurements, i.e., no robustness. As expected, the quality of the predicted image or equivalently the accuracy of motion estimation is reduced when the measurements are quantized. Similarly, in the case of disparity estimation, the influence of the quantization is negligible when the measurements are quantized with a 4-bit quantizer. We also compare our results with DFR-TV scheme that first reconstructs the images by solving an optimization problem based on $BPDN_p$, and then estimates a motion field from the reconstructed images. The performance of DFR-TV scheme for the 2-bit quantization scenario is shown in Fig. 6.14. Interestingly, when the measurements are quantized we see that the proposed scheme competes with the DFR-TV scheme; this is because of the poor image reconstruction performance in the DFR-TV scheme when the measurements are coarsely quantized (i.e., 2-bit quantizer).

We now carry out our experiments using the robust data cost functions given in Eqs. (6.40), (6.42) and (6.43). We use the CVX toolbox [151] to solve the optimization problems given in Eq. (6.39) and Eq. (6.41).



(a) (b)

Figure 6.15: Effect of the proposed robust data costs on the (a) image prediction quality and (b) disparity error in the Venus dataset. The measurements are quantized using a uniform 2-bit quantizer. When the measurements are quantized the images are reconstructed independently in DFR-TV scheme using $BPDN_p$ [48] .

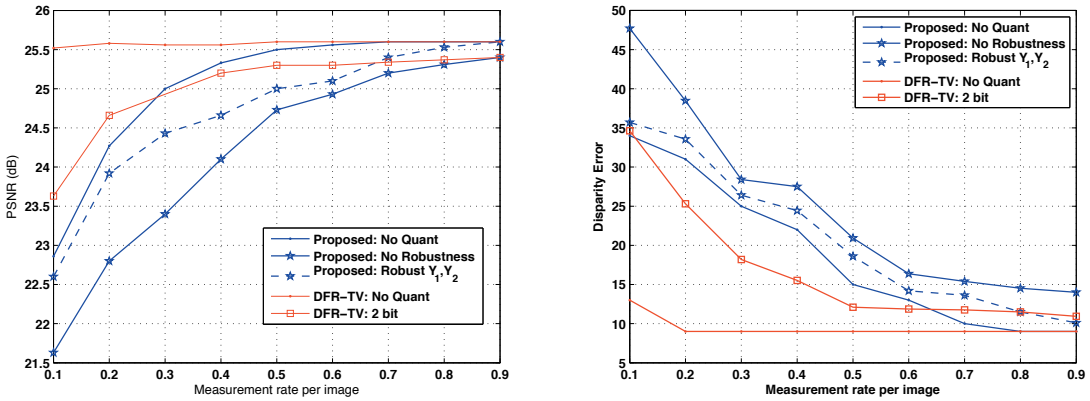


(a) (b)

Figure 6.16: Effect of robust data cost on the image prediction quality in the (a) Yosemite dataset and (b) Mequon dataset. The measurements are quantized using a uniform 2-bit quantizer. When the measurements are quantized the images are reconstructed independently in DFR-TV scheme using $BPDN_p$ scheme [48] .

We first quantify our results using a 2-bit quantizer. In particular, we consider the non-linearity effects only on Y_1. In order to consider this, we minimize the correlation energy function in Eq. (6.16) using the data cost function given in Eq. (6.40). The resulting experimental behavior is plotted in Fig. 6.15, Fig. 6.16(a) and Fig. 6.16(b) for the Venus, Yosemite and Mequon datasets respectively (marked as *Robust* Y_1). From the plots we observe that the correlation estimation performance is improved by considering the non-linearities of quantization on Y_1, when compared to the non-robust scheme that estimates a correlation model directly from the quantized measurements (marked as *No robustness*), especially at low measurement rates. We then carry out similar experiments to consider the non-linearity effects only on the second measurement vector Y_2, i.e., we minimize the energy function in Eq. (6.16) using the data cost $\hat{E}_{d,2}$ given in Eq. (6.42). The

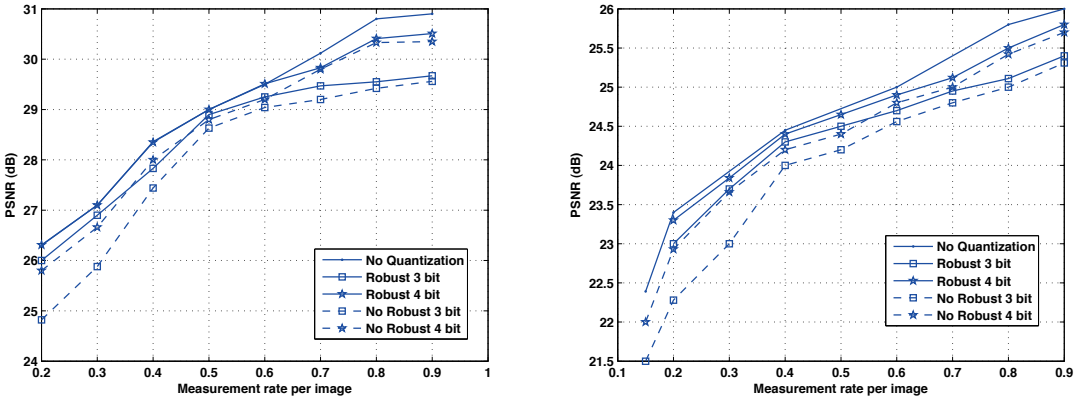
resulting performance is marked as *Robust* Y_2 in the plots shown in Fig. 6.15 and Fig. 6.16 for the respective datasets. From the plots we observe the benefit of our robust data term $\hat{E}_{d,2}$ that considers effectively the noise on the measurement vector Y_2. Further, it is interesting to note that the overall performance is similar for both robust data cost functions given in Eq. (6.40) and Eq. (6.42). In other words, due to the symmetric encoding scheme, we achieve same amount of gain (compared to the non-robust scheme) by considering effectively the quantization noises either on Y_1 or Y_2.



(a) (b)

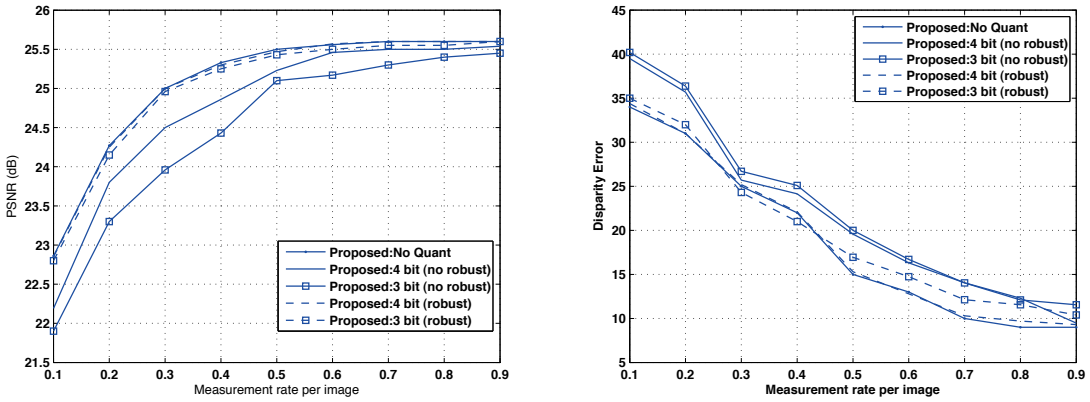
Figure 6.17: Effect of the proposed robust data costs on the (a) image prediction quality and (b) disparity error in the Tsukuba dataset. The measurements are quantized using a uniform 2-bit quantizer. When the measurements are quantized the images are reconstructed independently in DFR-TV scheme using $BPDN_p$ [48] .

We then consider the effects of quantization noises on both Y_1 and Y_2 using the data cost function given in Eq. (6.43). The resulting performance is shown in Fig. 6.15 and Fig. 6.17 for the Venus and Tsukuba stereo datasets (marked as *Robust* Y_1, Y_2). It is clear from the plots that our robust scheme leads to an improved estimation performance comparing to the non-robust scheme, especially at low to medium measurement rates. At high rates, the performance is not dramatically improved, due to the significant measurement quantization. Furthermore, from Fig. 6.15 we see that it is beneficial to consider the non-linearity effects on both Y_1 and Y_2 rather than considering it on one of them. Similar conclusions can be derived for the Yosemite and Mequon motion datasets shown in Fig. 6.16. It is interesting to note that our robust correlation estimation scheme outperforms the performance of DFR-TV scheme in the quantized scenario (see Fig. 6.16); this is certainly a promising and interesting result. Finally, we carry out similar experiments using the measurements that are quantized using 3 and 4 bits. In this scenario, we consider the quantization noise effects on both Y_1 and Y_2. The results are summarized in Fig. 6.18(a), Fig. 6.18(b) and Fig. 6.19 for the Yosemite, Grove and Tsukuba datasets respectively. From the plots, as expected, we see that the performance of our robust scheme outperforms the non-robust scheme. Furthermore, it is interesting to observe that our robust correlation estimation scheme competes with the performance of the non-quantization scheme, even when the measurements are coarsely quantized using 3 bits. These results demonstrate that the proposed robust estimation scheme provides an accurate correlation model even with significant quantization of the measurements.



(a) (b)

Figure 6.18: Effect of robust data cost on the image prediction quality in the (a) Yosemite dataset and (b) Grove dataset. The measurements are quantized using 3-bit and 4-bit uniform quantizers.



(a) (b)

Figure 6.19: Effect of the proposed robust data costs on the (a) image prediction quality and (b) disparity error in the Tsukuba dataset. The measurements are quantized using 3-bit and 4-bit uniform quantizers.

6.6 Extension to multi-view images

So far, we have focused on the correlation estimation between a pair of images. In this section, we describe our extension to more than two images. Let $I_1, I_2, \ldots, I_J$ denote the J correlated images, and the corresponding measurements generated from these images are given as $Y_1, Y_2, \ldots, Y_J$ respectively. We are interested in estimating a correlation model between J images from the corresponding measurements $Y_1, Y_2, \ldots, Y_J$. In this section, we focus on J correlated images that are captured in a multi-view imaging scenario. Nevertheless, our framework could also be extended to multiple images captured in a video sequence.

In multi-view imaging scenarios, the correlation model between J views is described using a depth image Z. As described in the previous chapter (see Section 5.6), we assume that the depth values Z are discretized such that the inverse depth $1/Z$ is uniformly sampled in the range $[1/Z_{max}, 1/Z_{min}]$, where Z_{min} and Z_{max}

are the minimal and maximal depth values in the scene respectively. The estimation problem is equivalent to finding a set of depth labels $l \in \mathcal{L}$ that effectively capture the depth information for each pixel $\mathbf{z}$ in the reference image, where $\mathcal{L}$ is a discrete set of labels. We propose to estimate a depth image in a regularized energy minimization framework as a trade-off between the data $E_{d,u}$ and smoothness $E_{s,u}$ costs. The data cost function estimates a depth label l for each pixel that best agrees with the J linear measurements, while the smoothness cost $E_{s,u}$ enforces consistency in the depth image. The energy model is given as

$$E_u(l) = E_{d,u}(l) + \lambda E_{s,u}(l), \tag{6.44}$$

where λ is a regularization constant.

The smoothness cost $E_{s,u}(l)$ enforces consistency in the depth values between the adjacent pixels $\mathbf{z}$ and $\mathbf{z}'$. For a given pair of adjacent pixels $\mathbf{z}$ and $\mathbf{z}'$, we propose to compute the smoothness penalty using the truncated absolute difference given as

$$V_{\mathbf{z},\mathbf{z}'} = min(|l(\mathbf{z}) - l(\mathbf{z}')|, \tau), \tag{6.45}$$

where τ is a constant that sets an upper bound on the penalty. The smoothness cost is then defined as the cumulative sum of the penalties given in Eq. (6.45) for all the adjacent pixels $\mathbf{z}, \mathbf{z}'$ in $\mathcal{N}$, i.e.,

$$E_{s,u}(l) = \sum_{\mathbf{z},\mathbf{z}' \in \mathcal{N}} min(|l(\mathbf{z}) - l(\mathbf{z}')|, \tau). \tag{6.46}$$

We now describe the data cost function that measures how well a particular depth label l fits with the J correlated measurements. Before deriving the expression for the data cost function $E_{d,u}$ in the compressed domain, we first derive a data cost expression $\tilde{E}_{d,u}$ as a function of the linear motion operator A in the image domain. Let $\mathcal{P}_i(\mathbf{z}, l)$ be a projection operator that projects the pixel $\mathbf{z} = (m, n)$ from the reference image I_1 to the i^{th} view through depth level $l \in \mathcal{L}$, as shown in Fig. 6.20. From Fig. 6.20 we see that the projection operator shifts the pixel $\mathbf{z} = (m, n)$ in reference image I_1 to the location $(m + \mathbf{m}^h(m, n)), l + \mathbf{m}^v(m, n))$ in view I_i, where $\mathbf{m}^h(m, n)$ and $\mathbf{m}^v(m, n)$ represent the horizontal and vertical motion components respectively. The amount of motion $\mathbf{m}^h$ and $\mathbf{m}^v$ for each pixel depends on the particular depth label l for a fixed camera configuration. Therefore, we can relate the reference view I_1 to the view I_i by a motion field. In other words, the images I_1 and I_i can be related as $I_i = A_i I_1$, where A_i is the matrix that shifts the pixels in the reference image I_1 to form I_i. It should be noted that the matrix A_i is a function of the depth label l. By extending the same argument to all views we can write the data cost between J images as

$$\tilde{E}_{d,u}(l) = \sum_{j=2}^{J} \|I_j - A_j I_1\|_2^2 \tag{6.47}$$

$$= \sum_{j=2}^{J} \sum_{m=1}^{N_1} \|I_{j,m}^T - A_j^m I_1\|_2^2, \tag{6.48}$$

where Eq. (6.48) follows from Eq. (6.7). It should be noted that the data cost $\tilde{E}_{d,u}$ in the multi-view scenario is similar to the stereo data function given in Eq. (6.19), except that the summation is carried out for all the views.

In our framework however, we do not have access to the original images but only to linear measurements computed from the images. Therefore, we approximate the data cost relating the J measurements using the relationship between the matrices A and B described in Section 6.2.3. The approximate data cost relating

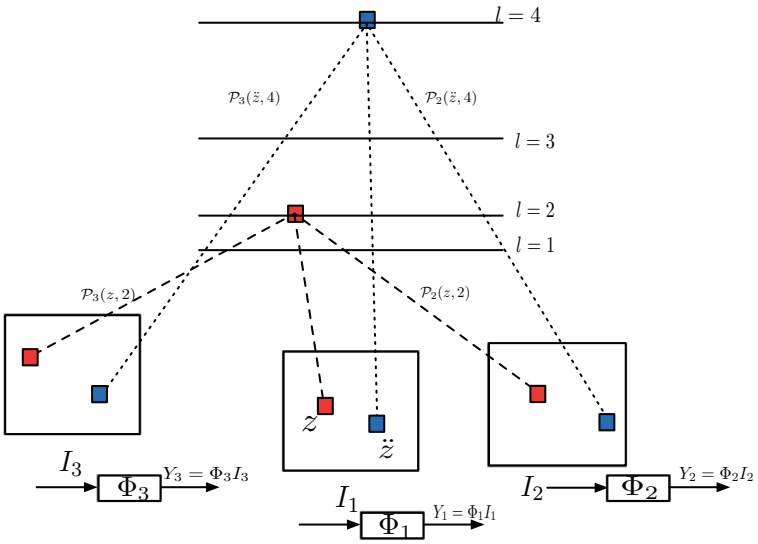


Figure 6.20: Depth estimation in the multi-view scenario. $\mathcal{P}_i(\mathbf{z}, l)$ represents a projection operator that projects the pixel $\mathbf{z}$ in view I_1 to the i^{th} view at depth label l. The pixel $\mathbf{z}$ in view I_1 and the projected pixel $\mathcal{P}_i(\mathbf{z}, l)$ in view I_i are related by a motion vector.

the J measurements is given as

$$\tilde{E}_{d,u}(l) = \sum_{j=2}^{J} \sum_{m=1}^{N_1} \|I_{j,m}^T - A_j^m I_1\|_2^2 \tag{6.49}$$

$$\approx \sum_{j=2}^{J} \sum_{m=1}^{N_1} \|Y_{j,m}^T - \phi_j^m A_j^m \Phi_1^T Y_1\|_2^2 \tag{6.50}$$

$$= E_{d,u}(l), \tag{6.51}$$

where the last equation is derived using the approximate relation between the matrices A and B, given as $B \approx \Phi_2 A \Phi_1^T$. Finally, the depth image can be estimated by minimizing Eq. (6.44) using Graph Cuts.

We now quantify the penalty of estimating the depth image in the compressed domain similarly to the analysis described in Section 6.3.2 with two images. The multi-view data cost function $E_{d,u}(l)$ in Eq. (6.50) is the cumulative sum of the stereo data cost function given in Eq. (6.23) for all the views. Using this property and repeating the steps in Proposition 3 and Proposition 4 one can prove the following results.

Corollary 1. *The penalty of estimating the depth image directly from the measurements is bounded. In particular we have* $|(1-\delta)^2 \tilde{E}_{d,u}(l) - \kappa_l| \leq E_{d,u}(l) \leq (1+\delta)^2 \tilde{E}_{d,u}(l) + \kappa_u$, *where* $\kappa_l = (J-1)C_l$ *and* $\kappa_u = (J-1)C_u$. *The terms* C_l *and* C_u *are given in Eq. (6.37) and Eq. (6.34) respectively.*

Corollary 2. *The penalty of estimation depth image directly from measurements monotonically decreases when the measurement rate increases. It further becomes negligible at high measurement rate.*

Finally, we modify the data cost $E_{d,u}(l)$ to efficiently handle the measurement quantization noise, where the quantized measurements are represented as $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_J$ respectively. As described in Section 6.5, we first handle the quantization noise in the measurement vector Y_1, and then the result is used to handle the noise in the measurement vectors $Y_2, \ldots, Y_J$. In more details, we first jointly estimate a measurement vector

$\tilde{Y}_1$ and an image $\bar{I}_1$ by solving the optimization problem given in Eq. (6.39). The estimated vector $\tilde{Y}_1$ is then used to jointly estimate a set of measurement vectors $\tilde{Y}_2, \ldots, \tilde{Y}_J$ as a solution to the following optimization problem:

$$(\tilde{Y}_2, \ldots, \tilde{Y}_J) = \min \sum_{j=2}^{J} \sum_{m=1}^{N_1} \|Y_{j,m}^T - \phi_j^m A_j^m \Phi_1^T \tilde{Y}_1\|_2 \quad \text{s.t.} \quad \tilde{Y}_{2,m} \in \mathcal{R}_{\hat{Y}_2}, \ldots, \tilde{Y}_{J,m} \in \mathcal{R}_{\hat{Y}_J}. \tag{6.52}$$

Using the estimated measurement vectors $\tilde{Y}_1, \tilde{Y}_2, \ldots, \tilde{Y}_J$, the robust data term is given as

$$\hat{E}_{d,u}(l) = \sum_{j=2}^{J} \sum_{m=1}^{N_1} \|\tilde{Y}_{j,m}^T - \phi_j^m A_j^m \Phi_1^T \tilde{Y}_1\|_2^2. \tag{6.53}$$

We then replace the data term $E_{d,u}(l)$ with the robust data term $\hat{E}_{d,u}(l)$ in the energy model given in Eq. (6.44) in order to estimate a robust depth solution from quantized measurements. We show later the experimental results that highlight the benefit of robust data term $\hat{E}_{d,u}(l)$.

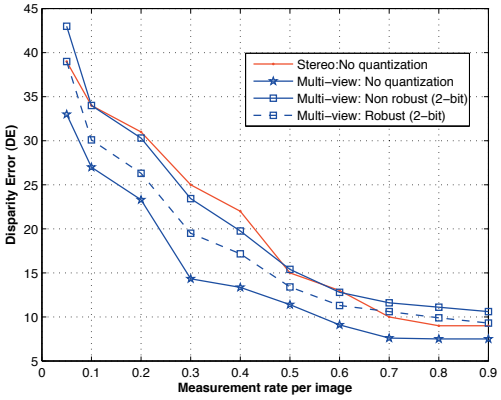
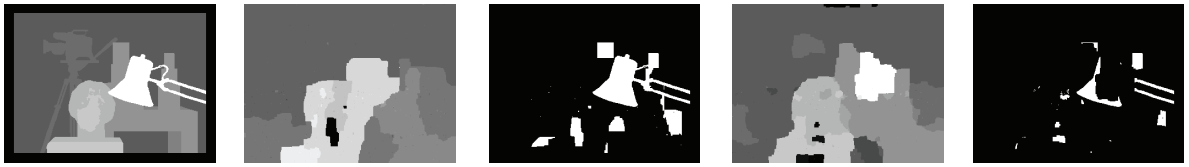


Figure 6.21: Comparison of the inverse depth error estimated in the multi-view and stereo scenarios for the Tsukuba dataset. Five views (center, left, right, bottom and top views) are used in the multi-view experiments. The effect of 2-bit measurements quantization on the performance is also illustrated in the multi-view scenario.

We finally evaluate the performance of our multi-view scheme using five images from the Tsukuba dataset (center, left, right, bottom and top views) [150]. For a given measurement rate, we estimate a depth image by solving the optimization given in Eq. (6.44) using a modified Graph Cut algorithm. The resulting performance is given in Fig. 6.21, where the x-axis represents the measurement rate per image and y-axis represents the corresponding depth error. As expected, from Fig. 6.21 we observe that the quality of the depth image is improved when the measurement rate is increased. In particular, we observe that the disparity error is approx 7.5% at a measurement rate of 0.7, and that the accuracy of the disparity image is not further improved for rates larger than 0.7. This is because the proposed scheme reaches a saturation point at high rates; this is consistent with our earlier observations in the stereo imaging frameworks. We then compare our results to a scheme that estimates a disparity image in the stereo imaging framework. The performance comparison between the multi-view and stereo scenarios is given in Fig. 6.21. It is clear from the plot that the depth error is small for a given measurement rate when all the views are available. For visual comparisons, we show in Fig. 6.22(b) and Fig. 6.22(d) the depth map obtained in stereo and multi-view scenarios, respectively for a measurement rate of 0.4 per view. The corresponding errors with respect to the groundtruth are shown in Fig. 6.22(c) and Fig. 6.22(e) respectively, where the white pixels



(a) $\mathbf{M}^h$ (b) $\mathbf{m}^h$ (c) $|\mathbf{M}^h - \mathbf{m}^h| > 1$ (d) $\mathbf{m}^h$ (e)$|\mathbf{M}^h - \mathbf{m}^h| > 1$

Figure 6.22: Comparison of the depth image estimated in the multi-view and stereo scenarios for the Tsukuba dataset: (a) Groundtruth disparity map; (b) estimated depth map in the stereo setup; (c) absolute error in the inverse depth map with DE=22%; (d) estimated depth map in the multi-view setup; (e) absolute error in the inverse depth map with DE=13.6%. Pixels with absolute error greater than one is marked as white. The depth image is estimated at a measurement rate of 0.4 per image.

denote errors. We see that the error is smaller in the multi-view scenario. In particular, the depth values in the lamp region are correctly estimated in the multi-view scenario when compared to the stereo scenario. Finally, it should be noted that the total number of measurements used in the multi-view scenario is higher than the one used in the stereo imaging case. The proposed multi-view scheme however gives a better depth image when more information is available.

Lastly, we study the performance of our multi-view scheme in the quantized scenario, where the measurements are uniformly quantized using a 2-bit quantizer. We first estimate a depth image directly from the quantized measurements using the data cost $E_{d,u}$ given in Eq. (6.50). The resulting performance is denoted as *Non robust* in Fig. 6.21. As expected, for a given measurement rate the performance is degraded in the quantized scenario when compared to the unquantized scenario. We then estimate a depth map using the robust data term given in Eq. (6.53) and the resulting performance is denoted as *Robust* in Fig. 6.21. We see that similarly to our earlier observations in the stereo scenarios, the proposed robust data term improves the performance when compared to the non-robust scheme that directly uses the quantized measurements for depth estimation.

6.7 Conclusions

In this chapter we have contributed a novel and robust framework for estimating a correlation model between multiple images directly from compressed linear measurements without any explicit image reconstruction steps. We have shown that the underlying correlation structure that relates the images remains linear in both image and compressed domains, when the sensing matrix is linear. Based on this observation, we have derived a new regularized cost function to estimate the motion field directly in the compressed domain. A robust correlation information is estimated by solving a regularized energy minimization problem that enforces consistency with the measurements and smoothness of the correlation estimate. We have theoretically proved that the accuracy of the correlation estimation improves monotonically with increasing measurement rates; this has been later verified with experiments too. Then, we have demonstrated experimentally that our novel framework is very promising in handling the measurement noise due to quantization. Finally, we have shown that the performance of our low complex correlation estimation scheme is competitive to the DFR-sparsity scheme; this is certainly an interesting and promising result especially in distributed analysis tasks that do not target reconstruction.

In the next chapter, we propose a novel joint reconstruction algorithm that uses our correlation estimation for decoding the correlated images. We will also see that our accurate correlation solution helps the joint reconstruction algorithm to improve the decoding quality of the images when compared to the independent reconstruction of images.

Chapter 7

Joint Reconstruction from Compressed Linear Measurements

7.1 Introduction

In Chapters 5 and 6, we have proposed correlation estimation algorithms in a framework where multiple simple CS sensors transmit compressed images in the form of quantized linear measurements. In this chapter, we build on these frameworks and propose efficient joint reconstruction solutions that combine the correlation information and the quantized measurements to decode the images.

In the first part of this chapter, we build on the asymmetric correlation estimation algorithm described in Chapter 5 and propose a joint reconstruction algorithm from quantized linear measurements. As described in Chapter 5, we first compute the most prominent visual features in the reference image and approximate them with geometric functions drawn from a parametric dictionary. Then, a correlation model is constructed by solving a regularized optimization problem that estimates the corresponding features in the compressed images along with the relative geometric transformations. When the correlation model is used for compressed image estimation based on warping, we have observed that the visual information is not accurately represented along the edges and high frequency components. In this chapter, we propose a reconstruction algorithm that captures the missing details and texture information in the predicted image from the information provided by quantized measurements. The joint reconstruction is based on a constrained optimization framework that reconstructs a smooth image which is as close as possible to the predicted image (that is obtained in Chapter 5). At the same time, we use additional constraints to enforce the reconstructed image to be consistent with the quantized measurements, where the consistency is measured using an l_p norm to effectively account the quantization non-linearities [48]. We solve this joint reconstruction problem using parallel proximal algorithms [136]. Our experimental results confirm that the proposed reconstruction scheme improves the quality of the predicted image and also improves the RD performance of the proposed distributed coding solution.

In the second part of this chapter, we propose a novel joint reconstruction algorithm that takes benefit of our correlation estimation (described in Chapter 6) for decoding the correlated images in symmetric sensing framework. The joint reconstruction is cast as an optimization problem where the reconstructed images have to satisfy sparsity priors, as well as consistency with the corresponding quantized measurements and the correlation information. We solve this joint reconstruction problem by effective proximal methods and show that accurate correlation estimation in distributed image representation permits to outperform independent decoding solutions in terms of image quality.

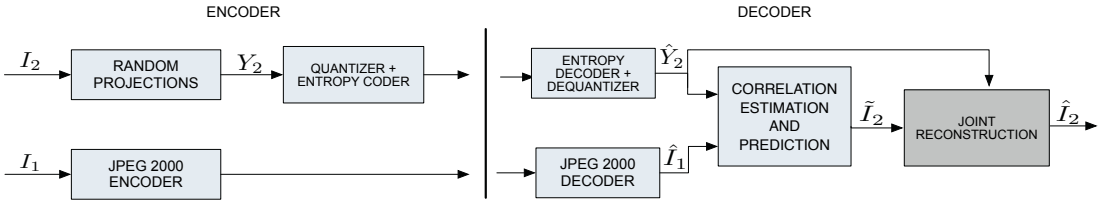


Figure 7.1: Schematic representation of the proposed asymmetric coding scheme. The images I_1 and I_2 are correlated through displacement of scene objects due to viewpoint change or motion of scene objects. The correlation between images is first estimated and then used to generate a side information $\tilde{I}_2$, as described in Chapter 5. The quality of the side information $\tilde{I}_2$ is improved using an additional joint reconstruction stage that decodes an image $\hat{I}_2$ from the quantized linear measurements $\hat{Y}_2$.

7.2 Asymmetric framework

We consider the framework shown in Fig. 7.1, where a pair of images I_1 and I_2 (with resolution N) represent a scene at different time instants or from different viewpoints; these images are correlated through the motion of visual objects. The correlated images I_1 and I_2 are independently encoded and the compressed visual information is jointly processed at a central decoder. In particular, we consider the scenario where the image I_1 is selected as the reference and it is used as a side information for decoding the second image I_2. The first image I_1 is encoded using standard coding solutions based on JPEG 2000, and the second image I_2 is represented in terms of small number of linear projections Y_2. At the joint decoder, we estimate the underlying correlation model between I_1 and I_2 from the compressed reference image $\hat{I}_1$ and quantized linear measurements $\hat{Y}_2$, using the algorithm described in Chapter 5. Recall that the correlation model is estimated from the linear measurements by solving the OPT-1 or OPT-2 optimization problems described in Sections 5.3 and 5.5.

We have shown experimentally that the quality of $\tilde{I}_2$ saturates at high rates as the high frequency components are not efficiently captured. We describe now a novel reconstruction algorithm based on an optimization framework that estimates the high frequency components from the measurements information (see Fig. 7.1). The key idea in the proposed reconstruction algorithm is to consider the predicted image $\tilde{I}_2$ obtained in Chapter 5 as a side information and combine it with the information from the measurements for the reconstruction of the improved image $\hat{I}_2$. We propose to reconstruct an image $\hat{I}_2$ that is not only consistent with the measurements Y_2, but also close to the image $\tilde{I}_2$ that is obtained by warping the reference image. By merging these two constraints the proposed optimization problem is given as

$$\hat{I}_2 = arg \min_{I_2} \|I_2\|_{TV} \;\; \text{s.t.} \;\; \|Y_2 - \Phi I_2\|_2 = 0, \|I_2 - \tilde{I}_2\|_2 \leq \epsilon_2. \tag{7.1}$$

In the above optimization, we use the prior based on total variation (TV) norm defined in Eq. (2.6) that works well for natural images [16]. Nevertheless, one could also use a sparsity prior that assumes sparse representation of the reconstructed image in a particular basis Ψ.

When the measurements are quantized, it is well known that the optimization problem given in Eq. (7.1) fails to meet the quantization consistency, i.e., the reconstructed image $\hat{I}_2$ is not consistent with the quantized measurements. Jacques *et al.* [48] have showed that the quantization consistency can be enforced with an l_p norm with $p > 2$ and not using l_2 norm in the first constraint. Inspired by Jacques *et al.* [48], when the measurements Y_2 are quantized, the above optimization problem can be modified as

$$\hat{I}_2 = arg \min_{I_2} \|I_2\|_{TV} \;\; \text{s.t.} \;\; \|Y_2 - \Phi I_2\|_p \leq \epsilon_1, \|I_2 - \tilde{I}_2\|_2 \leq \epsilon_2. \tag{7.2}$$

Finally, it should be noted the optimization problems given in Eq. (7.1) or Eq. (7.2) can also be used to decode video images in the distributed compressive sampling applications proposed in [101, 102]. In these schemes, the side information $\tilde{I}_2$ used for decoding the Wyner-Ziv image I_2 is constructed using the key frames based on motion compensated prediction (see Section 2.5.3). However, it is well known that the motion compensation prediction fails to accurately estimate the visual information in the texture and edge regions. Therefore, the optimization given in Eq. (7.1) can help in capturing the missing details in the side information and subsequently to reconstruct the Wyner-Ziv frame $\hat{I}_2$. The main advantage is that our joint reconstruction scheme efficiently handles the measurement quantization noise, while it is not efficiently considered in [101, 102].

7.2.1 Optimization methodology

We now describe the optimization methodology based on parallel proximal methods (PPXA) to solve Eq. (7.2). The optimization problem given in Eq. (7.2) can be visualized as

$$\min_{X\in\mathcal{H}} f(X) = \min_{X\in\mathcal{H}} f_1(X) + f_2(X) + f_3(X), \tag{7.3}$$

where $\mathcal{H} = \mathbb{R}^N$ is the Hilbert space. The functions f_1, f_2 and f_3 are given as

$$f_1(X) = \|X\|_{TV}, \tag{7.4}$$
$$f_2(X) = i_{c^p(\epsilon_1)}(X), \text{ where } c^p(\epsilon_1) = \{X \in \mathbb{R}^N |\ \|\hat{Y}_2 - \Phi X\|_p \leq \epsilon_1\}, \tag{7.5}$$
$$f_3(X) = i_{d(\epsilon_2)}(X), \text{ where } d(\epsilon_2) = \{X \in \mathbb{R}^N |\ \|X - \tilde{I}_2\|_2 \leq \epsilon_2\}, \tag{7.6}$$

where i_a represents the indicator function of the closed convex set a. In the rest of this section, we focus on computing the *prox* for the functions f_1, f_2 and f_3 given in Eqs. (7.4)-(7.6). For more details about the PPXA methods we refer the reader to Section 4.2.4. The *prox* for the function $f_1(X) = \|X\|_{TV}$ can be computed using Chambolle's algorithm [137]. The function f_2 can be represented using the combination of a function F and an affine operator G, with $F = i_{D^p(\epsilon_1)}$ and $G = \Phi X - \hat{Y}_2$, i.e., $f_2 = F \circ G$. The set $D^p(\epsilon_1)$ is the l_p ball with radius ϵ_1 that is represented as

$$D^p(\epsilon_1) = \{y \in \mathbb{R}^K |\ \|y\|_p \leq \epsilon_1\}. \tag{7.7}$$

As the measurement matrix Φ is a tight frame, i.e., $\Phi\Phi^* = c\mathbb{1}$, the $prox_{f_2}$ can be computed as

$$prox_{f_2}(X) = prox_{F\circ G}(X) = X + c^{-1}\Phi^*(prox_F - \mathbb{1})(G(X)). \tag{7.8}$$

The $prox_F$ for $p = 2$ can be computed using as

$$prox_F(y) = min(\epsilon_1 \frac{y}{\|y\|_2}, y). \tag{7.9}$$

Unfortunately, for $p > 2$ there is no closed form expression to compute the $prox_F(y)$. In such cases, one has to rely on the iterative scheme proposed in [48]. Finally, the function f_3 can be represented as $f_3 = F \circ G$, where $F = i_{D^2(\epsilon_2)}$ and $G = (\mathbb{1}X - \tilde{I}_2)$. The set $D^2(\epsilon_2)$ represents the l_2 ball with radius ϵ_2 (see Eq. (7.7)). The $prox_{f_3}$ can be computed as

$$prox_{f_3}(X) = prox_{F\circ G}(X) = X + (prox_F - \mathbb{1})(G(X)), \tag{7.10}$$

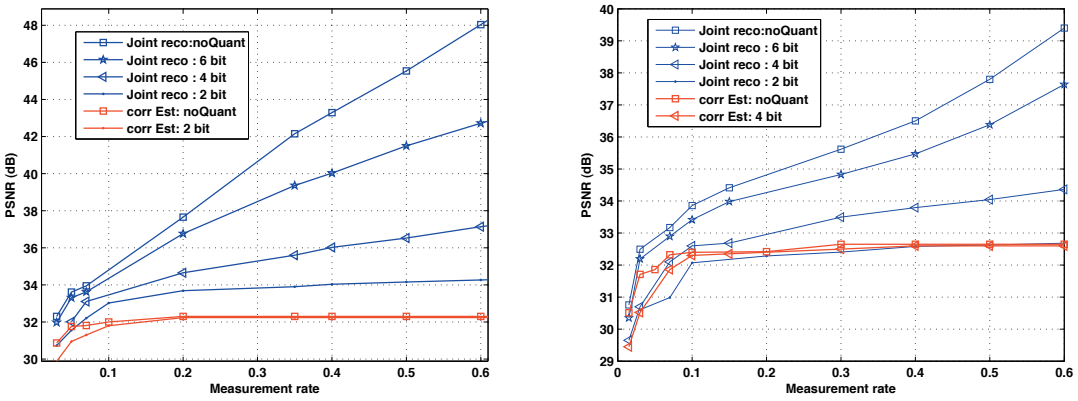
where $prox_F$ for the indicator function $F = i_{D^2(\epsilon_2)}$ can be computed using Eq. (7.9).

7.2.2 Experimental results

The performance of the joint reconstruction has been studied on three natural datasets: *Sawtooth* and *Plastic* stereo images, and *Foreman* video sequence (frames 2 and 3), as considered in Chapter 5. In our experiments, we set the quality of the reference image $\hat{I}_1$ to 33 dB for the Sawtooth and Plastic datasets and to 45 dB for the Foreman dataset. The measurements are generated using a block scrambled Hadamard transform with a block size 8 [30]. For a given measurement rate, we estimate the correlation model with OPT-2 optimization problem using the robust data cost in Eq. (5.8). In particular, we use the local optimization techniques based on constructive search strategy to estimate a robust correlation model with consistent image prediction (see Algo. 5 in Chapter 5). We consider this suboptimal correlation solution for simplicity; nevertheless one could also use graph-based solutions to estimate the correlation model (see Chapter 5 for details). The estimated correlation model is then used to predict the second image $\tilde{I}_2$ based on disparity or motion compensation respectively. We then reconstruct the second image $\hat{I}_2$ from the compressed measurements $\hat{Y}_2$ and the predicted image $\tilde{I}_2$ by solving Eq. (7.2).

Fig. 7.2 shows the benefit of additional reconstruction stage for the Plastic and Foreman datasets, in unquantized scenario. It should be noted that, when the measurements are not quantized we solve the optimization problem in Eq. (7.2) with $p = 2$. The parameter ϵ_1 is set to $1e-4$ and the parameter ϵ_2 is selected based on trial and error experiments such that the quality of reconstructed image $\hat{I}_2$ is maximized. The PSNR curves corresponding to the predicted image $\tilde{I}_2$ and the reconstructed image $\hat{I}_2$ are marked in red and blue respectively. From Fig. 7.2, it is clear that the quality of the predicted image $\tilde{I}_2$ saturates around a measurement rate of 0.2 (see *corr Est: no Quant*). By activating the reconstruction stage the quality of the reconstructed image $\hat{I}_2$ improves as the measurement rate increases; this proves that the reconstruction stage captures the details and texture components (see *Joint reco: no Quant*). Similar experimental findings are observed for the Sawtooth dataset shown in Fig. 7.3.

Now, we analyze the behavior of the proposed joint reconstruction scheme in the quantized scenario. We solve the optimization problem with $p = 8$ (selected based on trial and error experiments), and the parameter ϵ_1 is calculated based on the quantization bit rate and the value of p [48]. Fig. 7.2 and Fig. 7.3 compare the reconstruction quality of $\hat{I}_2$, when the measurements are quantized using 2, 4 and 6 bits. As expected, for a given measurement rate the quality of the reconstructed image $\hat{I}_2$ degrades when the quantization bit



(a) (b)

Figure 7.2: Influence of measurement quantization on the joint reconstruction performance: (a) Plastic dataset; (b) Foreman dataset. The PSNR is computed between I_2 and $\hat{I}_2$ in the joint reconstruction scheme and between I_2 and $\tilde{I}_2$ in the correlation estimation scheme.

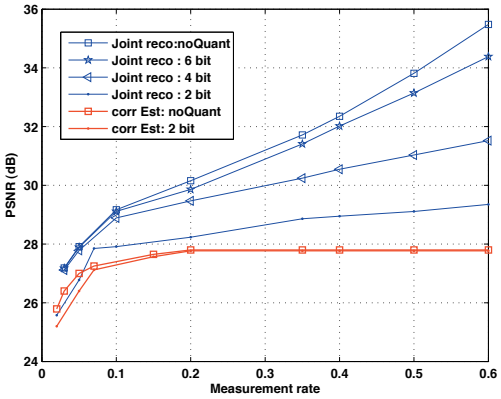
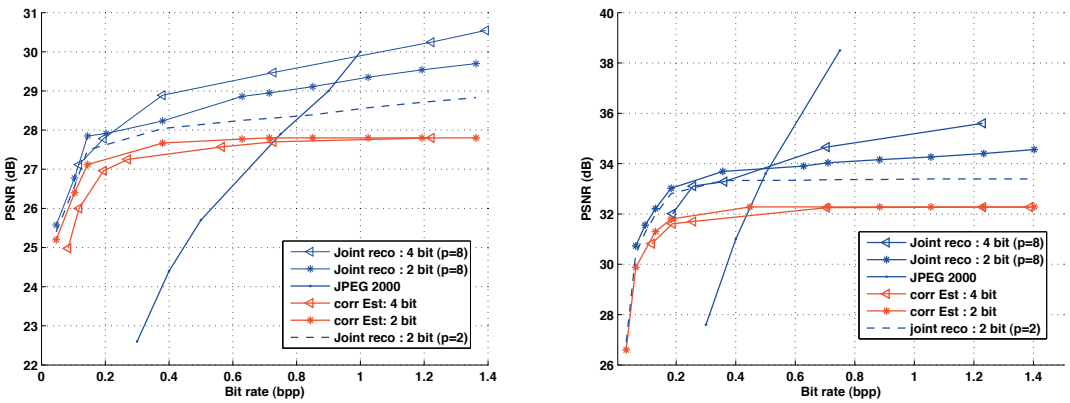


Figure 7.3: Influence of measurement quantization on the joint reconstruction performance in the Sawtooth dataset. The PSNR is computed between I_2 and $\hat{I}_2$ in the joint reconstruction scheme and between I_2 and $\tilde{I}_2$ in the correlation estimation scheme.

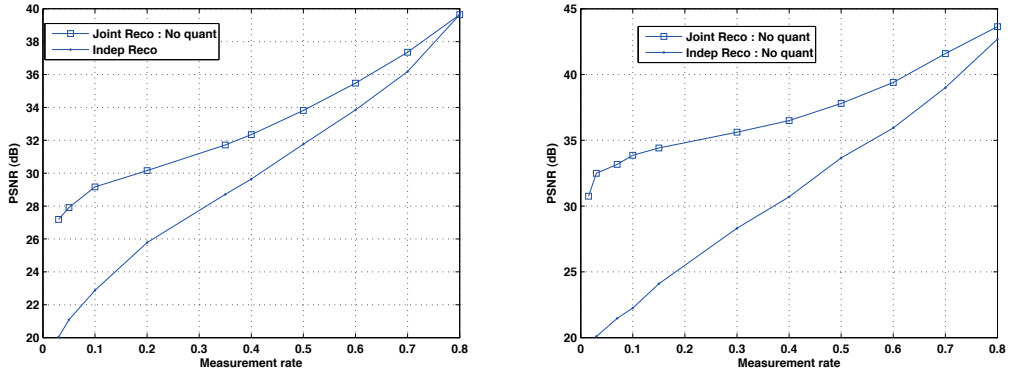
rate decreases.

We finally analyze the coding performance of the proposed scheme, where the bit rate is computed by entropy coding of the quantized measurements using an arithmetic encoder (see Fig. 7.1). Fig. 7.4 shows the coding performance of the proposed scheme when the measurements are quantized using 2 and 4 bits. For a given bit rate, it is clear from the plot that the proposed reconstruction scheme improves the quality of the warped image $\tilde{I}_2$. Also, as expected, the quality of the reconstructed image $\hat{I}_2$ increases with the bit rate and thus corrects the saturation of the predicted image $\tilde{I}_2$ at high rates. In addition, our solution shows significant improvement over independent coding solutions based on JPEG 2000 between low to medium bit rates. The choice of using $p > 2$ is proven effective in Fig. 7.4, when the measurements are quantized using a 2-bit quantizer. It is clear that the coding performance is improved with $p = 8$ instead of $p = 2$, as the former choice effectively handles the quantization noise while reconstructing the image $\hat{I}_2$.



(a) (b)

Figure 7.4: Influence of 2- and 4-bit rate quantizers on the joint reconstruction performance: (a) Sawtooth dataset; (b) Plastic dataset. The PSNR is computed between I_2 and $\hat{I}_2$ in the joint reconstruction scheme and between I_2 and $\tilde{I}_2$ in the correlation estimation scheme.



(a) (b)

Figure 7.5: Benefit of using side information on the joint reconstruction performance: (a) Sawtooth dataset; (b) Plastic dataset. The independent coding performance is obtained by solving Eq. (7.1) without activating the last constraint $\|I_2 - \tilde{I}_2\|_2 \leq \epsilon_2$.

At last, we highlight the benefit of using side information (i.e., the benefit of exploiting the correlation for the joint reconstruction) in Fig. 7.5. It is clear that the proposed joint reconstruction scheme significantly outperforms the independent reconstruction scheme at low rates; this is due to the efficient correlation estimation. When the rate increases, we notice that the impact of side information on the reconstruction quality of image $\hat{I}_2$ decreases. In other words, the benefit of exploiting the correlation between images is less beneficial at high rates.

7.3 Symmetric framework

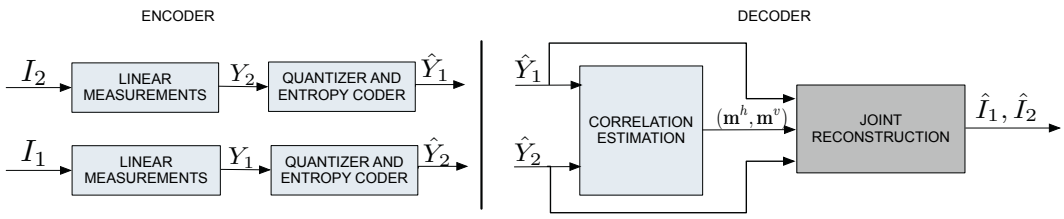


Figure 7.6: Schematic representation of the proposed symmetric coding scheme. The images I_1 and I_2 are correlated through displacement of scene objects due to viewpoint change or motion of scene objects. The correlation model is estimated directly in the compressed domain without any intermediate image reconstruction, as described in Chapter 6. The estimated correlation information $(\mathbf{m}^h, \mathbf{m}^v)$ is used for joint reconstruction to decode the images $\hat{I}_1$ and $\hat{I}_2$.

We now describe our symmetric distributed joint reconstruction framework for reconstructing multiple correlated images. The scenario is illustrated in Fig. 7.6 for two correlated images I_1 and I_2 and we describe later the extension to more than two images. As shown in Fig. 7.6, the images I_1 and I_2 are directly acquired in the form of linear measurements, respectively given as Y_1 and Y_2. The computed linear measurements are quantized and entropy coded. A central decoder jointly processes the compressed linear measurements $\hat{Y}_1$ and $\hat{Y}_2$, and decodes a pair of images $\hat{I}_1$ and $\hat{I}_2$.

In this setting, we propose to estimate the underlying correlation model between images directly in the compressed domain as described in Chapter 6. In this chapter, we propose a novel joint reconstruction algorithm that takes benefit of our estimated correlation information $(\mathbf{m}^h, \mathbf{m}^v)$ to jointly reconstruct the images as shown in Fig. 7.6. We propose to reconstruct a pair of images $\hat{I}_1$ and $\hat{I}_2$ by enforcing consistency with the compressed information and with the estimated correlation model. A pair of images $\hat{I}_1$ and $\hat{I}_2$ is reconstructed as a solution to the following constrained optimization problem:

$$(\hat{I}_1, \hat{I}_2) = arg \min_{(I_1, I_2)} (\|\Psi^* I_1\|_1 + \|\Psi^* I_2\|_1) \text{ s.t. } \|Y_1 - \Phi_1 I_1\|_2 = 0, \|Y_2 - \Phi_2 I_2\|_2 = 0, \|I_2 - A I_1\|_2^2 \leq \epsilon_2, \tag{7.11}$$

where Ψ represents a redundant dictionary or an orthonormal basis in which the image is assumed to be sparse and Ψ^* represents the conjugate transpose of Ψ. In this optimization framework we use the sparse prior for simplicity, however one could also use a TV prior as described in Section 7.2. From Eq. (7.11) it is clear that the images can be reconstructed independently if we solve the optimization without the last constraint $\|I_2 - A I_1\|_2^2 \leq \epsilon$; this corresponds to reconstructing the sparse images in Ψ that best agree with the given measurement information. By adding the last constraint we impose that the pair of images should also satisfy the correlation model in addition to the sparsity and data fidelity constraints.

When the measurements are quantized, we compute the measurement consistency in the first two constraints using the l_p distance, as described in Section 7.2. The optimization problem in Eq. (7.11) can be rewritten as

$$(\hat{I}_1, \hat{I}_2) = arg \min_{(I_1, I_2)} (\|\Psi^* I_1\|_1 + \|\Psi^* I_2\|_1) \text{ s.t. } \|\hat{Y}_1 - \Phi_1 I_1\|_p \leq \epsilon_1, \|\hat{Y}_2 - \Phi_2 I_2\|_p \leq \epsilon_1, \|I_2 - A I_1\|_2^2 \leq \epsilon_2, \tag{7.12}$$

where the measurement consistency is measured using the l_p norm with $p > 2$. It is easy to check that all the functions in Eq. (7.11) and Eq. (7.12), i.e., $\|\Psi^* I_j\|_1, \|\hat{Y}_j - \Phi_j I_j\|_p, \forall j \in \{1, 2\}$ and $\|I_2 - A I_1\|_2^2$ are convex. For more details related to the convexity of the last constraint $\|I_2 - A I_1\|_2^2$, we refer the reader to Proposition 1 in Chapter 4. Therefore the optimization problems in Eq. (7.11) and Eq. (7.12) are convex that allows to effectively solve using proximal solutions.

7.3.1 Optimization methodology

In this section, we describe the methodology to solve the optimization problem in Eq. (7.12) using parallel proximal splitting (PPXA) methods. For mathematical convenience we rewrite problem in Eq. (7.12) as

$$\min_X (\|\Psi^* E_1 X\|_1 + \|\Psi^* E_2 X\|_1) \text{ s.t. } \|E_1(Y - \Phi X)\|_2 \leq \epsilon_1, \|E_2(Y - \Phi X)\|_2 \leq \epsilon_1, \|[-A \quad \mathbb{1}] X\|_2^2 \leq \epsilon_2, \tag{7.13}$$

where $X = \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}$, $Y = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \end{bmatrix}$, $E_1 = \begin{bmatrix} \mathbb{1} & 0 \end{bmatrix}$, $E_2 = \begin{bmatrix} 0 & \mathbb{1} \end{bmatrix}$, and $\Phi = \begin{bmatrix} \Phi_1 & 0 \\ 0 & \Phi_2 \end{bmatrix}$.
The optimization problem in Eq. (7.13) can be visualized as

$$\min_{X \in \mathcal{H}} f(X) = \min_{X \in \mathcal{H}} f_1(X) + f_2(X) + f_3(X) + f_4(X) + f_5(X), \tag{7.14}$$

where $\mathcal{H} = \mathbb{R}^{2N}$ is the Hilbert space. The functions f_1, f_2, f_3, f_4 and f_5 are given as

$$f_1(X) = \|\Psi^* E_1 X\|_1, \tag{7.15}$$
$$f_2(X) = \|\Psi^* E_2 X\|_1, \tag{7.16}$$
$$f_3(X) = i_{c_1^p(\epsilon_1)}(X), \text{where } c_1^p(\epsilon_1) = \{X \in \mathbb{R}^{2N} \mid \|E_1(Y - \Phi X)\|_p \leq \epsilon_1\}, \tag{7.17}$$
$$f_4(X) = i_{c_2^p(\epsilon_1)}(X), \text{where } c_2^p(\epsilon_1) = \{X \in \mathbb{R}^{2N} \mid \|E_2(Y - \Phi X)\|_p \leq \epsilon_1\}, \tag{7.18}$$
$$f_5(X) = i_{d(\epsilon)}(X), \text{where } d(\epsilon) = \{X \in \mathbb{R}^{2N} \mid \|[-A \;\; \mathbb{1}]X\|_2^2 \leq \epsilon_2\}. \tag{7.19}$$

In the rest of this section, we focus on the methodology to compute the *prox* for functions $f_i, \forall i \in \{1, 2, 3, 4, 5\}$, while for more details about the PPXA algorithm we refer the reader to Section 4.2.4. The function $f_1(X) = \|\Psi^* E_1 X\|_1$ can be represented by the combination of functions F and G, where $F = \|.\|_1$ and $G = \Psi^* E_1 X = \Omega X$, i.e., $f_1 = F \circ G$. If Ψ is an orthonormal basis (i.e., $\Omega\Omega^* = \mathbb{1}$) the $prox_{f_1}$ can be computed using the following closed form expression

$$prox_{f_1}(X) = prox_{F \circ G}(X) = X + \Omega^*(prox_{\gamma F} - \mathbb{1})(G(X)), \tag{7.20}$$

[136, 138]. The $prox_{\gamma F}$ of $F = \|.\|_1$ can be computed using component-wise soft thresholding with threshold γ [138]. On the other hand, if Ψ is a general frame or redundant dictionary (i.e., $c_1\mathbb{1} \leq \Psi\Psi^T \leq c_2\mathbb{1}$) the *prox* can be computed using Eqs. (4.20)-(4.21). The $prox_{f_2}$ can be computed using Eq. (7.20) or Eqs. (4.20)-(4.21) by setting $F = \|.\|_1$ and $G = \Psi^* E_2 X = \Omega X$, depending on Ψ is an orthonormal basis or a redundant dictionary (general frame) respectively.

The function f_3 can be represented using the combination of a function F and an affine operator G where $F = i_{D^p(\epsilon_1)}$ and $G = E_1\Phi X - E_1 Y$. The set $D^p(\epsilon_1)$ is the l_p ball with radius ϵ_1 as defined in Eq. (7.7). As the operator Φ is a tight frame with frame constant c, the $prox_{f_3}$ can be computed as

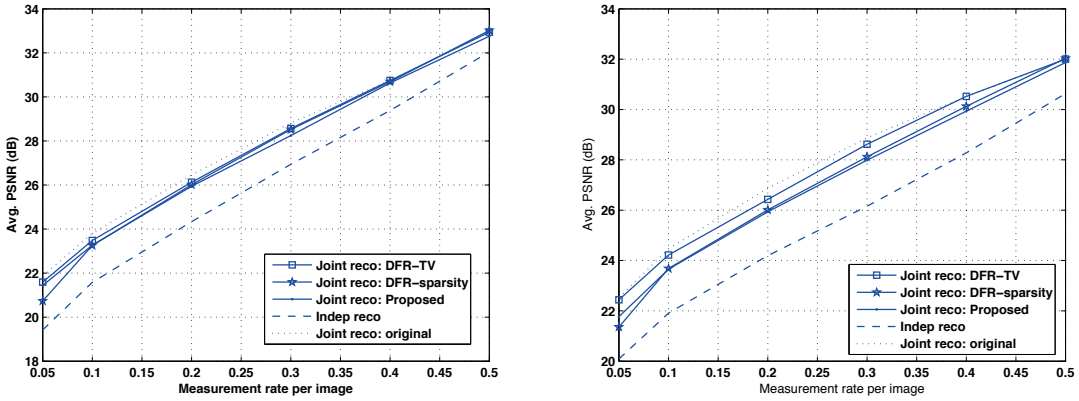
$$prox_{f_3}(X) = prox_{F \circ G}(X) = X + c^{-1}(E_1\Phi)^*(prox_F - \mathbb{1})(G(X)), \tag{7.21}$$

where $prox_F$ of $F = i_{D^p(\epsilon_1)}$ can be computed using Eq. (7.9) for $p = 2$. For $p > 2$, we use the iterative scheme proposed in [48] to compute the $prox_F$. Following the same analogy, the *prox* for the function f_4 can be solved using Eq. (7.21) by setting $F = i_{D^p(\epsilon_1)}$ and $G = E_2\Phi X - E_2 Y$. Finally, the *prox* computation of function f_5 can be computed using Eqs. (4.20)-(4.21).

7.3.2 Experimental results

In this section, we report the experimental behavior of our novel joint reconstruction algorithm of Eq. (7.13) in the correlated stereo and motion datasets that are used in Chapter 6. In particular, we perform joint reconstruction experiments using the correlation model estimated with different sets of measurement matrices, i.e., $\Phi_1 \neq \Phi_2$. We assume that the images are sparse in an orthonormal basis built on a wavelet representation. Hence, the $prox_{f_1}$ and $prox_{f_2}$ operators are computed using Eq. (7.20). The parameter ϵ_2 is selected based on a trial and error procedure that maximizes the reconstruction image quality of $\hat{I}_1$ and $\hat{I}_2$. In the rest of this section, we highlight the benefit of our joint reconstruction algorithm that effectively considers the underlying correlation model while jointly decoding a pair of images. We then study the importance of accurate correlation estimation and the impact of the measurement quantization noise in the joint reconstruction performance. Finally, we extend our proposed joint reconstruction scheme to decode more than two images.

We first analyze the performance of our joint reconstruction scheme using the correlation model estimated from the unquantized linear measurements for the stereo datasets (Tsukuba and Venus). We found experimentally that $\epsilon_2 = 14$ works well on Tsukuba and Venus datasets. We first compare our result with respect to an independent reconstruction scheme that decodes the images without exploiting the correlation.



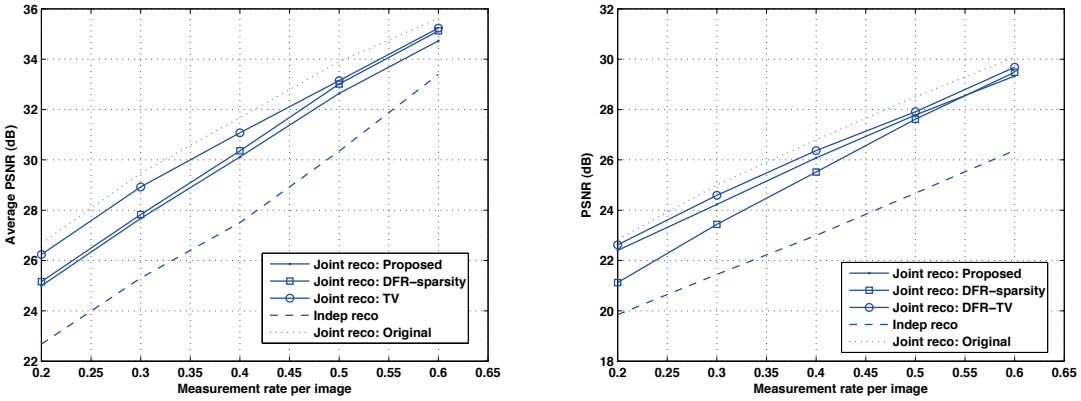
(a) (b)

Figure 7.7: Influence of the disparity accuracy on the joint reconstruction performance: (a) Tsukuba dataset; (b) Venus dataset. The joint reconstruction performance is also compared to the independent reconstruction scheme.

This corresponds to solving the optimization problem of Eq. (7.13) without activating the last constraint $\|I_2 - AI_1\|_2^2 \leq \epsilon_2$. Fig. 7.7 compares the average reconstruction quality between the joint (denoted as *Joint reco: Proposed*) and independent reconstruction schemes (denoted as *Indep reco*) for Tsukuba and Venus datasets. As expected, the joint reconstruction improves the reconstruction quality by 2 dB at low measurement rate and by approximately 1 dB at high rates when compared to the independent reconstruction scheme. We also observe in our experiments that the PSNR quality of both reconstructed images $\hat{I}_1$ and $\hat{I}_2$ are similar, i.e., the quality is balanced between the images. The disparity estimation thus proves to be useful in improving the image quality while reconstructing the images.

We then jointly reconstruct the images using the disparity estimated with the DFR-sparsity and DFR-TV schemes. Recall that the DFR scheme estimates a correlation model from the reconstructed reference images (see Section 6.4.1 for details). Fig. 7.7 compares the quality of reconstructed images in the proposed and DFR schemes. We see that the joint reconstruction performance of the proposed scheme is competitive with the joint reconstruction based on DFR-sparsity scheme in terms of image reconstruction quality. This is particularly obvious at a low rate 0.05 where the DFR-sparsity scheme fails to accurately estimate the disparity (see Section 6.4.1). However, the quality of the reconstructed images are marginally penalized (i.e., 0.2 dB and 0.4 dB for the Tsukuba and Venus datasets respectively) compared to the reconstruction achieved in the DFR-TV scheme. Finally, we carry out the same experiments in a scenario where the images are jointly reconstructed using a correlation model that is estimated from the original images. This scheme serves as a benchmark for the joint reconstruction, since the correlation is accurately known at the decoder. The corresponding results are denoted as *Joint reco:original* in Fig. 7.7. We see that the reconstruction quality achieved with the correlation estimated from compressed measurements converges to the upper-bound performance when the measurement rate increases; this further confirms the quality of the disparity estimation algorithm described in Chapter 6.

We now carry out similar experiments on the Yosemite, Mequon and Grove video datasets. The experiments are carried out with $\epsilon_2 = 3$ (selected based on trial and error procedure). Fig. 7.8(a) and Fig. 7.8(b) compare the performance of the proposed joint reconstruction scheme with respect to independent reconstruction for the Mequon and Grove datasets respectively. As expected, by exploiting the correlation between images we improve the reconstruction quality by 3-4 dB at low to medium measurement rates and by approximately 2-3 dB at high rates. We then compare our results with the joint reconstruction scheme based on DFR schemes. For the Mequon dataset (see Fig. 7.8(a)), we observe that the quality penalization is



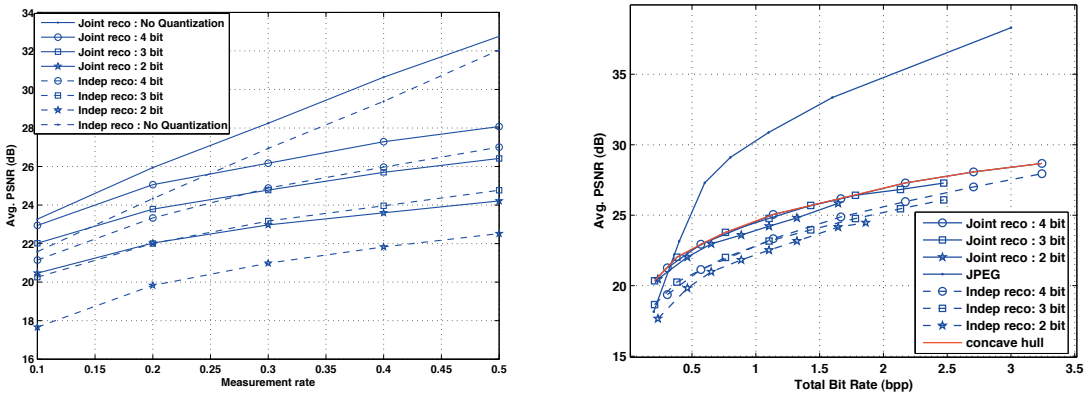
(a) (b)

Figure 7.8: Influence of the motion accuracy on the joint reconstruction performance: (a) Mequon dataset; (b) Grove dataset. The joint reconstruction performance is also compared to the independent reconstruction scheme.

close to 1 dB at low rates and 0.5 dB at high rates when compared to DFR-TV scheme. This is due to the fact that the DFR-TV scheme estimates an accurate correlation model due to good reconstruction quality of the images, as described in Section 6.4.2. However, the proposed scheme performs competitively to the joint reconstruction scheme based on DFR-sparsity. For the Grove dataset (see Fig. 7.8(b)), the proposed scheme competes with the image reconstruction quality of the DFR-TV scheme due to accurate correlation estimation. Finally, we compare our results to a joint reconstruction scheme that uses the true motion field estimated from the original images. As expected, the proposed scheme is far from the upper-bound by a margin of 0.8 dB for the Grove dataset (similar performance for Yosemite dataset) and of 1.5 dB for the Mequon dataset.

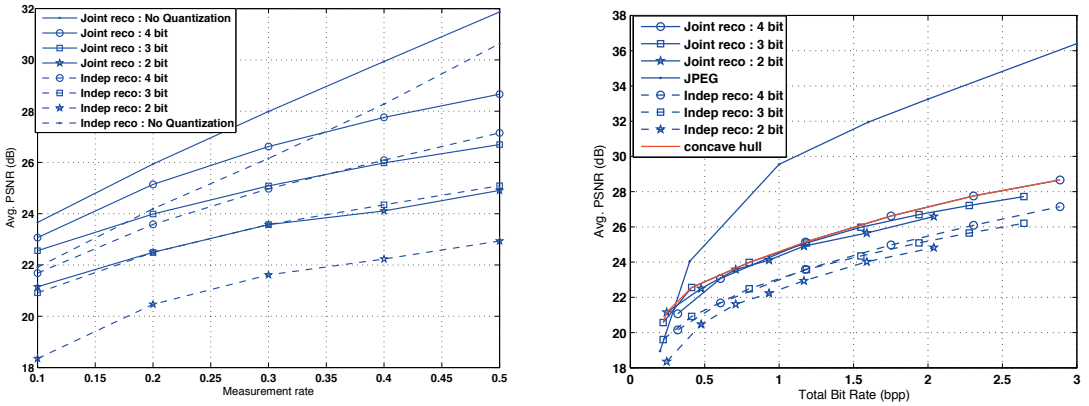
We then analyze the effect of the measurement quantization on the joint reconstruction performance. For the sake of simplicity, we study this influence using uniform quantizers. In the quantized scenarios, we estimate a correlation model using the robust data cost function given in Eq. (6.43). We then use the estimated robust correlation model to solve the joint reconstruction problem in Eq. (7.13) with $p = 8$. Figs. 7.9(a)-7.12(a) show the resulting joint reconstruction performance (shown in *dashed* lines) for the Tsukuba, Venus, Grove and Mequon datasets respectively, when the measurements are quantized using $2, 3$ and 4 bits. We first observe from Figs. 7.9(a)-7.12(a) that the quality of the decoded images degrades in the quantized scenario compared to the unquantized scenario (marked as *Joint reco:No Quantization*). Especially, the degradation is significant when the measurements are quantized coarsely using a 2-bit quantizer. We now compare the quality difference between the joint and independent scheme for a given quantization bit rate. For 3- and 4-bit rate quantizers we observe that the gain with respect to the corresponding independent scheme remains approximately the same as in unquantized scenario, due to the robust correlation estimation from quantized measurements. This shows that the measurement quantization noise is efficiently handled in our distributed representation scheme. However, when the measurements are coarsely quantized using 2 bits the gain with respect to the independent scheme reduces (by 0.5-1 dB) when compared to the unquantized scenario.

We now analyze the RD performance of our symmetric scheme when the bit rate is computed by entropy coding of the quantized measurements. Figs. 7.9(b)- 7.12(b) show the resulting joint reconstruction performances (shown in *dashed line*) when the measurements are quantized using 2, 3 and 4 bits. From the plots we observe that, at low bit rates the 2-bit quantizer performs better than 3- and 4-bit rate quantizers in terms of reconstruction quality. Though the quality of the reconstruction image degrades with 2-bit quantizers as



(a) (b)

Figure 7.9: Influence of the 2-, 3- and 4-bit rate quantizers on the quality of the reconstructed images $\hat{I}_1$ and $\hat{I}_2$ in the Tsukuba dataset. The quality evolution is given for (a) measurement rate and (b) bit rate.

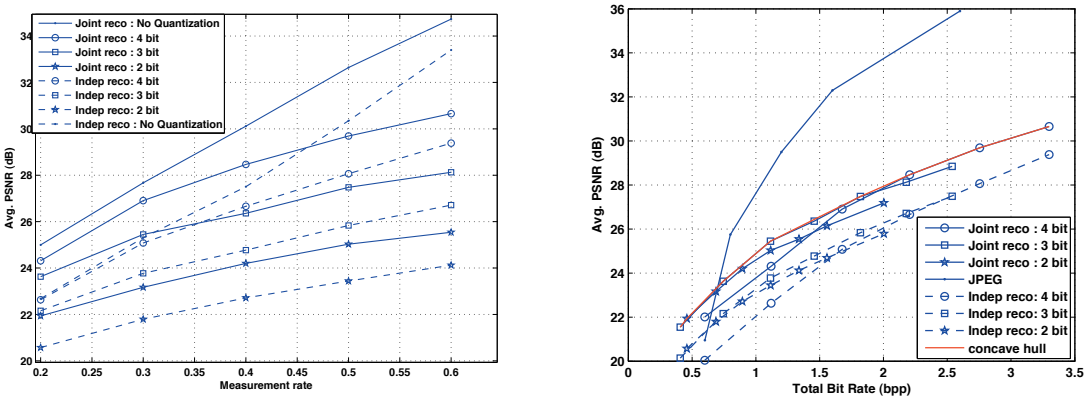


(a) (b)

Figure 7.10: Influence of the 2-, 3- and 4-bit rate quantizers on the quality of the reconstructed images $\hat{I}_1$ and $\hat{I}_2$ in the Venus dataset. The quality evolution is given for (a) measurement rate and (b) bit rate.

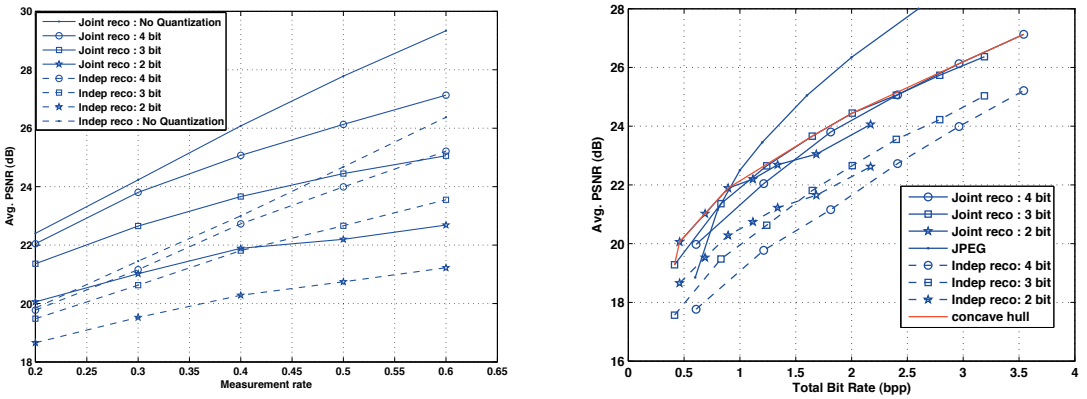
shown before, it is largely compensated by the reduction in bit rate in the rate-distortion performance. This means that the proposed distributed representation scheme is relatively robust to quantization, so that it is possible to attain good rate-distortion performance by drastic quantization of the measurements. At high bit rates, we see that the 4-bit quantizer performs better than the 2- and 3-bit rate quantizers. When we take the convex hull of the RD performances (which corresponds to implementing a proper rate allocation strategy), we observe that the proposed joint reconstruction scheme outperforms independent coding solutions based on JPEG at low rates. However, at high bit rates the proposed scheme performs significantly worse than the JPEG coding solutions due to the high entropy of the linear measurements. Several researchers have noted it already that CS is actually not good for coding or compression applications [44, 45, 46].

We finally extend our proposed symmetric joint reconstruction scheme to more than two images. Es-



(a) (b)

Figure 7.11: Influence of the 2-, 3- and 4-bit rate quantizers on the quality of the reconstructed images $\hat{I}_1$ and $\hat{I}_2$ in the Mequon dataset. The quality evolution is given for (a) measurement rate and (b) bit rate.



(a) (b)

Figure 7.12: Influence of the 2-, 3- and 4-bit rate quantizers on the quality of the reconstructed images $\hat{I}_1$ and $\hat{I}_2$ in the Grove dataset. The quality evolution is given for (a) measurement rate and (b) bit rate.

pecially, we describe here the extension to multi-view imaging scenarios, however the scheme can also be extended to the video imaging scenarios. Given the measurements $Y_j, \forall j = \{1, 2, \ldots, J\}$ and the correlation information among J images (in the form $A_i, \forall j = \{2, \ldots, J\}$, as described in Section 6.6), we propose to jointly reconstruct the J multi-view images as a solution to the following optimization problem:

$$\min_{I_1,\ldots,I_J} \left(\sum_{i=1}^{J} \|\Psi^* I_i\|_1 \right) \quad \text{s.t.} \quad \sum_{i=1}^{J} \|Y_i - \Phi_i I_i\|_2 = 0, \; \sum_{i=2}^{J} \|I_i - A_i I_1\|_2^2 \leq \epsilon. \tag{7.22}$$

From the above equation, we see that the proposed reconstruction algorithm estimates J sparse images that are consistent with both the measurement and correlation informations. Furthermore, it can be noted

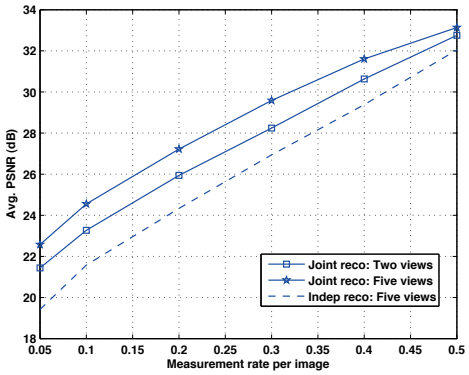


Figure 7.13: Joint reconstruction performance comparison between the multi-view and stereo imaging scenarios in the Tsukuba multi-view dataset. Five views (center, left, right, bottom and top views) are jointly reconstructed in the multi-view scenario.

that the above optimization problem is an extension to the one described in Eq. (7.11), except that the sparsity constraints and the measurement and correlation consistency objectives are applied more than images. The optimization problem in Eq. (7.22) can be rewritten as

$$\min_{X} \left(\sum_{i=1}^{J} \|\Psi^* E_i X\|_1 \right) \text{ s.t. } \quad \|E_1(Y - \Phi X)\|_2 = 0, \|E_2(Y - \Phi X)\|_2 = 0, \ldots, \|E_J(Y - \Phi X)\|_2 = 0, \\ \|DX\|_2^2 \leq \epsilon, \tag{7.23}$$

where $X = \begin{bmatrix} I_1 & I_2 & \cdots & I_J \end{bmatrix}^T$, $Y = \begin{bmatrix} Y_1 & Y_2 & \cdots & Y_J \end{bmatrix}^T$, $E_1 = \begin{bmatrix} \mathbb{1} & 0 & \cdots & 0 \end{bmatrix}$, $E_J = \begin{bmatrix} 0 & 0 & \cdots & \mathbb{1} \end{bmatrix}$. The matrices Φ and D are given as $\Phi = \begin{bmatrix} \Phi_1 & 0 & \ldots & 0 \\ 0 & \Phi_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \Phi_J \end{bmatrix}$, $D = \begin{bmatrix} -A_2 & \mathbb{1} & 0 & \ldots & 0 & 0 \\ 0 & -A_3 & \mathbb{1} & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -A_J & \mathbb{1} \end{bmatrix}$. It is easy to check that the optimization problem in Eq. (7.23) is convex. Therefore, the solution can be estimated with parallel proximal splitting methods. In particular, the *prox* operators for the objective function and constraints in Eq. (7.23) can be easily computed by following the steps described in Section 7.3.1.

We analyze the multi-view joint reconstruction performance with the Tsukuba multi-view dataset (center, left, right, bottom and top views) [150]. Fig. 7.13 shows the quality of the reconstructed images in terms of average PSNR for the proposed joint reconstruction scheme and an independent decoding scheme. From Fig. 7.13 we observe that for a given measurement rate the proposed joint reconstruction scheme outperforms the independent scheme by a margin of 3 dB. Finally, we compare the joint reconstruction performance between the multi-view and stereo imaging scenarios in Fig. 7.13. We observe that the quality of the reconstructed images in multi-view scenario is 1.5 dB better than the stereo scenario; this is due to better correlation estimation in the former case. Therefore, it is clear that the quality of decoded images can be improved when more views are available at the decoder.

7.4 Conclusions

Our first contribution in this chapter is a novel and robust *asymmetric* distributed joint reconstruction algorithm for decoding correlated images from quantized linear measurements. We have formulated a convex optimization algorithm that ensures the reconstructed image to be consistent with the quantized measurements and also with the predicted image obtained by warping the reference image. We have enforced the measurement consistency using an l_p norm with $p > 2$ in order to account for the non-linearities of the quantization noise. We have experimentally shown that the proposed joint reconstruction solution is effective in capturing the texture and details in the predicted image. We have then shown that the RD performance of our asymmetric coding scheme computed on quantized linear measurements is far superior to state-of-the-art independent coding solutions based on JPEG 2000.

Our next contribution in this chapter is a novel and robust *symmetric* joint representation algorithm that decodes the images by exploiting the correlation model which is directly estimated in the compressed domain. We have formulated a new constraint convex optimization problem which encourages the sparsity prior of the reconstructed images and also uses both the correlation and the compressed linear measurements information. We have then shown that our optimization framework can be easily extended to decode multi-view images. Finally, we have shown experimentally that our proposed joint decoding solution reconstructs a better quality image compared to the independent decoding solutions for various multi-view and video datasets. This certainly positions our low complexity distributed coding frameworks for an effective 3D scene communication in multi-view or video imaging scenarios.

Chapter 8

Conclusions

8.1 Thesis achievements

This thesis has addressed the problem of effective distributed 3D scene representation and joint reconstruction for multiple correlated images, whose correlation is mainly driven by the displacement of objects. We have considered the use of various sensing methodologies for the scene representation which include planar, omnidirectional and compressed sensing cameras. Since the geometry of the visual representation is different for each sensor, we have proposed algorithms to process the data by properly considering the specific geometry of the imaging system.

First, we have proposed an innovative solution for the distributed representation and coding of correlated images captured in omnidirectional camera networks, with equivalent computational resources and transmission capabilities. We have proposed to work directly in the spherical framework in order to efficiently model the geometry of the omnidirectional images. The joint decoder effectively exploits the intra- and inter-view correlation based on image prediction and disparity compensation, respectively and improves on the coding performance of the independent coding solutions. Therefore, our rate balanced coding scheme certainly provides an interesting solution for simple networks of omnidirectional cameras, since it neither requires camera calibration nor hierarchy between sensors.

Next, we have studied the problem of distributed joint representation in a framework, where the correlated images are encoded independently at similar rates using standard coding schemes (e.g., SPIHT). In this setting, we have contributed a novel joint reconstruction algorithm that jointly decodes the images by taking benefit of the inter-view correlation. We have shown that our reconstruction algorithm is convex and we have proposed effective solutions based on proximal splitting methods. Simulation results confirm that the proposed joint representation algorithm is successful in improving the reconstruction quality of the compressed images for both planar and omnidirectional datasets. However, quality enhancement of the compressed images are not noticeable at low rates. We have provided insights into this scenario and we have described that this is due to the standard rate allocation scheme which fails to provide a compressed representation for an accurate disparity estimation. We have then contributed a novel rate allocation method based on modified SPIHT coding principles that allow for an accurate correlation estimation from the highly compressed images. Finally, for a given target bit rate, we have shown that there exists an effective trade-off between the accurate correlation estimation and the quality of image reconstruction due to different encoding optimization objectives; this implies that one has to encode different image characteristics for an effective distributed scene analysis or distributed coding of visual information captured in camera networks.

The next contribution of this thesis is a novel distributed representation scheme that estimates the underlying correlation model between a compressed reference image and a set of compressed images given in the form of quantized linear measurements. We have related the prominent features among multiple images

using a geometry-based correlation model. The depth or motion field is computed by solving a new regularized optimization problem under local geometrical transform constraints. We have also proposed innovative solutions to cope up with the measurement noise due to quantization and it has been successfully verified by experiments. Experimental results also demonstrate that the proposed methodology computes a good estimation of dense depth or motion field for various multi-view and video datasets. To the best of our knowledge, we are the first ones to show that the RD performance computed on compressed linear measurements (in distributed settings) outperform state-of-the-art independent and distributed coding schemes due to accurate and robust correlation estimation. This clearly positions our scheme as an effective solution for distributed image processing with low encoding complexity.

Then, we have contributed a novel and robust framework that estimates the correlation among multiple images directly in the compressed domain without reconstructing the images. We have proposed linear representations for disparity and motion models and we have shown that the correlation can be estimated in the compressed domain; thanks to the distance preserving property of the sensing matrices. We have proposed to estimate a robust correlation model in an energy minimization framework that is solved effectively using Graph Cuts. We have experimentally justified that our novel algorithm estimates a good correlation model even when the measurements are highly quantized. Simulation results further confirm that, our low complexity correlation estimation solution competes with the solution of a scheme that estimates the correlation model after reconstructing the images. Therefore, our correlation estimation framework from compressed measurements certainly provides an effective solution for distributed scene analysis or coding applications in low complexity sensor networks.

Finally, we have proposed novel and robust joint reconstruction algorithms (asymmetric and symmetric) that take benefit of the correlation estimation to decode the images from quantized linear measurements. The joint reconstruction is cast as an optimization problem that decodes sparse or smooth images by enforcing consistency with both the quantized linear measurements and correlation estimation. We have shown that the proposed optimization problems are convex and solved effectively using parallel proximal splitting methods. Simulation results confirm that the proposed joint representation scheme effectively exploits the inter-view correlation to achieve high compression gains at low bit rates when compared to state-of-the-art independent coding solutions; this illustrates the potential of our scheme in distributed multi-view or video coding applications. Therefore, our joint representation scheme certainly provides a low-complexity coding solution for the effective representation of 3D scenes in multi-view or video imaging scenarios.

8.2 Future directions

In this thesis, we have addressed the problem of designing efficient and robust distributed algorithms to jointly process the compressed visual information from various sensing methods, i.e., planar, omnidirectional and CS cameras. We have shown experimentally that our novel correlation estimation algorithms effectively handles the compressed images that allows for an improved joint reconstruction image quality. The frameworks presented in this thesis opens new exciting directions for further research.

In Chapter 3, we have used half of the entropy coded visual information from both images for the correlation estimation at the joint decoder. However, we did not consider the other half of compressed visual information given in terms of parity bits for the disparity estimation. Recently, Varodayan *et al.* [86] demonstrated that the underlying disparity model between images can be learned from the compressed syndrome bits based on Expectation-Maximization (EM) principles. Therefore, it would be interesting to merge the entropy and Slepian-Wolf coded visual information to estimate the correlation model at the decoder. We can benefit from both forms of compressed visual informations for the correlation estimation; this might lead to improved correlation estimation accuracy. Also, it would be interesting to extend our rate balanced omnidirectional coding scheme to encode correlated planar images.

In Chapter 4, we have developed a rate allocation scheme that identifies and encodes the low contrast

regions in the images, which permits to estimate an accurate correlation information for a given bit rate. However, the decoder estimates the correlation model directly from the compressed images, without properly considering the quantization noise or compression artifacts. Different cost functions have been proposed in [159] to efficiently consider the radiometric variations (e.g., brightness and contrast variations) between images. However, it is well known that these cost functions fail to efficiently handle the compression artifacts. Therefore, developing robust matching terms to estimate an accurate correlation model from highly compressed images would be an interesting topic of future research. For a good starting point, one could borrow ideas from the robust data fidelity functions proposed in Chapters 5 and 6.

In Chapters 5 and 6, we have proposed correlation estimation algorithms that build the underlying scene geometry from the compressed linear measurements. In our energy minimization framework however, we did not consider the presence of occlusions between multi-view images. Intuitively, the sets of occluded pixels can be identified by including additional terms in our framework, in order to enforce global visibility and uniqueness constraints [160, 161]. Extending our correlation estimation scheme for efficient occlusion handling can be a very interesting and challenging future research problem. Finally, it would be interesting to extend our low complexity distributed solutions presented in Chapters 5, 6 and 7 to the case of dynamic scenes simultaneously recorded by two or more cameras. The main goal in this case is to estimate a correlation model in terms of 3D scene flow that determines both the 3D geometry and the 3D motion of the scene at every pixel from compressed linear measurements.

Appendix A

Appendix

A.1 Motion field estimation from atom transforms

We describe here the methodologies to estimate the dense transformation field $\mathbf{f}$ from the sets of atoms $\{g_{\gamma_i}\}$ and $\{g_{\gamma'_i}\}$ that are related with the geometric transformations $\{F^i\}$. That is, the i^{th} pair of atoms g_{γ_i} and $g_{\gamma'_i}$ are related by

$$g_{\gamma'_i} = F^i(g_{\gamma_i}) = g_{\delta\gamma_i \circ \gamma_i}, \tag{A.1}$$

where $\gamma_i = (t_x, t_y, \theta, s_x, s_y)$, $\gamma'_i = (t'_x, t'_y, \theta', s'_x, s'_y)$ and $\delta\gamma_i = (t_x - t'_x, t_y - t'_y, \theta - \theta', s'_x/s_x, s'_y/s_y)$. The scenario is illustrated in Fig. A.1 using $K = 4$ atoms where we see that it is trivial to find the transformation solution $\mathbf{f}$ for the pixels in the non-overlapping regions. However, the pixels in overlapping regions have multiple transformation values $\{\delta\gamma_i\}$ that represent possible solutions. For example, as shown in Fig. A.1, the pixel $\mathbf{z}$ in the overlapping region $g_{\gamma_1} \cap g_{\gamma_2}$ has two solutions given by $\delta\gamma_1$ and $\delta\gamma_2$. Therefore, it is clear that the dense transformation field estimation from the sets of geometrically transformed atoms is an ill-posed problem. We propose here two methodologies based on energy minimization framework to estimate a dense transformation field $\mathbf{f}$ from the geometrically transformed atoms $\{g_{\gamma_i}\}$ and $\{g_{\gamma'_i}\}$.

Our first framework effectively considers all the transformations $\{\delta\gamma_{k(\mathbf{z})}\}$ at the given pixel $\mathbf{z}$ in order to generate a smooth transformation field, where $k(\mathbf{z}) \in \{1, 2, \cdots, K\}$ (recall that K is the number of atoms). Our solution is based on an energy minimization framework that estimates a transformation field that is

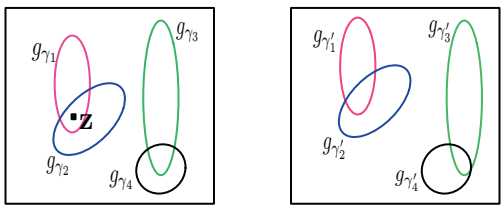


Figure A.1: Transformation field estimation from sets of atom transformations. The same colored atoms between images are related with a transformation F^i as $g_{\gamma'_i} = F^i(g_{\gamma_i}) = g_{\delta\gamma_i \circ \gamma_i}$. The pixels in the non-overlapping regions have a unique transformation solution, and the pixels in overlapping regions have more than one solution. For example, the transformation field solutions at the pixel $\mathbf{z}$ are $\delta\gamma_1$ and $\delta\gamma_2$.

consistent with the given transformation values $\{\delta\gamma_{k(\mathbf{z})}\}, \forall \mathbf{z}$. At the same time, we ensure coherency in the transformation field in order to model consistent motion of visual objects. The proposed regularized energy model is represented as

$$\mathcal{E}(\mathbf{f}, k) = \mathcal{E}_d(\mathbf{f}, k) + \lambda_1 \mathcal{E}_s(\mathbf{f}, k) + \lambda_2 \mathcal{E}_I(\mathbf{f}, k), \tag{TFE-1}$$

where k represents the dense atom index image with $k(\mathbf{z}) \in \{1, 2, \cdots, K\}$. The parameters λ_1 and λ_2 balance the data term $\mathcal{E}_d$ and the smoothness terms $\mathcal{E}_s$, $\mathcal{E}_I$.

Now, we describe the three costs functions used in our framework. The data term $\mathcal{E}_d(\mathbf{f}, k)$ finds the amount of disagreement between the estimated solution $\mathbf{f}$ and the given transformation model $\{\delta\gamma_{k(\mathbf{z})}\}$ given in Eq. (A.1). In other words, the data term finds how well a particular solution $\mathbf{f}$ fits with the given observation model $\{\delta\gamma_{k(\mathbf{z})}\}$. The data term is defined as

$$\mathcal{E}_d(\mathbf{f}, k) = \sum_{\mathbf{z} \in \mathcal{Z}} \mathcal{C}(\mathbf{f}(\mathbf{z}), k(\mathbf{z})), \tag{A.2}$$

where $\mathcal{Z}$ denote the total number of pixels, i.e., $|\mathcal{Z}| = N$, where N represents the resolution of the image. The term $\mathcal{C}(\mathbf{f}(\mathbf{z}), k(\mathbf{z}))$ is given as

$$\mathcal{C}(\mathbf{f}(\mathbf{z}), k(\mathbf{z})) = \| g_{\gamma'_{k(\mathbf{z})}} - g_{\mathbf{f}(\mathbf{z}) \circ \gamma_{k(\mathbf{z})}} \|_2^2 \, \mathcal{B}_{k(\mathbf{z})}(\mathbf{z}). \tag{A.3}$$

where $\mathcal{B}_i$ be a binary function that assigns the value 1 if a particular pixel $\mathbf{z} = (m, n)$ belongs to the support of the atom g_{γ_i} and ∞ otherwise. Mathematically, it reads as

$$\mathcal{B}_i(\mathbf{z}) = \begin{cases} 1 & \text{if } \mathbf{z} \in \mathcal{Z}_i, \\ \infty & \text{otherwise.} \end{cases} \tag{A.4}$$

Recall that $\mathcal{Z}_i$ represents the set of pixels in the support of the atom g_{γ_i} (see Eq. (5.12)). From Eq. (A.3), it is clear that $\mathcal{C}(\mathbf{f}(\mathbf{z}), k(\mathbf{z})) = 0, \forall \mathbf{z} \in \mathcal{Z}_i$ when the transformation label $\mathbf{f}(\mathbf{z}) = \delta\gamma_i$ where $i = k(\mathbf{z})$, i.e., $\| g_{\gamma'_i} - g_{\mathbf{f}(\mathbf{z}) \circ \gamma_i} \|_2^2 = 0$. This follows from Eq. (A.1) where i^{th} atom pairs g_{γ_i} and $g_{\gamma'_i}$ are related by $g_{\gamma'_i} = g_{\delta\gamma_i \circ \gamma_i}$. Therefore, the proposed fidelity term $\mathcal{C}(\mathbf{f}(\mathbf{z}), k(\mathbf{z}))$ assigns zero for all pixels in the support of atom g_{γ_i} when the transformation label $\mathbf{f}(\mathbf{z})$ satisfies with the observation model. Furthermore, it is worth mentioning that for a pixel $\mathbf{z}$ in the overlapping region the cost $\mathcal{C}(\mathbf{f}(\mathbf{z}), k(\mathbf{z}))$ is zero for several labels (see Fig. A.1). Due to this ambiguity, it is not possible to solve the problem uniquely based only on the data cost $\mathcal{E}_d$. We therefore rely on two additional smoothness prior terms $\mathcal{E}_s$ and $\mathcal{E}_I$ to resolve the ambiguity. Now, we explain the smoothness terms $\mathcal{E}_s$ and $\mathcal{E}_I$.

The smoothness cost $\mathcal{E}_s$ represents the penalty of assigning different transformations to adjacent pixels, so that it results in a coherent deformation field. Instead of enforcing the smoothness constraint directly on the transformation field, we propose to smooth the underlying motion field estimated from the sets of atom transformations, since the correlation between images usually takes the form of motion or disparity vectors in video or multi-view images (see Section 5.3.3). For a given transformation field $\mathbf{f}$, we first generate a dense motion field using Eq. (5.9), and then we compute the smoothness cost using Eq. (5.10).

The third term $\mathcal{E}_I$ smoothens the atom indexes of the neighboring pixels, so that the transformation values of adjacent pixels tend to come from the same atom. We propose to measure the cost $\mathcal{E}_I$ as

$$\mathcal{E}_I(\mathbf{f}, k) = \sum_{\mathbf{z}, \mathbf{z}' \in \mathcal{N}} I_{\mathbf{z}, \mathbf{z}'}(k(\mathbf{z}), k(\mathbf{z}')) \tag{A.5}$$

$$= \sum_{\mathbf{z}, \mathbf{z}' \in \mathcal{N}} min(\mathcal{T}_{k(\mathbf{z})}, \mathcal{T}_{k(\mathbf{z}')}) \delta_{k(\mathbf{z}), k(\mathbf{z}')}, \tag{A.6}$$

where $\mathbf{z}$ and $\mathbf{z}'$ are the adjacent pixels with atom indices $k(\mathbf{z})$ and $k(\mathbf{z}')$ respectively. The term $\mathcal{T}_{k(\mathbf{z})}$ represents the total number of pixels in the support of the atom $g_{\gamma_{k(\mathbf{z})}}$, i.e., $\mathcal{T}_{k(\mathbf{z})} = |\mathcal{Z}_{k(\mathbf{z})}|$ (see Eq. (5.12)).

From Eq. (A.6), it is clear that $I_{\mathbf{z},\mathbf{z}'}$ is zero if the pixels $\mathbf{z}$ and $\mathbf{z}'$ belong to the same atom, i.e., $k(\mathbf{z}) = k(\mathbf{z}')$. However, $I_{\mathbf{z},\mathbf{z}'}$ is nonzero in the regions where the atoms intersect and the amount of penalty is then given as $min(\mathcal{T}_{k(\mathbf{z})}, \mathcal{T}_{k(\mathbf{z}')})$. By taking the minimum cost between $\mathcal{T}_{k(\mathbf{z})}$ and $\mathcal{T}_{k(\mathbf{z}')}$ we enforce less penalty to the thin atoms when compared to the fat atoms. This encourages the presence of thin atoms in the final solution, so that the transformations along the edges of the objects can be effectively captured.

Finally, the dense transformation solution $(\hat{\mathbf{f}}, \hat{k})$ is estimated by minimizing the energy function $\mathcal{E}(\mathbf{f}, k)$ given as

$$(\hat{\mathbf{f}}, \hat{k}) = \arg\min_{\mathbf{f},k} \ \mathcal{E}(\mathbf{f}, k). \tag{A.7}$$

The problem given in Eq. (A.7) is non-convex, and therefore a solution to Eq. (A.7) can be estimated by using strong optimization techniques based on Graph Cuts [130, 133] or Belief Propagation [132].

We now describe the second approach to generate the dense transformation field. This approach is also based on an energy minimization framework, where the energy is given as

$$\tilde{\mathcal{E}}(\mathbf{f}) = \tilde{\mathcal{E}}_d(\mathbf{f}) + \lambda_1 \mathcal{E}_s(\mathbf{f}), \tag{TFE-2}$$

where $\mathcal{E}_s$ represents the smoothness term discussed earlier. The data term $\tilde{\mathcal{E}}_d$ reads as

$$\tilde{\mathcal{E}}_d(\mathbf{f}) = \sum_{\mathbf{z}\in\mathcal{Z}} \tilde{\mathcal{C}}(\mathbf{f}(\mathbf{z})) = \sum_{\mathbf{z}\in\mathcal{Z}} \sum_{i=1}^{K} g_{\gamma_i}(\mathbf{z}) \| g_{\gamma_i'} - g_{\mathbf{f}(\mathbf{z})\circ\gamma_i} \|_2^2, \tag{A.8}$$

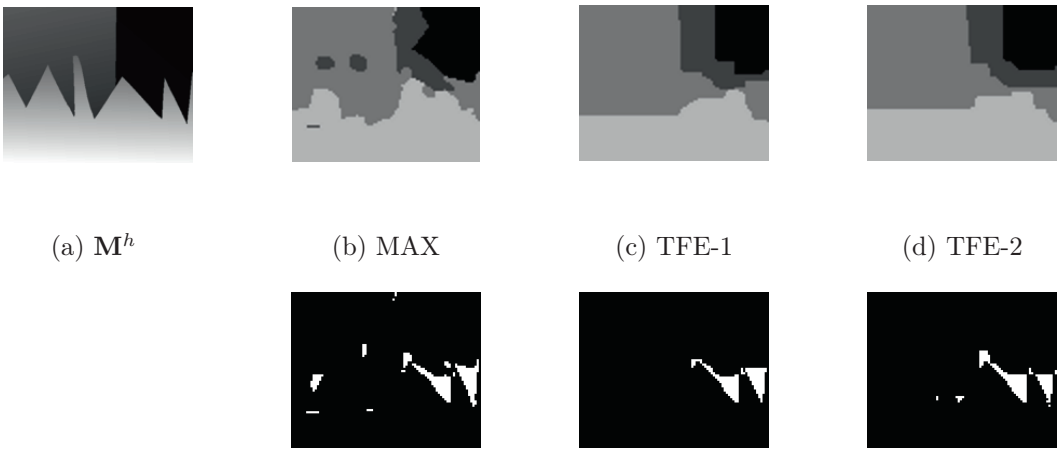
where $g_{\gamma_i}(\mathbf{z})$ represents the response of the i^{th} atom at location $\mathbf{z}$. The correlation solution $\acute{\mathbf{f}}$ can be estimated as a solution to following minimization problem:

$$\acute{\mathbf{f}} = \arg\min_{\mathbf{f}} \ \tilde{\mathcal{E}}(\mathbf{f}). \tag{A.9}$$

The solution to the above equation can be obtained using energy minimization techniques based on Graph Cuts or Belief Propagation. Finally, by comparing Eq. (A.7) and Eq. (A.9), it can be noted that the size of the search space in TFE-2 is reduced by an order K when compared to the TFE-1 scheme. Therefore, the TFE-2 problem is relatively less complex compared to TFE-1. However, it should be noted that MAX scheme (see Eq. (5.13)) is the simplest one to build the transformation field, as the TFE-1 and TFE-2 schemes are based on solving a regularized optimization problem. In the experimental results section, we compare the performances among these three schemes and we show that there is a clear trade-off between the performance and the complexity.

A.1.1 Performance analysis

We now compare the performances among the three transformation field estimation schemes based on MAX, TFE-1 and TFE-2. For simplicity, we compare the performances in a scenario where the sets of atoms $\{g_{\gamma_i'}\}$ are estimated from the 2-bit quantized measurements using the local optimization methodology. In particular, we estimate atoms $\{g_{\gamma_i'}\}$ with OPT-2 optimization problem described in Section 5.5. After estimating the sets of atoms $\{g_{\gamma_i'}\}$ for a given measurement rate, we estimate a dense disparity model based on MAX, TFE-1 and TFE-2 schemes. The optimization problems TFE-1 and TFE-2 are solved using Graph Cuts. Figs. A.2(b), (c) and (d) show the disparity solutions estimated with the MAX, TFE-1 and TFE-2 schemes respectively, for a measurement rate of 0.2 that is further quantized with a 2-bit quantizer. Compared to the groundtruth image given in Fig. A.2(a), we see that the TFE-1 and TFE-2 optimization schemes improve the disparity estimation solution when compared to the solution obtained using the MAX scheme given in Fig. A.2(b). Especially, we observe the improvement along the vertical edges in the center of the disparity image. We also see that the TFE-1 scheme estimates a better solution than TFE-2, where the DE is 3.13% for the former and 3.37% for the latter. This is because the TFE-1 scheme performs an



(a) $\mathbf{M}^h$ (b) MAX (c) TFE-1 (d) TFE-2

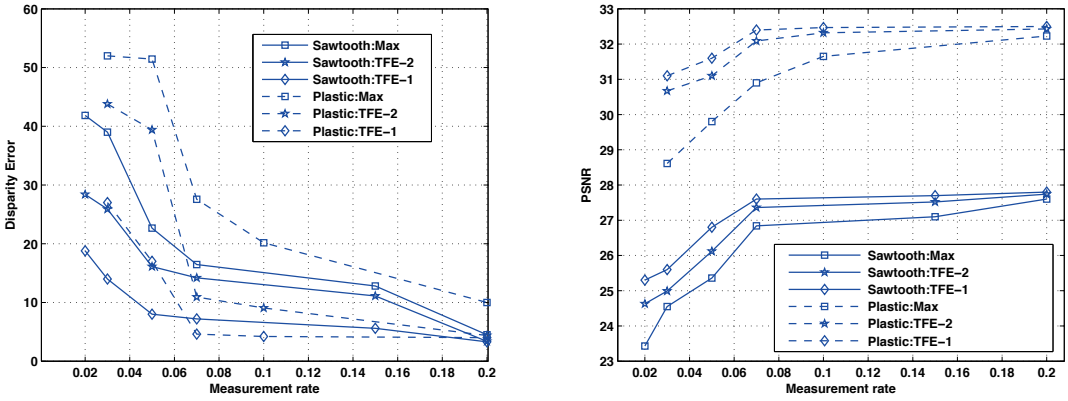
(f) DE: 4.8% (g) DE: 3.13% (h) DE: 3.37%

Figure A.2: Performance comparison of the disparity results among the MAX, TFE-1 and TFE-2 schemes in the Sawtooth dataset. (a) Groundtruth disparity field $\mathbf{M}^h$ between views 1 and 5. Top row: Disparity fields estimated with MAX, TFE-1 and TFE-2 schemes respectively. Bottom row: Respective disparity errors with DE=4.8%, 3.13%, 3.37%. The white pixels denote an error larger than 1.

exhaustive search by effectively considering all the possible transformation solutions at a given pixel as discussed above. Fig. A.3(a) and Fig. A.3(b) compare the accuracy of the disparity estimation solution between the MAX, TFE-1 and TFE-2 schemes in terms of disparity error and image prediction quality respectively. For a given measurement rate, we see from Fig. A.3 that the disparity estimation solution based on TFE-1 gives better results compared to the TFE-2 and MAX schemes. Therefore, it is clear that there is a trade-off between the complexity and the performance. However, from Fig. A.3 we see that the MAX scheme performance is reasonably good if one advantageously considers the low complexity of this approach.

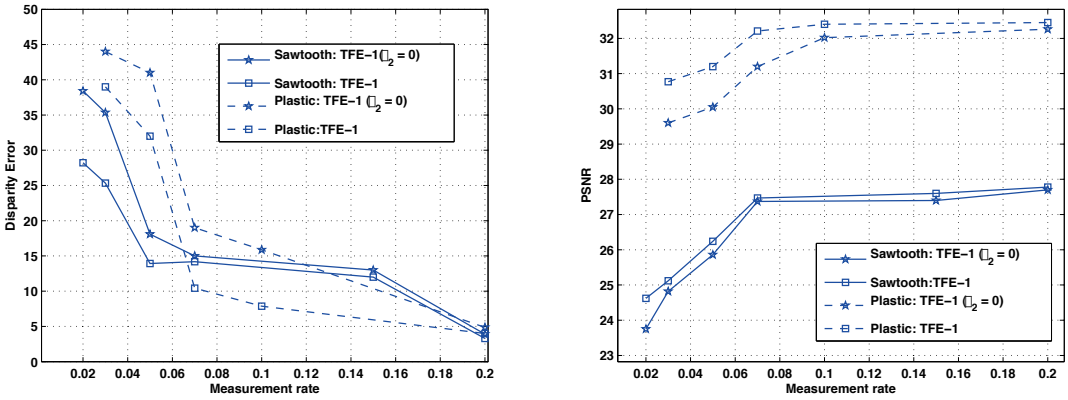
We now illustrate the benefit of using the smoothness priors, and the effect of regularization parameters in the TFE-1 and TFE-2 schemes. We first highlight the benefit of activating the smoothness term $\mathcal{E}_I$ in the optimization TFE-1. We first solve the TFE-1 scheme without activating the $\mathcal{E}_I$ term, i.e., $\lambda_2 = 0$. Fig. A.4 compares the performance of this setting with respect to the TFE-1 scheme where all the terms are activated, i.e., $\lambda_1 \neq 0$ and $\lambda_2 \neq 0$. It is clear from Fig. A.4 that the disparity error and the prediction image quality are improved by activating the term $\mathcal{E}_I$. While carrying out the experiments, we have further observed a degradation in performance when the TFE-1 is solved only with the $\mathcal{E}_I$ term and not using the smoothness $\mathcal{E}_s$ term, i.e., $\lambda_1 = 0$ and $\lambda_2 \neq 0$. Finally, we study the sensitivity of the parameter λ_1 in TFE-2 scheme. Fig. A.5 shows the disparity error (left side) and image prediction quality (right side) for various choices of λ_1 between 0 to 180 with a step-size 5. From Fig. A.5, we see that the TFE-2 scheme gives a similar disparity image with DE $\approx 3.5\%$ for λ_1 between 40 to 80. Similarly, the quality of the predicted image is approximately 27.8 dB for λ_1 between 20 to 70, which illustrates the insensitivity of our proposed scheme to the exact value of λ_1. When the parameter λ_1 is increased beyond 100, we notice that the disparity becomes over smooth; this leads to a bad disparity estimation solution and also to a poor image prediction quality.

We now compare the estimated motion field among the MAX, TFE-1 and TFE-2 schemes using a similar experimental setup in the stereo imaging framework, i.e., we compare the performances in a scenario



(a) (b)

Figure A.3: Performance comparison for the disparity image computed using MAX, TFE-1 and TFE-2 schemes for the Sawtooth and Plastic datasets. For a given measurement rate the accuracy of the disparity estimation is evaluated in terms of (a) disparity error and (b) quality of image prediction. These plots are generated from 2-bit quantized measurements.



(a) (b)

Figure A.4: Performance comparison for the disparity image computed using the TFE-1 scheme with and without $\mathcal{E}_I$. For a given measurement rate the accuracy of the disparity estimation is evaluated in terms of (a) disparity error and (b) quality of image prediction. These plots are generated from 2-bit quantized measurements.

where the sets of atoms $\{g_{\gamma_i'}\}$ are estimated from 2-bit quantized measurements by solving OPT-2 scheme using local optimization methodology. Figs. A.6(b), (c) and (d) compare the motion fields for the Foreman sequence that are estimated from 2-bit quantized measurements at the rate 0.15. Comparing Figs. A.6(b), and (c) we see that the motion field in the face region is better represented with the TFE-1 scheme compared to the MAX scheme. This can be clearly observed in the prediction errors given in Figs. A.6(f) and (g) when the motion fields respectively in Figs. A.6(b) and (c) are used for image prediction. We then compare in

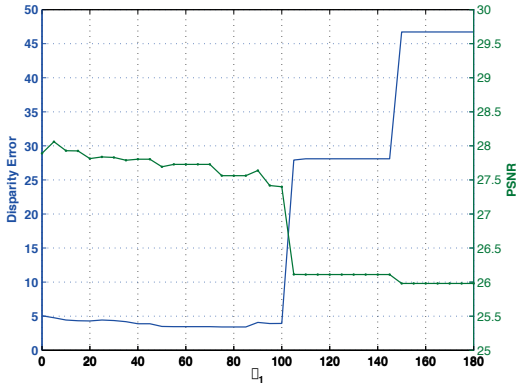
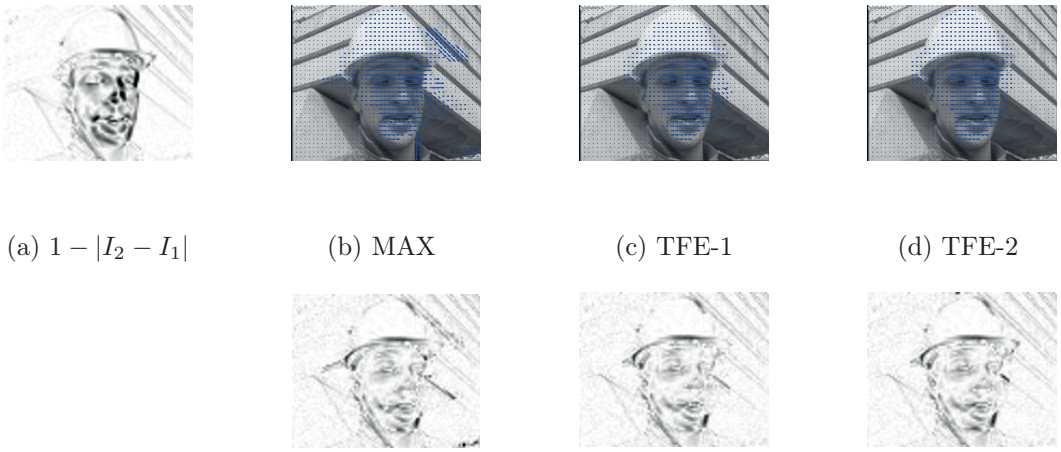


Figure A.5: Effect of the regularization parameter λ_1 on the disparity estimation performance with the TFE-2 scheme in the Sawtooth dataset. The left and right side of the y-axis show the disparity error and the quality of the predicted image respectively.



(a) $1 - |I_2 - I_1|$ (b) MAX (c) TFE-1 (d) TFE-2

(f) MSE: 41.90 (g) MSE: 34.85 (h) MSE: 38.17

Figure A.6: Performance comparison among the MAX, TFE-1 and TFE-2 schemes in the Foreman dataset. (a) Inverted absolute error between original images I_1 and I_2. MSE between I_2 and I_1 is 79.5. Top row: Motion fields computed with MAX, TFE-1 and TFE-2 schemes respectively. Bottom row: Respective prediction errors w.r.t. I_2. The error is inverted so that the white pixels correspond to no error.

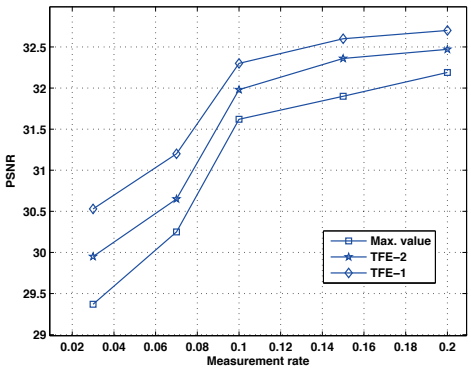


Figure A.7: Performance comparison of the quality of the predicted image using the motion field computed with MAX, TFE-1 and TFE-2 schemes in the Foreman dataset.

Fig. A.7, the prediction image quality among the three schemes for measurement rates from 0.03 to 0.2. As observed earlier in disparity scenario, we notice (see Fig. A.7) that the complex TFE-1 scheme outperforms the TFE-2 and MAX schemes, where the gain reaches up to 1 dB at low rates. However, it is interesting to note that the MAX scheme estimates a reasonably good solution with low computational complexity; this encourages to use this approach in our correlation estimation OPT-1 and OPT-2 problems described in Chapter 5.

Bibliography

[1] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[2] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer, 2003.

[3] L. G. Shapiro and G. C. Stockman, *Computer Vision*. Prentice Hall, 2001.

[4] M. Flierl and B. Girod, "Multiview video compression," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 66–76, 2007.

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

[6] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, pp. 471 – 480, 1973.

[7] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense stereo," *International Journal on Computer Vision*, vol. 47, pp. 7–42, 2002.

[8] S. Baker, S. Roth, D. Scharstein, M. Black, J. Lewis, and R. Szeliski, "A database and evaluation methodology for optical flow," *International Journal of Computer Vision*, vol. 1, no. 92, pp. 1–31, 2011.

[9] E. H. Adelson and J. R. Bergen, *Computational Models of Visual Processing*. Cambridge, MA: MIT Press, 1991.

[10] C. Geyer and K. Daniilidis, "Catadioptric projective geometry," *International Journal of Computer Vision*, vol. 45, no. 3, pp. 223–243, 2001.

[11] Y. Yagi, "Omnidirectional sensing and its applications," *IEICE Transactions on Information And Systems*, vol. E82-D, no. 3, pp. 568–579, 1999.

[12] S. Nayar, "Omnidirectional vision," in *Proc. International Symposium of Robotics Research*, 1997.

[13] T. Svoboda, T. Pajdla, and V. Hlavac, "Epipolar geometry for panoramic cameras," in *Proc. European Conference on Computer Vision*, 1998.

[14] S. Nene and S. K. Nayar, "Stereo with mirrors," in *Proc. European Conference on Computer Vision*, 1998.

[15] D. Donoho, "Compressed sensing," *IEEE Trans. Information Theory*, vol. 52, pp. 1289–1306, 2006.

[16] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Information Theory*, vol. 52, pp. 489–509, 2006.

[17] H. Mamaghanian, N. Khaled, D. Atienza, and P. Vandergheynst, "Compressed sensing for real-time energy-efficient ecg compression on wireless body sensor nodes," *IEEE Transcations on Biomedical Engineering*, vol. 58, no. 12, pp. 120–129, 2011.

[18] E. J. Candes and J. Romberg, "Practical signal recovery from random projections," in *Proc. SPIE Computational Imaging*, 2005.

[19] G. Peyrè, "Literature review on sparse optimization," 2008. [Online]. Available: http://www.ceremade.dauphine.fr/ peyre/cs-tv/

[20] J. A. Tropp and S. J. Wright, "Computational methods for sparse solution of linear inverse problems," *Proc. of IEEE*, vol. 98, no. 6, pp. 948–958, 2010.

[21] A. Rakotomamonjy, "Surveying and comparing simultaneous sparse approximation (or group-lasso) algorithms," *Signal Processing: Image Communication*, vol. 91, no. 7, pp. 1505–1526, 2011.

[22] M. Wakin, J. Laska, M. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. Kelly, and R. Baraniuk, "An architecture for compressive imaging," *Proc IEEE International Conference on Image Proc*, 2006.

[23] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk, "Single-pixel imaging via compressive sampling," *Signal Processing Magazine, IEEE*, vol. 25, no. 2, pp. 83–91, 2008.

[24] R. Fergus, A. Torralba, and W. Freeman, "Random lens imaging," MIT-CSAIL-TR-2006-058, Tech. Rep., 2006.

[25] R. Marcia and R. Willett, "Compressive coded aperture superresolution image reconstruction," *Proc IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[26] J. Romberg, "Sensing by random convolution," *Proc IEEE Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, 2007.

[27] L. Jacques, P. Vandergheynst, A. Bibet, V. Majidzadeh, A. Schmid, and Y. Leblebici, "CMOS compressed imaging by random convolution," *Proc IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[28] J. Romberg, "Compressive sensing by random convolution," *SIAM J. Imaging Sciences*, vol. 2, no. 4, pp. 1098–1128, 2009.

[29] R. Robucci, L. K. Chiu, J. Gray, J. Romberg, P. Hasler, and D. Anderson, "Compressive sensing on a CMOS separable transform image sensor," *Proc IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[30] L. Gan, T. T. Do, and T. D. Tran, "Fast compressive imaging using scrambled hadamard ensemble," in *Proc. European Signal and Image Processing Conference*, 2008.

[31] T. Do, T. Tran, and L. Gan, "Fast compressive sampling with structurally random matrices," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

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

[33] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press, 2008.

[34] W. B. Pennebaker and J. L. Mitchell, "JPEG: Still image compression standard," in *Van Nostrand Reinhold*, 1993.

[35] C. A. Christopoulos, A. N. Skodras, and T. Ebrahimi, "The JPEG 2000 still image coding system: An overview," *IEEE Trans. on Consumer Electronics*, vol. 46, no. 4, pp. 1103–1127, 2000.

[36] E. L. Pennec and S. Mallat, "Sparse geometric image representation with bandelets," *IEEE Trans. Image Processing*, vol. 14, no. 4, pp. 423–438, 2005.

[37] M. N. Do and M. Vetterli, "The contourlet transform: an efficient directional multiresolution image representation," *IEEE Trans. Image Processing*, vol. 14, no. 12, pp. 2091–2106, 2005.

[38] R. M. Figueras, P. Vandergheynst, and P. Frossard, "Low-rate and flexible image coding with redundant representations," *IEEE Trans. Image Processing*, vol. 15, pp. 726–739, 2006.

[39] I. Tosic, I. Bogdanova, P. Frossard, and P. Vandergheynst, "Multiresolution motion estimation for omnidirectional images," in *Proc. European Signal Processing Conference*, 2005.

[40] I. Bogdanova, J. Vandergheynst, P and. Antoine, L. Jacques, and M. Morvidone, "Discrete wavelet frames on the sphere," in *Proc. European Signal Processing Conference*, 2004.

[41] Y. Wiaux, J. D. McEwen, P. Vandergheynst, and O. Blanc, "Exact reconstruction with directional wavelets on the sphere," *Monthly Notices of the Royal Astronomical Society*, vol. 388, no. 2, pp. 770–788, 2008.

[42] P. Schröder and W. Sweldens, "Spherical wavelets: Efficiently representing functions on the sphere," in *Proc. of ACM SIGGRAPH*, 1995.

[43] I. Tosic, P. Frossard, and P. Vandergheynst, "Progressive coding of 3-d objects based on overcomplete decompositions," *IEEE Trans. on Circuts and Systems for Video Technology*, vol. 16, no. 11, pp. 1338–1349, 2006.

[44] A. Fletcher, S. Rangan, and V. Goyal, "On the rate-distortion performance of compressed sensing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[45] S. Sarvotham, D. Baron, and R. Baraniuk, "Measurements vs. bits: Compressed sensing meets information theory," in *Proc. Allerton Conference on Communication, Control and Computing*, 2006.

[46] A. Schulz, L. Velho, and E. A. B. da Silva, "On the empirical rate-distortion performance of compressive sensing," in *Proc. IEEE International Conference on Image Processing*, 2009.

[47] P. Boufounos and R. G. Baraniuk, "1-bit compressive sensing," in *Proc. International Conference on Information Science and Systems*, 2008.

[48] L. Jacques, D. K. Hammond, and M. J. Fadili, "Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine," *IEEE Trans. on Information Theory*, vol. 57, pp. 559–571, Jan 2011.

[49] A. Zymnis, S. Boyd, and E. Candes, "Compressed sensing with quantized measurements," *IEEE Signal Processing Letters*, vol. 17, no. 2, pp. 149–152, 2010.

[50] W. Dai, H. V. Pham, and O. Milenkovic. (2009) Distortion-rate functions for quantized compressive sensing. [Online]. Available: http://arxiv.org/abs/0901.0749

[51] J. Sun and V. Goyal, "Optimal quantization of random measurements in compressed sensing," in *Proc. IEEE International Symposium on Information Theory*, 2009.

[52] L. Wang, X. Wu, and G. Shi, "Progressive quantization of compressive sensing measurements," in *Proc. IEEE Data Compression Conference*, 2011.

[53] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin, and R. G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," in *Proc. Asilomar Conference on Signal System and Computing*, 2005.

[54] ——, "Universal distributed sensing via random projections," in *Proc. Information Processing in Sensor Networks*, 2006.

[55] S. Li and K. Fukumori, "Spherical stereo for the construction of immersive vr environment," in *Proc. IEEE Virtual Reality*, 2005.

[56] A. Torii, A. Imiya, and N. Ohnishi, "Two- and three- view geometry for spherical cameras," in *Proc. of the 6th Workshop on Omnidirectional Vision, Camera Networks and Non-classical cameras*, 2005.

[57] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artifical Intelligence*, vol. 17, pp. 185–203, 1981.

[58] B. Lucas and T. Kanade, "An iterative image registration technique with an application in stereo vision," in *Proc. DARPA IU Workshop*, 1981.

[59] A. Puri, R. V. Kollarits, and B. G. Haskell, "Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4," *Journal of Signal Processing Image Communication*, vol. 10, pp. 201–234, 1997.

[60] M. E. Lukacs, "Predictive coding of multi-viewpoint image sets," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 1986.

[61] W. Woo and A. Ortega, "Overlapped block disparity compensation with adaptive windows for stereo image coding," *IEEE Trans. on Circuts and Systems for Video Technology*, vol. 10, no. 4, pp. 194–200, 2000.

[62] I. Dinstein, M. G. Kim, J. Tselgov, and A. Henik, "Compression of stereo images and the evaluation of its effects on 3-d perception," in *Proc. Applications of Digital Image Processing XII, SPIE*, 1989.

[63] M. G. Perkins, "Data compression of stereopairs," *IEEE Trans. on Communications*, vol. 40, no. 4, pp. 684–696, 1992.

[64] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Trans. on Circuts and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[65] A. Smolic, P. Merkle, K. Müller, C. Fehn, P. Kauff, and T. Wiegand, *Compression of multi-view video and associated data*. Springer, 2008.

[66] M. Cossalter, G. Valenzise, M. Tagliasacchi, and S. Tubaro, "Joint compressive video coding and analysis," *IEEE Transactions on Multimedia*, vol. 12, no. 3, pp. 168–183, 2010.

[67] V. Stankovic, L. Stankovic, and S. Cheng, "Compressive video sampling," in *Proc. European Signal and Image Processing Conference*, 2008.

[68] J. Y. Park and M. B. Wakin, "A multiscale framework for compressive sensing of video," in *Proc. Picture Coding Symposium*, 2009.

[69] N. Vaswani, "Kalman filtered compressed sensing," in *Proc. IEEE International Conference on Image Processing*, 2008.

[70] O. D. Escoda, G. Monaci, R. M. Figueras, P. Vandergheynst, and M. Bierlaire, "Geometric video approximation using weighted matching pursuit," *IEEE Trans. Image Processing*, vol. 18, pp. 1703–1716, 2009.

[71] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side-information at the decoder," *IEEE Transactions on Information Theory*, vol. 22(1), pp. 1 –10, 1976.

[72] Z. Xiong, A. D. Liveris, and S. Cheng, "Distributed source coding for sensor networks," *IEEE Signal Processing Magazine*, vol. 21, pp. 80–94, 2004.

[73] S. S. Pradhan, J. Kusama, and K. Ramachandran, "Distributed compression in a dense microsensor network," *IEEE Signal Processing Magazine*, vol. 19, no. 2, pp. 51–60, 2002.

[74] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS)," *IEEE Transactions on Information Theory*, vol. 49(3), pp. 626–643, 2003.

[75] J. Garcia-Frias and Y. Zhao, "Compression of correlated binary sources using turbo codes," *IEEE Communication Letter*, vol. 5(10), pp. 417– 419, 2001.

[76] A. Aaron and B. Girod, "Compression with side information using turbo codes," in *Proc. IEEE Data Compression Conference*, 2002.

[77] Z. Liveris, A.Dand Xiong and C. Georghiades, "Compression of binary sources with side information at the decoder using LDPC codes," *IEEE Communication Letter*, vol. 6(1), pp. 440–442, 2002.

[78] M. Grangetto, E. E. Magli, and G. Olmo, "Distributed arithmetic coding for the Slepian-Wolf problem," *IEEE Trans. on Signal Processing*, vol. 57, no. 6, pp. 2245–2257, 2009.

[79] A. Aaron, S. Rane, E. Setton, and B. Girod, "Transform-domain Wyner-Ziv codec for video," in *Proc. SPIE Visual Communication and Image Processing*, 2004.

[80] R. Puri, A. Majumdar, and K. Ramachandran, "PRISM: A video coding paradigm with motion estimation at the decoder," *IEEE Trans. on Image Processing*, vol. 16, no. 10, pp. 2436–2448, 2007.

[81] X. Artigas, J. Ascenso, M. Dalai, S. Klomp, D. Kubasov, and M. Ouaret, "The DISCOVER codec: Architecture, techniques and evaluation," in *Proc. Picture Coding Symposium*, 2007.

[82] C. Brites and F. Pereira, "Correlation noise modeling for efficient pixel and transform domain Wyner-Ziv video coding," *IEEE Trans. on Ciruits and Systems for Video Techonolgy*, vol. 18, no. 9, pp. 1177–1190, 2008.

[83] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. of the IEEE*, vol. 93, pp. 71 – 83, 2005.

[84] C. Guillemot, F. Pereira, L. Torres, T. Ebrahimi, R. Leonardi, and J. Ostermann, "Distibuted monoview and multiview video coding," *IEEE Signal Processing Magazine*, vol. 24, no. 5, pp. 67–76, 2007.

[85] X. Zhu, A. Aaron, and B. Girod, "Distributed compression for large camera arrays," in *Proc. IEEE Statistical Signal Processing*, 2003.

[86] D. Varodayan, Y. C. Lin, A. Mavlankar, M. Flierl, and B. Girod, "Wyner-Ziv coding of stereo images with unsupervised learning of disparity," in *Proc. Picture Coding Symposium*, 2007.

[87] T. Tillo, B. Penna, P. Frossard, and P. Vandergheynst, "Distributed coding of spherical images with jointly refined decoding," in *Proc. IEEE Multimedia Signal Processing*, 2005.

[88] V. Thirumalai, I. Tosic, and P. Frossard, "Distributed coding of multiresolution omnidirectional images," in *Proc. IEEE International Conference on Image Processing*, 2007.

[89] N. Gehrig and P. Dragotti, "Distributed compression of multi-view images using a geometric approach," in *Proc. IEEE International Conference on Image Processing*, 2007.

[90] N. Gehrig and P. L. Dragotti, "Geometry-driven distributed compression of the plenoptic function: Performance bounds and constructive algorithms," *IEEE Trans. on Image Processing*, vol. 18, no. 3, pp. 457–470, 2009.

[91] R. Wagner, R. Nowak, and R. Baraniuk, "Distributed image compression for sensor networks using correspondence analysis and super-resolution," in *Proc. IEEE International Conference on Image Processing*, 2003.

[92] I. Tosic and P. Frossard, "Geometry based distributed scene representation with omnidirectional vision sensors," *IEEE Trans. Image Processing*, vol. 17, pp. 1033–1046, 2008.

[93] F. M. J. Willems, "Totally asynchronous Slepian-Wolf data compression," *IEEE Trans. Information Theory*, vol. 34(1), pp. 35 – 44, 1988.

[94] S. Pradhan and K. Ramchandran, "Distributed source coding: Symmetric rates and applications to sensor networks," in *Proc. IEEE Data Compression Conference*, 2000.

[95] V. Stankovic, A. D. Liveris, Z. Xiong, and C. N. Georghiades, "Design of Slepian-Wolf codes by channel code partitioning," in *Proc. IEEE Data Compression Conference*, 2004.

[96] P. Tan and J. Li, "A practical and optimal symmetric Slepian-Wolf compression strategy using syndrome formers and inverse syndrome formers," in *Proc. Allerton Conference on Communication, Control and Computing*, 2005.

[97] M. Sartipi and F. Fekri, "Distributed source coding in the wireless sensor networks using LDPC codes: The entire Slepian-Wolf rate region," in *Proc. IEEE Wireless Communications and Networking Conference*, 2005.

[98] M. Grangetto, E. Magli, and G. Olmo, "Symmetric distributed arithmetic coding of correlated sources," in *Proc. IEEE Multimedia Signal Processing*, 2007.

[99] F. Yang, Q. Dai, and G. Ding, "Multi-view images coding based on multiterminal source coding," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007.

[100] N. Gehrig and P. Dragotti, "Distributed compression in camera sensor networks," in *Proc. IEEE Multimedia Signal Processing*, 2004.

[101] L. W. Kang and C. S. Lu, "Distributed compressive video sensing," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[102] T. Do, Y. Chen, D. T. Nguyen, N. Nguyen, L. Gan, and T. D. Tran, "Distributed compressed video sensing," in *Proc. IEEE International Conference on Image Processing*, 2009.

[103] J. P. Nebot, Y. Ma, and T. Huang, "Distributed video coding using compressive sampling," in *Proc. Picture Coding Symposium*, 2009.

[104] M. Trocan, T. Maugey, J. E. Fowler, and B. Pesquet-Popescu, "Disparity-compensated compressed-sensing reconstruction for multiview images," *Proc. IEEE International Conference on Multimedia and Expo*, 2010.

[105] M. Trocan, T. Maugey, E. W. Tramel, J. E. Fowler, and B. Pesquet-Popescu, "Multistage compressed-sensing reconstruction of multiview images," in *Proc. IEEE International workshop on Multimedia Signal Processing*, 2010.

[106] M. B. Wakin, "A manifold lifting algorithm for multi-view compressive imaging," in *Proc. Picture Coding Symposium*, 2009.

[107] C. Fu, X. Ji, and Q. Dai, "Compressed multi-view imaging with joint reconstruction," in *Proc. IEEE Data Compression Conference*, 2011.

[108] X. Li, Z. Wei, and Z. Xiao, "Compressed sensing joint reconstruction for multi-view images," *Electronics Letters*, vol. 46, no. 23, pp. 1548–1550, Nov 2010.

[109] R. Zamir, S. Shamai, and U. Erez, "Nested linear/lattice codes for structured multiterminal binning," *IEEE Trans. on Infomation Theory*, vol. 48(6), pp. 1250 –1276, 2002.

[110] D. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. on Infomation Theory*, vol. 45(3), pp. 399 – 431, 1999.

[111] D. Mackay and R. Neal, "Near shannon limit performance of low density parity check codes," *Electronics Letters*, vol. 33(6), pp. 457 – 458, 1997.

[112] Z. Arican and P. Frossard, "Dense disparity estimation from omnidirectional images," in *Proc. IEEE International Conference on Advanced Video and Signal based Surveillance*, 2007.

[113] C. Geyer and K. Daniilidis, "Catadioptric projective geometry," *International Journal on Computer Vision*, vol. 45 (3), pp. 223 – 243, 2001.

[114] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Trans. on Communication*, vol. 31(4), pp. 532 – 540, 1983.

[115] W. Byerly, *An Elementary Treatise on Fourier's Series, and Spherical, Cylindrical, and Ellipsoidal Harmonics, with Applications to Problems in Mathematical Physics.* New York: Dover, 1959.

[116] D. Healy, D. Rockmore, P. Kostelec, and S. Moore, "FFTs for the 2-sphere - improvements and variations," *Journal of Fourier Analysis and Application*, vol. 9(3), pp. 341 – 385, 2003.

[117] J. L. Salinas and R. L. Baker, "Laplacian pyramid encoding: optimum rate and distortion allocations," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1989.

[118] Q. Xu and Z. Xiong, "Layered Wyner-Ziv video coding," *IEEE Trans. on Image Processing*, vol. 15(12), pp. 3791 – 3802, 2006.

[119] R. M. Neal, "Methods for constructing LDPC codes." [Online]. Available: http://www.cs.utoronto.ca/pub/radford/LDPC-2001-05-04/pchk.html

[120] S. Cheng and Z. Xiong, "Successive refinement for the Wyner-Ziv problem and layered code design," *IEEE Trans. on Signal Processing*, vol. 53(8), pp. 3269–3281, 2005.

[121] A. Jagmohan, A. Sehgal, and N. Ahuja, "Two-channel predictive multiple description coding," in *Proc. IEEE International Conference on Image Processing*, 2005.

[122] G. Rath and C. Guillemot, "Compressing the laplacian pyramid," in *Proc. IEEE Multimedia Signal Processing*, 2006.

[123] M. B. Schenkel, C. Luo, P. Frossard, and F. Wu, "Joint decoding of stereo image pairs," in *Proc. IEEE International Conference on Image Processing*, 2010.

[124] P. Lai, A. Ortega, C. Dorea, P. Yin, and C. Gomila, "Improving view rendering quality and coding efficiency by suppressing compression artifacts in depth-image coding," *Proc. of Visual Communic. and Image Proc.*, 2009.

[125] Y. Morvan, "Acquisition, compression and rendering of depth and texture for multi-view video," Ph.D. dissertation, Eindhoven University of Technology, 2009.

[126] R. Krishnamurthy, B. Chai, H. Tao, and S. Sethuraman, "Compression and transmission of depth maps for image-based rendering," *in Proc. IEEE International Conference on Image Proc.*, 2001.

[127] K. Klimaszewski, K. Wegner, and M. Domanski, "Distortions of synthesized views caused by compression of views and depth maps," in *Proc. 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*, 2009.

[128] A. Said and W. Pearlman, "A new, fast, and efficient image codec using set partitioning in hierarchical trees," *IEEE Trans. on Circuts and Systems for Video Technology*, vol. 6(3), pp. 243–250, 1996.

[129] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 4, pp. 401–406, 1998.

[130] O. Veksler, "Efficient graph based energy minimization methods in computer vision," Ph.D. dissertation, Cornell University, 1999.

[131] S. T. Barnard, "Stochastic stereo matching over scale," *International Journal of Computer Vision*, vol. 3, no. 1, pp. 17–32, 1989.

[132] P. Felzenszwalb and D. Huttenlocher, "Efficient belief propagation for early vision," *International Journal on Computer Vision*, vol. 70, no. 1, pp. 41–54, 2006.

[133] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, Jan 2002.

[134] V. Kolmogorov and R. Zabih, "What energy functions can be minimized via graph cuts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 26, no. 2, pp. 65–81, 2004.

[135] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, New York, 2004.

[136] P. L. Combettes and J.-C. Pesquet, "Proximal splitting methods in signal processing," *In Fixed Point Algorithms for Inverse Problems in Science and Engineering. Springer,*, 2010. [Online]. Available: http://arxiv.org/abs/0912.3522v4

[137] A. Chambolle, "An algorithm for total variation minimization and applications," *Jour. Math. Imag. Vis.*, pp. 89–97, 2004.

[138] M. Fadili and J. Starck, "Monotone operator splitting for optimization problems in sparse recovery," *Proc. IEEE International Conference on Image Processing*, pp. 1461–1464, 2010.

[139] G. Peyre, S. Bougleux, and L. D. Cohen, "Non-local regularization of inverse problems," in *Proc. European Conference on Computer Vision*, 2008.

[140] T. Tosic and P. Frossard, "Graph-based regularization for spherical signal interpolation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.

[141] D. Zhou and B. Scholkopf, "Regularization on discrete spaces," in *Pattern Recognition*. Springer, 2005, pp. 361–368.

[142] A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Comm. Pure and Appl. Math*, vol. 45, pp. 485–560, 1992.

[143] L. Zhang and S. M. Seitz, "Estimating optimal parameters for mrf stereo from a single image pair," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 29, no. 2, pp. 331–342, 2007.

[144] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. on Signal Processing*, vol. 41, no. 12, pp. 3445–3462, 1993.

[145] D. Taubman, "High performance scalable image compression with ebcot," *IEEE Trans. on Image Processing*, vol. 9, no. 7, pp. 1158–1170, 2000.

[146] G. Bjontegaard, "Calculation of average PSNR differences between RD curves," 13th VCEG-M33 Meeting, Austin, TX, USA, Tech. Rep., 2001.

[147] D. Varodayan, D. Chen, M. Flierl, and B. Girod, "Wyner-Ziv coding of video with unsupervised motion vector learning," *EURASIP Signal Processing: Image Communication*, vol. 23, pp. 369–378, 2008.

[148] G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, 1993.

[149] P. Jost, P. Vandergheynst, and P. Frossard, "Tree-based pursuit: Algorithm and properties," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4685–4697, 2006.

[150] V. Kolmogorov and R. Zabih, "Multi-camera scene reconstruction via graph cuts," *Computer Vision—ECCV 2002*, pp. 8–40, 2002.

[151] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," http://cvxr.com/cvx, Apr. 2011.

[152] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for markov random fields with smoothness-based priors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 6, pp. 1068–1080, Jun 2008.

[153] R. Baraniuk, M. Davenport, and R. DeVore, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation, Springer*, vol. 28, pp. 253–263, Jan 2008.

[154] T. Do, L. Gan, Y. Chen, N. Nguyen, and T. Tran, "Fast and efficient dimensionality reduction using structurally random matrices," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009.

[155] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems," *IEEE Journal of Selected Topics in Signal Processing*, vol. 1, pp. 586–597, 2007.

[156] D. K. Hammond, L. Jacques, M. J. Fadili, G. Puy, and P. Vandergheynst, "The basis pursuit dequantizer (bpdq) toolbox," July 2009. [Online]. Available: http://wiki.epfl.ch/bpdq

[157] S. Becker, J. Bobin, and E. Candes, "NESTA: A fast and accurate first-order method for sparse recovery," Caltech, Tech. Rep., 2009.

[158] Y. L. Montagner, E. Angelini, and J.-C. Olivo-Marin, "Comparison of reconstruction algorithms in compressed sensing applied to biological imaging," in *Proc. IEEE International Symposium on Biological Imaging*, 2011.

[159] H. Hirschmuller and D. Scharstein, "Evaluation of stereo matching costs on images with radiometric differences," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 9, pp. 1582–1509, 2009.

[160] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions via graph cuts," in *Proc. of the International Conference on Computer Vision*, 2001.

[161] D. Min and K. Sohn, "Cost aggregation and occlusion handling with wls in stereo matching," *IEEE Trans. Image Processing*, vol. 17, no. 8, pp. 1431–1442, 2008.

Curriculum Vitae

Address

Vijayaraghavan Thirumalai
Ecole Polytechnique Fédérale de Lausanne (EPFL),
Signal Processing Laboratory- LTS4,
EPFL-STI-IEL-LTS4,
Station-11, Lausanne-1015, Switzerland.

Contact Information

Email: vijayaraghavan.thirumalai@epfl.ch
Tel: 0041 21 693 2708
Fax: 0041 21 693 7600
Web: http://lts4.epfl.ch/vijay

RESEARCH INTERESTS

Signal and Image Processing; Image and Video coding; 3D coding; Sparse approximations; Multi-view geometry; Omnidirectional/spherical signal processing; Distributed signal processing and coding; Compressed sensing; Information Theory.

EDUCATION

- **Ph.D. program in Electrical Engineering** — Aug 2007 - present
  Swiss Federal Institute of Technology in Lausanne (EPFL),
  School of Electrical Engineering, Lausanne
  Research Topic: Low complexity distributed 3D scene representation

- **M.Sc. program in Communication Systems** — Oct. 2005 - May 2007
  Swiss Federal Institute of Technology in Lausanne (EPFL),
  School of Computer and Communication systems, Lausanne

- **M.Tech program in Electronics and Instrumentation** — Aug. 2003 - July 2005
  Indian Institute of Science (IISc),
  Department of Instrumentation, Bangalore-12, India
  Secured First rank in Department

- **B.E. program in Electronics and Instrumentation** — Aug. 1999 - May 2003
  Manonmaniam Sundaranar University (MSU)
  National College of Engineering, Kovilpatti, India
  Secured First rank in University

RESEARCH EXPERIENCE

Graduate Research at EPFL, Advisor: Prof. Pascal Frossard Aug. 2006 - present

- Designed a novel and robust dictionary-based distributed coding algorithm for compressing correlated images from compressed linear measurements.
- Developed a novel distributed joint representation algorithm based on a convex optimization framework for decoding correlated images from quantized linear measurements.
- Developed novel symmetric and asymmetric distributed coding algorithms for compressing correlated spherical images captured in omnidirectional camera networks.
- Developed a novel and robust distributed algorithm to estimate the correlation model directly from the set of compressed images that are given in the form of quantized linear measurements.
- Developed a novel rate allocation scheme that permits to estimate an accurate disparity information from the highly compressed planar or omnidirectional images.

Graduate Research at IISc, Advisor: Dr. Rajan Kanhirodan Jun. 2004 - May 2005

- Designed and implemented a wavelet-based codec for fast retrieval of 2D images from 3D compressed information.

AWARDS

- **EPFL Fellowship for the M.Sc. program** 2005
  for the academic year Oct. 2005 - May 2007
- **Fellowship from the Ministry of Human resources, India for M.Tech. program** 2003
  for the academic year Aug. 2003 - July 2005
- **Secured third rank in Graduate Aptitude Test in Engineering (GATE) exam** 2003
  in all India level
- **Gold Medal of Academic Excellence from Manonmaniam Sundaranar University** 2003
  for securing first rank in the University
- **Best outgoing student award** 2003
  from the Faculty of Electronics and Instrumentation
- **Certificates of Merit from the Collegiate Education Department, Tamil Nadu** 1997
  for the 100 meritorious candidate

TEACHING EXPERIENCE

- Teaching Assistant for the Graduate course *Image Communications*, EPFL, spring 2008 & 2009.
- Supervisor for one M.Sc. thesis and three intern projects.

LANGUAGES

- *English :* fluent
- *Tamil :* mother tongue
- *Hindi :* basic
- *French :* basic

Personal Publications

Journal Papers

J.1 V. Thirumalai and P. Frossard, "Correlation Estimation from Compressed Images," accepted to Journal of Visual Communication and Image Representation, 2011.

J.2 V. Thirumalai and P. Frossard, "Distributed Representation of Geometrically Correlated Images with Compressed Linear Measurements," revised version under review in IEEE Transactions on Image Processing, 2011.

J.3 V. Thirumalai, I. Tosic and P. Frossard, "Symmetric Distributed Coding of Stereo Omnidirectional Images," Signal Processing: Image Communication, vol. 23(5), pp. 379-390, 2008.

J.4 V. Thirumalai, and R. Kanhirodan, "Image Coding of 3D Volume Using Wavelet Transform for Fast Retrieval of 2D Images," IEE Proc. Vision, Image and Signal Processing, vol. 153(4), pp. 507-511, 2006.

Conference Papers

C.1 V. Thirumalai and P. Frossard, "Image Reconstruction from Compressed Linear Measurements with Side Information," Proc. IEEE International Conference on Image Processing, 2011.

C.2 V. Thirumalai and P. Frossard, "Dense Disparity Estimation from Linear Measurements," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2011.

C.3 V. Thirumalai and P. Frossard, "Joint reconstruction of correlated images from compressed linear measurements," Proc. European Signal Processing Conference, 2010.

C.4 V. Thirumalai and P. Frossard, "Motion Estimation from Compressed Linear Measurements," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2010.

C.5 V. Thirumalai and P. Frossard, "Bit Rate Allocation for Disparity Estimation from Compressed Images," Proc. Picture Coding Symposium, 2009.

C.6 V. Thirumalai, I. Tosic and P. Frossard, "Balanced Distributed Coding of Omnidirectional Images," Proc. Visual Communication and Image Processing, 2008.

C.7 V. Thirumalai, I. Tosic and P. Frossard, "Distributed coding of multiresolution omnidirectional images," Proc. IEEE International Conference on Image Processing, 2007.

C.8 V. Thirumalai and R. Kanhirodan, "Wavelet Transform Codec for Fast Retrieval of Slices of 3D objects," Proc. IEEE International Conference on Instrumentation and Measurement Technology, 2005.

Technical Reports

T.1 V. Thirumalai and P. Frossard, "Correlation Estimation from Compressed Images," arXiv:1107.4667v1, 2011.

T.2 V. Thirumalai and P. Frossard, "Distributed Representation of Geometrically Correlated Images with Compressed Linear Measurements," TR-LTS-2010-005, 2010.

T.3 V. Thirumalai, I. Tosic and P. Frossard, "Symmetric Distributed Coding of Stereo Omnidirectional Images," TR-ITS-2008-013, 2008.