On Unifying Sparsity and Geometry for Image-Based 3D Scene Representation

THÈSE N^{O} 4380 (2009)

PRÉSENTÉE LE 1ER MAI 2009
À LA FACULTÉ ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE TRAITEMENT DES SIGNAUX 4
PROGRAMME DOCTORAL EN INFORMATIQUE, COMMUNICATIONS ET INFORMATION

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Ivana TOŠIĆ

acceptée sur proposition du jury:

Prof. P. Fua, président du jury
Prof. P. Frossard, directeur de thèse
Dr P. L. Dragotti, rapporteur
Prof. A. Ortega, rapporteur
Prof. P. Vandergheynst, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE
Suisse
2009

Mojoj sestri, najvećoj duši na ovom svetu.

Aknowledgments

The work presented in this thesis would have never been possible without the support of many people, both in my professional and personal life. The biggest thanks I owe to my thesis advisor, Prof. Pascal Frossard, who has supported my passion for research and my ideas during all these years. He always had time for stimulating scientific discussions that helped me to find my research path and, more importantly, to be happy with the path I have chosen. I am immensely happy and grateful that I got a chance to have Pascal as my advisor, and to share his knowledge, thoughts, and the joy of doing research. I also greatly appreciate his dedication in reading and correcting this manuscript.

I would like to thank my committee members, Prof. Pierre Vandergheynst, Prof. Antonio Ortega, Prof. Pier Luigi Dragotti, and Prof. Pascal Fua, for their valuable opinions and advices on my thesis and suggestions for improving the final version. I am also grateful for their interesting and promising ideas for future research. Many thanks to Nikos, Tamara, Mirjana, Meri, Yannick, Radan and Jack for carefully reading and correcting the final version of my thesis manuscript.

A big thanks goes to all the members of LTS4, known as "the greatest lab in the world". Many thanks to my great colleagues Dan, Ivana, Jean-Paul, Zorana, Tamara, Zafer, Vijay, Eirina, Efi, Elif, Luigi, Marija, Nikos, Jacob, Jari and Yannick. They have all contributed to make the life in the lab more fun, more interesting and more complete. I would especially like to thank Ivana Radulović, for being the driving "fun" force of the lab, for her big heart, and for our unforgettable conference trips. Special thanks also goes to Zorana, Tamara and Marija, who made me feel like home during the long working hours. I would like also to thank Vijay and Tamara for working with me on joint projects, which brought many fruitful discussions and results. I would like to thank my students for their help in my research.

I thank my office mate Meri, who has always been a wonderful friend and colleague, and responsible for the pleasant atmosphere in our office. Many thanks to all the members of the Signal Processing Institute (ITS), for numerous stimulating discussions and ideas, and all the fun we had during my stay in ITS. I would like to thank Prof. Murat Kunt, the founder of ITS, for creating this unique signal processing research environment, and for giving me a chance to be a part of it. Special thanks goes to Marianne and Rosie for their enormous help with all administrative issues. I would like to thank all the members of the Redwood Center for Theoretical Neuroscience in UC Berkeley, and its director Prof. Bruno Olshausen, for being such wonderful hosts during my stay there. I am grateful to all of them for the exchange of exciting ideas that opened some new research perspectives for me. Special thanks goes to Gianluca, for his enormous help with finding my research path in Berkeley, and for being a wonderful and supportive friend throughout my stay.

Many thanks, with all my heart, goes to my serbian friends in Lausanne: Ana, Danica, Mirjana, Zorana, Ivana A., Jugoslava, Tamara, Jelena, Marija, for all the wonderful and cheerful moments we have shared during all these years. My life in Lausanne would have never been complete without them. Many thanks to all other serbian friends in Lausanne, who are not mentioned here, but they know how much they mean to me. I would also like to thank Cane for all the happy days we spent together. Thanks a lot to my serbian friends in Serbia, Jelena Velev, Jelena Nikolić and

Divna, who, although far away, have always been close to me with their hearts. Many thanks to my uncle Dragan and aunt Milica for being my small family in Switzerland.

A huge thanks goes to Jack, for his endless love and support during the last, hardest year of my PhD studies. I would also like to thank him for his valuable help with my research and for reading and correcting my thesis.

Finally, my deepest gratitude I owe to my family, my sister Lidija, my mom Liljana, and my dad Ljubinko, for always being there for me. Their support and love have made me what I am, and helped me through the most difficult periods of my life. I would also like to thank my brother-in-law Radan and my small nephew Marko for making the atmosphere always happy and cheerful.

Abstract

Demand has emerged for next generation visual technologies that go beyond conventional 2D imaging. Such technologies should capture and communicate all perceptually relevant three-dimensional information about an environment to a distant observer, providing a satisfying, immersive experience. Camera networks offer a low cost solution to the acquisition of 3D visual information, by capturing *multi-view* images from different viewpoints. However, the camera's representation of the data is not ideal for common tasks such as data compression or 3D scene analysis, as it does not make the 3D scene geometry explicit. Image-based scene representations fundamentally require a multi-view image model that facilitates extraction of underlying geometrical relationships between the cameras and scene components. Developing new, efficient multi-view image models is thus one of the major challenges in image-based 3D scene representation methods.

This dissertation focuses on defining and exploiting a new method for multi-view image representation, from which the 3D geometry information is easily extractable, and which is additionally highly compressible. The method is based on sparse image representation using an overcomplete dictionary of geometric features, where a single image is represented as a linear combination of few fundamental image structure features (edges for example). We construct the dictionary by applying a unitary operator to an analytic function, which introduces a composition of geometric transforms (translations, rotation and anisotropic scaling) to that function. The advantage of this approach is that the features across multiple views can be related with a single composition of transforms. We then establish a connection between image components and scene geometry by defining the transforms that satisfy the multi-view geometry constraint, and obtain a new geometric multi-view correlation model.

We first address the construction of dictionaries for images acquired by omnidirectional cameras, which are particulary convenient for scene representation due to their wide field of view. Since most omnidirectional images can be uniquely mapped to spherical images, we form a dictionary by applying motions on the sphere, rotations, and anisotropic scaling to a function that lives on the sphere. We have used this dictionary and a sparse approximation algorithm, Matching Pursuit, for compression of omnidirectional images, and additionally for coding 3D objects represented as spherical signals. Both methods offer better rate-distortion performance than state of the art schemes at low bit rates.

The novel multi-view representation method and the dictionary on the sphere are then exploited for the design of a distributed coding method for multi-view omnidirectional images. In a distributed scenario, cameras compress acquired images without communicating with each other. Using a reliable model of correlation between views, distributed coding can achieve higher compression ratios than independent compression of each image. However, the lack of a proper model has been an obstacle for distributed coding in camera networks for many years. We propose to use our geometric correlation model for distributed multi-view image coding with side information. The encoder employs a coset coding strategy, developed by dictionary partitioning based on atom shape similarity and multi-view geometry constraints. Our method results in significant rate savings compared to independent coding. An additional contribution of the proposed correlation model is that it gives information about the scene geometry, leading to a new camera pose estimation method using an extremely small amount of data from each camera.

Finally, we develop a method for learning stereo visual dictionaries based on the new multi-view image model. Although dictionary learning for still images has received a lot of attention recently, dictionary learning for stereo images has been investigated only sparingly. Our method maximizes the likelihood that a set of natural stereo images is efficiently represented with selected stereo dictionaries, where the multi-view geometry constraint is included in the probabilistic modeling. Experimental results demonstrate that including the geometric constraints in learning leads to stereo dictionaries that give both better distributed stereo matching and approximation properties than randomly selected dictionaries. We show that learning dictionaries for optimal scene representation based on the novel correlation model improves the camera pose estimation and that it can be beneficial for distributed coding.

Keywords: multi-view imaging, omnidirectional imaging, sparse approximations, multi-view geometry, distributed coding, dictionary learning

Version Abrégée

De nouvelles technologies ont récemment vu le jour et permettent d'aller au-delà de l'imagerie bidimensionnelle classique. Ces technologies devraient permettre de capturer et de communiquer les informations 3D essentielles d'une scène, offrant ainsi à un observateur à distance une expérience immersive. Les réseaux de caméras offrent une solution peu coûteuse pour l'acquisition d'informations visuelles 3D, en capturant plusieurs images de différents points de vue. Cependant, la représentation usuelle de l'information des caméras n'est pas idéale pour des tâches telles que la compression de données ou l'analyse de scène 3D, car elle n'exploite pas la géométrie de la scène 3D explicitement. Les représentations de scènes basées sur des images nécessitent un modèle d'images de vues multiples qui facilite à la fois l'extraction de relations géométriques entre les caméras, de même que les composants principaux d'une scène. Ainsi, développer des nouveaux modèles d'images de vues multiples est l'un des grands défis dans la résolution de méthodes de représentation de scène 3D basée sur des images.

Cette thèse porte son attention sur la définition et l'exploitation d'une nouvelle représentation d'images de vues multiples, très compressible, à partir de laquelle les informations de la géométrie 3D peuvent être extraites. La méthode est basée sur la représentation parcimonieuse d'images utilisant un dictionnaire redondant d'éléments géométriques, où une chaque image est représentée comme une combinaison linéaire de quelques éléments fondamentaux représentant la structure de l'image (les contours par exemple). Le dictionnaire est construit à partir d'une fonction analytique qui subit une composition de transformations géométriques (translations, rotations, et changement d'échelle anisotropique). L'avantage de cette approche est que les caractéristiques des différentes vues peuvent être représentées par une seule composition de transformations. Nous avons ensuite établi un lien entre les composants d'une image et la géométrie de la scène en définissant les transformations qui satisfont les contraintes géométriques de vues multiples. Nous avons proposé un nouveau modèle de corrélation géométrique pour des vues multiples.

Nous nous sommes particulièrement intéressés à des images acquises par des caméras omnidirectionnelles, qui sont particulièrement efficaces pour la représentation de scène en raison de leur large champ de vue. Comme la plupart des images omnidirectionnelles peuvent être associées à des images sphériques, nous avons construit un dictionnaire en appliquant les rotations sur la sphère, et les échelles anisotropes à une fonction définie sur la sphère. Nous avons utilisé ce dictionnaire et un algorithme d'approximation parcimonieuse, le Matching Pursuit, pour la compression d'images omnidirectionnelles, et en outre, pour le codage d'objets 3D représentés comme des signaux sphériques. Ces deux méthodes permettent d'obtenir des performances début-distorsion meilleures que l'état de l'art à bas débit.

La nouvelle méthode de représentation d'images multiples et le dictionnaire sur la sphère sont ensuite exploités pour la conception d'une méthode de codage distribué pour les vues multiples d'images omnidirectionnelles. Dans un scénario distribué, les caméras compressent les images acquises sans communiquer les unes avec les autres. En modélisant correctement la corrélation entre les points de vue, le codage distribué peut obtenir de meilleurs taux de compression qu'une compression indépendante de chaque image. Toutefois, l'absence d'un modèle précis a été jusqu'à présent un obstacle pour le codage distribué dans des réseaux de caméras. Nous proposons d'utiliser notre modèle de corrélation géométrique de vues multiples pour le codage distribué d'images. L'encodeur utilise une stratégie de codage en coset, mise au point par un partitionnement du dictionnaire basé sur la similitude de formes des atomes et les contraintes géométriques des vues multiples. Notre méthode se traduit par d'importants gains de compression par rapport au codage indépendant. Une contribution supplémentaire du modèle de corrélation proposé est qu'il donne des informations sur la géométrie de la scène, conduisant à une nouvelle méthode d'estimation de poses de caméras en utilisant seulement une petite quantité de données de chaque camera.

Enfin, nous avons développé une méthode d'apprentissage de dictionnaires stéréo sur la base du nouveau modèle d'images de vues multiples. Bien que l'apprentissage de dictionnaires pour les images ait reçu beaucoup d'attention récemment, l'apprentissage de dictionnaires pour des images stéréo n'a pas été étudié en détail. Notre méthode maximise la probabilité que des images stéréo naturelles soient efficacement représentées par un dictionnaire appris, où la contrainte géométrique de vues multiples est incluse dans la modélisation probabiliste. Les résultats expérimentaux montrent qu'en incluant les contraintes géométriques dans l'apprentissage, on obtient une meilleure correspondance entre des images de caméras distribuées et de meilleures propriétés d'approximation, qu'avec dictionnaires choisis au hasard. Nous montrons que l'apprentissage de dictionnaires pour la représentation optimale d'une scène basée sur le nouveau modèle de corrélation, améliore l'estimation des poses des caméras et peut être bénéfique pour le codage distribué.

Mots-clés: images de vues multiples, imagerie omnidirectionnelle, approximation parcimonieuse, géométrie de vues multiples, codage distribué, apprentissage de dictionnaires

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Thesis outline . . . 3
- 1.3 Summary of thesis contributions . . . 4

2 Single- and Multi-View Image Representation **7**
- 2.1 Image representation and compression . . . 7
  - 2.1.1 Transform coding . . . 7
  - 2.1.2 Sparse image approximation . . . 8
  - 2.1.3 Structured parametric dictionaries for image representation . . . 9
- 2.2 Multi-view imaging . . . 10
  - 2.2.1 Main multi-view correlation models . . . 10
  - 2.2.2 Compression of multi-view images . . . 13
- 2.3 Spherical imaging . . . 14
  - 2.3.1 Omnidirectional vision sensors . . . 14
  - 2.3.2 Spherical camera model . . . 16
  - 2.3.3 Image processing on the sphere . . . 17
  - 2.3.4 Calibration of catadioptric cameras . . . 18
- 2.4 Omnidirectional multi-camera systems . . . 19
  - 2.4.1 Epipolar geometry for paracatadioptric cameras . . . 19
  - 2.4.2 Estimation of extrinsic parameters . . . 20
- 2.5 Conclusion . . . 20

3 Sparse Approximations on the 2D Sphere **23**
- 3.1 Introduction . . . 23
- 3.2 Redundant dictionary on the 2-D sphere . . . 24
  - 3.2.1 Preliminaries . . . 24
  - 3.2.2 Generating functions . . . 25
  - 3.2.3 Motion on the sphere . . . 26
  - 3.2.4 Anisotropic refinement of atoms on the sphere . . . 26
- 3.3 Matching Pursuit algorithm for sparse approximation . . . 28
- 3.4 Omnidirectional image compression . . . 28
  - 3.4.1 Omnidirectional image coder . . . 29
  - 3.4.2 Implementation . . . 29
  - 3.4.3 Experimental compression performance evaluation . . . 31
- 3.5 3D object compression . . . 33
  - 3.5.1 3D-SMP coder . . . 35
  - 3.5.2 Implementation . . . 36
  - 3.5.3 Compression performance evaluation . . . 38
- 3.6 Conclusion . . . 40

4 Correlation in Multi-View Images 41
4.1 Introduction . . . 41
4.2 State of the art . . . 41
4.3 Multi-view geometric correlation model . . . 43
4.4 Geometric transforms in omnidirectional images . . . 45
4.4.1 Multi-view image expansions on the sphere . . . 45
4.4.2 Shape and epipolar constraints . . . 46
4.5 Disparity map estimation from atom transforms . . . 47
4.6 Evaluation of the proposed correlation model . . . 49
4.7 Conclusion . . . 53

5 Distributed Scene Coding 55
5.1 Introduction . . . 55
5.2 Distributed coding of correlated sources . . . 56
5.2.1 Slepian-Wolf theorem . . . 56
5.2.2 Wyner-Ziv theorem . . . 58
5.3 Wyner-Ziv coding with the geometric correlation model . . . 58
5.3.1 Coding with side information . . . 59
5.3.2 Coset design . . . 59
5.3.3 Central decoder . . . 61
5.3.4 Occlusion-resilient coding . . . 63
5.4 Experimental Coding Results . . . 64
5.4.1 Experimental settings . . . 64
5.4.2 Rate-distortion analysis of the Wyner-Ziv image . . . 65
5.4.3 Overall rate-distortion performance . . . 67
5.4.4 Performance of the occlusion-resilient Wyner-Ziv coder . . . 70
5.5 Related Work . . . 71
5.6 Conclusion . . . 74

6 Coarse Scene Structure Estimation 75
6.1 Introduction . . . 75
6.2 Identification of corresponding features . . . 76
6.3 Camera pose recovery . . . 77
6.3.1 Eight-point algorithm . . . 77
6.3.2 Robust estimation with RANSAC . . . 78
6.4 Depth estimation . . . 79
6.5 Experimental results . . . 79
6.5.1 Setup . . . 79
6.5.2 Disparity and depth estimation results for a synthetic 3D scene . . . 79
6.5.3 Disparity and camera pose estimation results for a natural scene . . . 80
6.5.4 Improving camera pose estimation results with RANSAC . . . 81
6.6 Conclusion . . . 84

7 Learning of Stereo Visual Dictionaries 85
7.1 Introduction . . . 85
7.2 Preliminaries . . . 86
7.2.1 Stereo image model . . . 86
7.2.2 Epipolar distance measure . . . 87
7.3 Maximum likelihood dictionary learning from natural images . . . 88
7.4 Maximum-likelihood learning of stereo dictionaries . . . 90
7.4.1 Problem formulation . . . 90
7.4.2 Distribution of coefficient vectors **a** and **b** . . . 90
7.4.3 Probability of epipolar matching of atoms . . . 92
7.4.4 The energy function . . . 93

7.5 Energy minimization by Expectation-Maximization 94
  7.5.1 Minimization with respect to coefficients 95
  7.5.2 Minimization with respect to atom scale parameters 96
7.6 Experimental results 96
  7.6.1 Experimental setup 96
  7.6.2 Scale learning results 99
  7.6.3 Verification of the learned dictionaries in the distributed setting 101
  7.6.4 Benefits of learning for distributed coding and camera pose estimation . . . 104
7.7 Related work on dictionary learning 105
7.8 Conclusion 105

8 Conclusion **107**

A Appendix **111**
A.1 Recovery of sparse correlated signals by thresholding 111
  A.1.1 Sparse signal correlation model 112
  A.1.2 Single signal thresholding 113
  A.1.3 Thresholding of correlated signals 113
  A.1.4 Verification of the results 116
    A.1.4.1 Randomly generated 1D signals 116
    A.1.4.2 Seismic 1D signals 117
A.2 Derivation of the conditional probabilities $P(b_r|a_l, \phi_l, \psi_r)$ and $P(a_l|b_r, \phi_l, \psi_r)$ for the ML based stereo learning method 118
A.3 Approximation rate of the MVMP algorithm 119

Curriculum Vitae **133**

Personal Publications **137**

List of Figures

1.1 Distributed cameras capturing the same 3D scene from different viewpoints. 2
1.2 a) Picture of the omnidirectional camera with a parabolic mirror. b) An image captured by omnidirectional camera in a). 3

2.1 Gaussian atoms on the plane: a) centered; b) translated by (20,40) pixels, rotated by $\pi/3$ and scaled by $s_x = 0.5, s_y = 0.2$. 10
2.2 Epipolar geometry for the pinhole camera model. 11
2.3 Block-based epipolar matching. Cameras are aligned such that image projections of an object in the 3D scene result in two displaced blocks. 12
2.4 Example of the perspective correlation model: a) image of the rabbit Zeka; b) perspective transform of the image in a). 12
2.5 Joint coding scheme for a camera network. Two cameras are used for illustration, but the network can include more cameras. 13
2.6 Distributed coding scheme for a camera network. Two cameras are used for illustration, but the network can include more cameras. 13
2.7 (a) Omnidirectional system with a parabolic mirror: the parabolic mirror is placed at the parabolic focus $\mathcal{F}_1$; the other focus $\mathcal{F}_2$ is at infinity [Gey01]. (b) Omnidirectional image "Zeka" captured by the parabolic catadioptric camera in (a). 15
2.8 Cross-section of mapping an omnidirectional image on the sphere [Gey01]. 15
2.9 Spherical projection model: (a) The projection of a point $\mathbf{X} \in \mathbb{R}^3$ to a point $\mathbf{x}$ on the spherical image. (b) The projection of a line $\mathbf{l}$ on the plane $\pi \in \mathbb{R}^3$ to a great circle on the spherical image [Tor05]. 17
2.10 Spherical image obtained from omnidirectional image "Zeka". 17
2.11 Epipolar geometry for the spherical camera model. 20

3.1 Stereographic Projection. A point on the 2-D sphere can be uniquely mapped on the plane tangent to the North Pole. 25
3.2 Anisotropic atoms: a) on the North pole ($\tau = 0$, $\nu = 0$), $\psi = 0$, $\alpha = 4$, $\beta = 4$; b) $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{2}$, $\psi = 0$, $\alpha = 4$, $\beta = 4$; c) $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{2}$, $\psi = \frac{\pi}{4}$, $\alpha = 8$, $\beta = 2$; d) Low frequency atom: $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{4}$, $\psi = \frac{\pi}{4}$, $\alpha = 4$, $\beta = 4$. 27
3.3 Omni-SMP coding scheme. 29
3.4 Original images, resolution $2B = 256$: (a) spherical image Room; (b) unfolded spherical image Room; (c) omnidirectional image Lab mapped to a spherical image; (d) unfolded omnidirectional image Lab. 31
3.5 Rate-distortion performance for the (a) Room image, (b) Lab image. 32
3.6 Decoded Room image: (a) Omni-SMP coder at 0.057 bpp, PSNR=29.43 dB; (b) JPEG coder at 0.057 bpp, PSNR=27.56 dB; (c) Omni-SMP coder at 0.088 bpp, PSNR= 31.06 dB; (b) JPEG coder at 0.088 bpp, PSNR=30.25 dB. 33
3.7 Decoded Lab image: (a) Omni-SMP coder at 0.058 bpp, PSNR=30.44 dB; (b) JPEG coder at 0.058 bpp, PSNR=25.50 dB; (c) Omni-SMP coder at 0.089 bpp, PSNR=31.94 dB; (d) JPEG coder at 0.089 bpp, PSNR=30.29 dB. 33

3.8 3D-SMP encoding scheme. 35
3.9 3D-SMP decoding scheme. 35
3.10 Generating the connectivity matrix: a) Sphere connectivity; b) Connectivity on the Venus model. 36
3.11 Anisotropic atoms: a) on the North pole $\tau = 0$, $\nu = 0$, $\psi = 0$, $\alpha = 8$, $\beta = 8$; b) $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{2}$, $\psi = 0$, $\alpha = 8$, $\beta = 8$; c) $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{2}$, $\psi = \frac{\pi}{4}$, $\alpha = 16$, $\beta = 4$; d) Low frequency atom: $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{4}$, $\psi = \frac{\pi}{4}$, $\alpha = 8$, $\beta = 8$. 37
3.12 a) Original Venus model; b) Interpolated Venus model at resolution $2B = 128$, PSNR=65.7983dB, $L^2 = 5.1296 \cdot 10^{-4}$; c) Interpolated Venus model at resolution $2B = 256$, PSNR=70.1930dB, $L^2 = 3.09278 \cdot 10^{-4}$; d) Original Rabbit model; e) Interpolated Rabbit model on three spheres at resolution $2B = 128$, PSNR=64.1790dB, $L^2 = 6.18091 \cdot 10^{-4}$. 37
3.13 Venus reconstructed after decoding (resolution $2B = 256$): a) 100 coefficients (0.58 KB); b) 200 coefficients (1.03 KB); c) 300 coefficients (1.5 KB); d) 400 coefficients (1.94 KB); e) Input model. 38
3.14 Rabbit reconstructed after decoding (three spheres, resolution $2B = 128$): a) 100 coefficients (0.64 KB); b) 200 coefficients (1.05 KB); c) 300 coefficients (1.45 KB); d) 400 coefficients (1.84 KB); e) Input model. 38
3.15 Rate-distortion performance for the Venus model (resolution $2B = 256$) a) PSNR, b) L^2 error. 39
3.16 Rate-distortion performance for the Rabbit model (resolution $2B = 128$) a) PSNR, b) L^2 error. 39
3.17 Venus with a filesize of 1287B: a) MP; b) PGC. 40

4.1 Example of atom transforms in approximation of two views of a 3D scene: a) a simple 3D scene b) first column: two captured spherical images of the scene; second column: sparse approximations of the two views with one Gaussian atom per view. The atom in the approximation of the second view is related to the atom in the first view by the following transform: $\tau_2 - \tau_1 = \pi/128$, $\nu_2 - \nu_1 = -12\pi/128$, $\psi_2 - \psi_1 = -\pi/16$, $\alpha_2/\alpha_1 = 1$, $\beta_2/\beta_1 = 0.8$. 45
4.2 Selection of positions of atoms that satisfy epipolar constraints. 47
4.3 Example of the distance $d(\mathbf{z}_1, \mathcal{C}_{\mathbf{z}_2})$ that is equal to the Euclidean distance of a point to the epipolar line, and denoted as d. 49
4.4 Shapes scene, original images of resolution $2B = 64$: a) view 1, b) view 2. 49
4.5 Approximated images of the Shapes scene: a) view 1 with six atoms, b) view 2 with six atoms. 50
4.6 Lab Scene, spherical original images of resolution $2B = 128$: a) view 1; b) view 2. . 51
4.7 Lab Scene, unfolded original images of resolution $2B = 128$: a) view 1; b) view 2. . 51
4.8 Lab Scene: Approximated images: a) view 1 with eight atoms, b) view 2 with eight atoms. 51
4.9 Two views from the "Mede" database. 53
4.10 Scatter plot of the epipolar distance and the inner product for atom correspondences, for 10 randomly selected image pairs from the "Mede" database. Circles denote the true correspondences and the dots denote the atom pairs that do not represent correspondences. 54
4.11 "Mede" database: a) histogram of the epipolar distance for true correspondences, b) histogram of the inner product for true correspondences. 54

5.1 Distributed coding of 3D scenes. Multiple correlated views $\{y_i\}$ of the scene are encoded independently, and decoded jointly by the central decoder. 55
5.2 Distributed source coding block scheme. 56
5.3 Achievable rate region for the Slepian-Wolf problem. 57
5.4 Graph-coloring interpretation of a corner point of the Slepian-Wolf region [Cov91]. 57
5.5 Wyner-Ziv coding scheme. 58

5.6 Block diagram for the Wyner-Ziv codec. 59
5.7 Samples of atoms in the same Shape coset. 60
5.8 Illustration of the epipolar coset design. The atoms with positions marked with different shapes are put in different epipolar cosets since they belong to the same epipolar line. 60
5.9 (a) Reference image; (b) Wyner Ziv image with the disparity map displayed by the disparity vectors. For better visibility, we show the disparity map that is down-sampled by four. The direction and length of the vectors show how pixels from the Wyner-Ziv image map to the pixels on the reference image. 62
5.10 Occlusion-resilient Wyner-Ziv coder. 64
5.11 Original Room images of resolution $2B = 128$. 65
5.12 Original Lab images. The natural omnidirectional images partially cover the sphere due to the boundaries of the mirror in the omnidirectional camera. The images are cropped to focus on the captured part of the scene. 65
5.13 Rate-distortion performance for the Room image set. 66
5.14 DSC results for the Room images: (a) decoded reference image $\hat{y}_1$ (PSNR=30.95dB); (b) transformed reference image y_{tr}; (c) decoded Wyner-Ziv image $\hat{y}_2$ at 0.0534bpp; (d) decoded second image $\hat{y}_2^{MP}$ when encoded with MP at 0.0534bpp. 67
5.15 Rate-distortion performance for the Lab image set (VQ Position cosets). Image y_2 is used as the Wyner-Ziv image and image y_1 as the reference image. 68
5.16 DSC results for the Lab images: (a) decoded reference image $\hat{y}_1$ (PSNR=29.4dB); (b) transformed reference image y_{tr}; (c) decoded Wyner-Ziv image $\hat{y}_2$ at 0.035 bpp; (d) decoded second image $\hat{y}_2^{MP}$ when encoded with MP at 0.035 bpp. 68
5.17 Comparison of rate-distortion performance for distributed coding and joint encoding (VQ coset design). 69
5.18 Influence of the camera distance on the DSC performance for the Lab image set: (a) Comparison of the rate-distortion performance for the reference camera distance $cd = 5$ and $cd = 10$. (b) Comparison of rate-distortion performance for distributed coding and joint encoding for $cd = 5$ and $cd = 10$. 69
5.19 DSC results for the Lab images: (a) decoded reference image $\hat{y}_3$ (PSNR=31.82dB); (b) transformed reference image y_{tr}; (c) decoded Wyner-Ziv image $\hat{y}_2$ at 0.035 bpp; (d) decoded second image $\hat{y}_2^{MP}$ when encoded with MP at 0.035 bpp. 70
5.20 Room image set: (e) $1 - |e_1| = 1 - |\hat{y}_1 - y_2|$, where e_1 is the difference between the decoded reference image and the original Wyner-Ziv image (white pixel denotes no error); (f) $1 - |e_2| = 1 - |\hat{y}_2 - y_2|$, where e_2 is the residue after DSC decoding. . . . 71
5.21 Lab image set ($cd = 5$): (a) $1 - |e_1| = 1 - |\hat{y}_1 - y_2|$, where e_1 is the residue without transform compensation (white pixel denotes no error); (b) $1 - |e_2| = 1 - |\hat{y}_2 - y_2|$, where e_2 is the residue after DSC decoding. 71
5.22 Lab image set ($cd = 10$): (a) $1 - |e_3| = 1 - |\hat{y}_3 - y_2|$, where e_3 is the residue without transform compensation (white pixel denotes no error); (b) $1 - |e_2| = 1 - |\hat{y}_2 - y_2|$, where e_2 is the residue after DSC decoding. 71
5.23 (a) Laplacian distribution fitted to the histogram of the residue e_3 for the Lab scene with $cd = 10$ (b) Laplacian distribution fitted to the histogram of the residue e_2 with $cd = 10$ (fitting is performed with the Matlab statistics toolbox, using the maximum likelihood estimator). 72
5.24 Rate-distortion performance of the occlusion-resilient Wyner-Ziv encoder for the image y_2. 72
5.25 Room scene: Visual comparison at very low bit rate of the decoded image y_2 between the (a) occlusion-resilient Wyner-Ziv coding at $0.0527bpp$ and $PSNR = 26.47dB$, and (b) Matching Pursuit coding at $0.0519bpp$ and $PSNR = 25.14dB$. 73
5.26 Lab scene: Visual comparison at very low bit rate of the decoded image y_2 between the (a) occlusion-resilient Wyner-Ziv coding at $0.0431bpp$ and $PSNR = 26.76dB$, and (b) Matching Pursuit coding at $0.0436bpp$ and $PSNR = 25.92dB$. 73

6.1 Example of an erroneous least squares linear fit due to a gross outlier [Fis81]. a) Set of data points representing measurements of a linear model. Point 7 is a gross outlier. b) Least square fitting using all points is given with a dashed line, while the ideal model is given with a solid line, showing the big error in the model estimation. 78
6.2 Original synthetic images: a) y_1, b) y_2. Approximated synthetic images: c) $\tilde{y}_1$, with 3 atoms, d) $\tilde{y}_2$ with 6 atoms. 80
6.3 a) Estimation of y_2 from y_1, using the disparity map: $\hat{y}_2$; b) $y_1 - y_2$; c) $\hat{y}_2 - y_2$. . . 80
6.4 Lab images. 81
6.5 Unwrapped Lab images (128 × 128): a) y_3; b) y_1; c) y_2. 82
6.6 $\hat{y}_2$: reconstructed y_2 using the disparity map from y_3 and y_1. 82
6.7 Differences between: a) y_3 and y_2, energy: 89.2; b) y_1 and y_2, energy: 77.0; c) $\hat{y}_2$ and y_2, energy: 49.7. 82

7.1 Epipolar geometry between stereo atoms. 88
7.2 Two views from the "Mede" database, no plants. 97
7.3 Two views from the "Mede" database, with plants. 98
7.4 Example of a pair of stereo patches and their MVMP selected atoms: a) left patch, b)-d) the first three atoms in the MVMP decomposition of the left patch; e) right patch, f)-h) the first three atoms in the MVMP decomposition of the right patch. . 98
7.5 Initial and learned scales of the atoms for the left and the right dictionaries. All atoms are on the North pole. 100
7.6 Distributed setting for the correspondence matching from sparse representations of stereo images. 101
7.7 Performance of the learned dictionaries in a distributed setting: a) Cumulative correspondence number (CCN) curves for initial dictionary and learned dictionaries, for $\rho = 0, 1, 3, 5$; b) MP energy decay for y_L and y_R. 102
7.8 Performance of the learned dictionaries in a distributed setting, with averaging over random initial dictionaries: a) Average Cumulative correspondence number (CCN) curve for 100 random initial dictionaries and CCN curves for learned dictionaries, for $\rho = 0, 1, 3, 5$; b) MP energy decay for y_L and y_R. 103

A.1 A set of sensors that observe the same event and record correlated observations. . 111
A.2 Number of false negatives versus the sparsity m, for different values of the maximal scaling parameters. 116
A.3 Number of false negatives versus the the sparsity m for different SNR, the maximum scale is set constant $s = 1.5$. 117
A.4 Seismic signals y_1 and y_2 captured at two neighboring locations and their respective approximations p_1 and p_2 with one Gabor atom per signal. The signals y_1 and y_2 are correlated by a shift on the x axis, which is correctly captured by shifted Gabor atoms. 118

List of Tables

4.1 Shapes scene: Semantic meaning of atoms in sparse representations of two views. . 50
4.2 Shapes Scene: Inner products for atom pairs from two views. Values of the inner products for the corresponding atom pairs are displayed in bold. 50
4.3 Shapes scene: Average symmetric epipolar atom distance for atom pairs from two views. Values of the epipolar distance for corresponding atom pairs are displayed in bold. 51
4.4 Lab scene: Semantic meaning of atoms in sparse representations of two views. . . . 52
4.5 Lab Scene: Inner products atom pairs from two views. Values of the inner products for the corresponding atom pairs are displayed in bold. 52
4.6 Lab scene: Average symmetric epipolar atom distance for atom pairs from two views. Values of the epipolar distance for corresponding atom pairs are displayed in bold. 52

6.1 Comparison of estimated depths with the ground truth for the simple scene. 80
6.2 The Euclidean error values for randomly chosen examples of the estimated and the target translation vectors. The table illustrates that for error values of 0.2 and above, the estimated translation vector cannot reliably point to the direction of the camera movement. 83
6.3 Translation estimation errors for the "Mede" database, for pose estimation using atom matching and RANSAC. 84
6.4 Translation estimation errors for the "Mede" database when images are rotated for an arbitrary angle between 10 and 30 degrees, for pose estimation using atom matching and RANSAC. 84

7.1 Initial and learned scale parameters for the left and the right image, for different values of the parameter ρ. 100
7.2 Comparison of camera pose estimation with random initial and respectively learned dictionaries, for target translation $\mathbf{T}^{\mathsf{T}} = [1 \ 0 \ 0]$. 104
7.3 Comparison of camera pose estimation with random initial and respectively learned dictionaries, for target translation $\mathbf{T}^{\mathsf{T}} = [0 \ 1 \ 0]$. 104

List of Main Notations

$\mathbb{R}$	the field of real numbers
H	Hilbert space
$SO(3)$	the rotation group in $\mathbb{R}^3$
S^2	2-D unit sphere
θ	zenith angle in the spherical coordinate system
φ	azimuth angle in the spherical coordinate system
B	bandwidth of a spherical signal
$\mathcal{D}$	overcomplete dictionary (or simply dictionary)
$\mathbf{\Phi}$	matrix of a dictionary
ϕ	atom in a dictionary
g	generating function of a structured dictionary
g_γ	atom in a structured dictionary
γ	atom index (a set of atom parameters)
Γ	a set of indexes γ of atoms in a structured dictionary
τ	motion of an atom on the sphere along θ
ν	motion of an atom on the sphere along φ
ψ	rotation of an atom on the sphere around its axis
α, β	anisotropic scaling parameters for an atom on the sphere
$T(\cdot)$	atom transform
$Q(\cdot)$	linear coordinate transform
$U(\cdot)$	unitary atom transform
$\mathbf{R}$	rotation matrix in $\mathbb{R}^3$
$\mathbf{T}$	translation vector in $\mathbb{R}^3$
d_{EA}	symmetric epipolar atom distance
$\|\cdot\|_k$	l_k norm, where $k = 0, 1, 2, \infty$. If k is omitted, $\|\cdot\|$ refers to the l_2 norm

Chapter 1

Introduction

1.1 Motivation

Despite significant advancements in video compression, camera and display technologies, visual information representation in modern technologies does not convey three-dimensional scenes in a perceptually satisfying way. In media communications, for example, even though the quality of the television broadcast has drastically increased in the last years with the development of digital and high-definition television, our perceptual systems have no difficulty differentiating between the broadcast and a real scene. Still, there are numerous situations in life when we would like to experience the three-dimensional vision of the displayed scene and have a feeling of presence in the scene. In other words, we would like our visual technology to be able to represent the 3D structure of the scene of interest, and to do that in a realistic manner. Besides video and image communications, such technology would be beneficial for designing devices that can sense and analyze their 3D surroundings using vision.

There are two main differences between the way we acquire images with cameras and the way we see the world around us. The most important difference is that most cameras are monocular, making the recovery of the scene's depth information an arduous task. However, with the decreasing costs of cameras it has become possible to capture images from different viewpoints at the same time using a network of cameras distributed in a scene, as depicted in Figure 1.1. Having multiple views of a scene allows us to reconstruct the 3D geometry by relying on the fundamental principles of stereo, or more generally, multiple view geometry [Sun03, Ana06, Sze99, Har97]. Therefore, multiple images of the scene, also called *multi-view images*, carry enough information for realistic representation of the 3D information.

We must not neglect the importance of our peripheral vision in perception of the environment surrounding us. In hemispherical cinemas, even when visual information is projected on a hemisphere and presents essentially only 2D information, we still get a much greater impression of immersion in the scene compared to just looking at a flat screen. While our eyes capture radial light rays and have a wide field of view, conventional cameras assume parallel rays and have a limited field of view due to camera construction and the planarity assumption. Therefore, future directions in realistic representation of 3D scenes would certainly benefit from alternative ways of capturing visual information, such as use of omnidirectional cameras, which record the light with a hemispherical field of view.

In order to have a realistic representation of a 3D scene, we need to address these two main differences, and develop novel visual technologies that will be similar to our sensing of the environment. However, we stand in front of many technological challenges related to the processing of visual information collected by a network of cameras and omnidirectional vision sensors.

Although multi-view geometry gives us the geometrical relations between pixels in multiple views, we do not know how to use them in the most efficient way to perform tasks such as

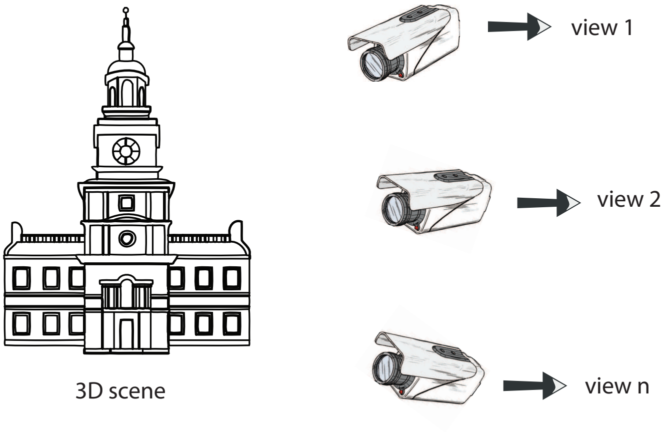
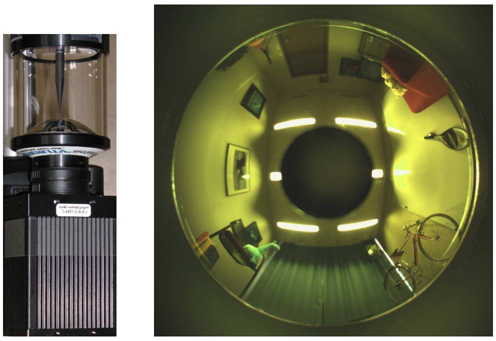


Figure 1.1: *Distributed cameras capturing the same 3D scene from different viewpoints.*

compression of acquired images, which is important in mutli-view image transmission over existing communication networks. Multi-view compression is currently a topic of intensive research in many applications, such as Free-viewpoint Television (FTV) [Fuj04] and 3D television 3DTV [Smo07, Kub07, Smo06, Mat04]. As the number of cameras becomes large, the required bandwidth for transmission of all views quickly overflows the capacity of existing communication networks. To reduce the required transmission bit rate, we need to efficiently compress multi-view images by exploiting the large amount of redundancy that they inherently contain, since they capture mostly the same visual information, just under a different viewing angle. However, state of the art methods still lack efficient models for the representation of multiple views that exploits multi-view correlation, thereby removing inter-view redundancy. Moreover, conventional methods for compressing correlated sources assume communication between sources either through a network of connections between sources or through a central encoder. Communication between cameras is usually undesirable since it consumes power and bandwidth. On the other hand, a central encoder makes the framework less robust because it is a single point of failure. Therefore, an important and interesting challenge in multi-view scene representation is the compression of multi-view images in a distributed manner.

Omnidirectional vision necessitates tackling a host of signal processing problems to deal with images of a radial and spherical nature. The first research interests for omnidirectional imaging appeared with applications such as video surveillance [Bou04], autonomous robot navigation [Yag95] and telepresence. Recently, omnidirectional imaging became an interesting and increasingly popular framework for 3D scene representation, as it requires only a small number of camera sensors for capturing a 3D scene. An example of an omnidirectional camera that consists of a conventional camera and a parabolic mirror is shown in Figure 1.2(a), while an image captured by this camera is displayed in Figure 1.2(b). With a single point of projection, omnidirectional cameras record the light field in its radial form. This permits processing of the visual information without the discrepancies introduced by the erroneous Euclidean assumptions in planar imaging. Moreover, any perspective image on any designated image plane or any panoramic image can be generated from the captured omnidirectional image [Bak03, Yag95]. As we can see in Figure 1.2 b), the acquired image has a special structure: straight lines in space project to curves on the image. How to properly take advantage of this structure in compression and image processing algorithms is the subject of research in the field of omnidirectional imaging.

These problems motivated the work done in this thesis, and directed our focus to the development of a framework for efficient scene representation with images from multiple distributed cameras. In particular, we consider omnidirectional cameras, which are interesting and advantageous due to their wide field of view and the single center of projection that permits to accurately capture the scene geometry.



(a) (b)

Figure 1.2: *a) Picture of the omnidirectional camera with a parabolic mirror. b) An image captured by omnidirectional camera in a).*

1.2 Thesis outline

Although many image representation methods exist for still and time-varying images, the development of multi-view image representation methods is in its beginnings. The goal of this thesis is to develop a multi-view image representation model that can enable efficient multi-view compression and inherently carry geometry information about the structure of the 3D scene. Certainly, the most difficult task in defining such a method is to understand and model the correlation between views, since it is the source of redundancy in images. With such a model, lossy coding of multi-view images would keep the most important visual information in all images, from which we can easily extract the main 3D structure of the scene using multi-view geometry. Therefore, we seek a multi-view image representation method that is elegant, compact, and relies on multi-view geometry principles.

The first chapter of the thesis reviews the current state of the art methods for image representation and compression, in the single-view and more importantly, the multi-view case. We also give some background theory on projective geometry and the processing of omnidirectional images. We give an overview of a unifying framework that maps omnidirectional images, captured by sensors of different construction and geometry, to spherical images. The chapter concludes with the theory of multi-view geometry for spherical camera models.

Chapter 3 is a study of the representation and compression of omnidirectional images in the single-view case. We formulate the problem of omnidirectional image representation in a spherical framework, since images from most of the existing omnidirectional cameras can be mapped to spherical images. We further propose to represent spherical images as sparse decompositions over a redundant dictionary of oriented and anisotropic atoms defined on the sphere. These types of atoms are geometric in that they have the capability to represent important image structures such as edges. The sparse representation obtained with the Matching Pursuit algorithm on spherical signals is then used for efficient compression of still omnidirectional images. Other spherical signals can be also coded and compressed with our method and we show that on the example of 3D objects. We show that for low bit rates the proposed methods outperform the state of the art codecs, both numerically and visually.

The rest of the thesis focuses on the problem of multi-view image representation. Chapter 4 introduces a new multi-view correlation model based on sparse image decompositions over structured parametric dictionaries of the geometric atoms defined in Chapter 3. Due to their geometric properties, atoms that approximate the same 3D object in the scene exist in different views under different local transforms introduced by the viewpoint change. Therefore, we define a novel correlation model that relates the atoms across different views by local geometric transforms: translation, rotation and anisotropic scaling. For the case of omnidirectional images, we define the local atom transforms that satisfy the multi-view geometry constraint. Our model has multiple advantages over existing multi-view correlation models since it is able to cope with diverse transforms of image projections of 3D objects across views, and deals with the transforms locally.

With the introduced multi-view correlation model, compression in camera networks can be achieved in a distributed manner, without requiring cameras to communicate or have a central encoder. The approach is based on the principles of distributed source coding whose theoretical foundations are given by the Slepian-Wolf [Sle73] and Wyner-Ziv [Wyn76] theorems, and where the knowledge of the correlation model between sources is crucial for constructing efficient encoders. Chapter 5 describes a novel distributed coding scheme based on the multi-view correlation model presented in Chapter 4. The detailed description of the encoder and decoder designs is given, and a method to deal with occlusions in the scene is proposed. We show that at low rates, when the image structure is encoded, the proposed distributed coding method significantly outperforms independent coding, and performs close to the joint encoding. This shows that alleviating the need for inter-camera communication induces only a negligible loss in compression performance.

We then show in Chapter 6 that the proposed correlation model can be used for coarse estimation of the 3D geometry of the scene, like depth and camera pose estimation, when very low bit rate representations of multi-view images are available. Namely, the geometric properties of atoms enable easy extraction of the scene geometry from the sparse image decompositions, using only a small number of atoms and hence a low bit rate image representation.

Chapter 7 addresses the important problem of dictionary design for efficient representation of multi-view images. A maximum-likelihood method for learning stereo dictionaries is proposed, based on the correlation model introduced previously in the thesis. The novel stereo dictionary learning method includes a multi-view geometry constraint in the probabilistic modeling in order to obtain dictionaries that have improved image approximation properties and lead to enhanced scene geometry representation. We learn dictionaries that give both better approximation performance and improved multi-view feature matching. Moreover, we discuss and demonstrate the benefits of dictionary learning for distributed coding and camera pose estimation.

Our last contribution in this thesis is a theoretical analysis of the recovery conditions for signals correlated by the proposed model using the distributed thresholding algorithm, which is an interesting alternative to other sparse approximation algorithms because of its low complexity. In order to preserve the flow within Chapters 4 to 7, we migrate this theoretical analysis to Appendix A.1.

Finally, concluding remarks and future perspectives are summarized in Chapter 8.

1.3 Summary of thesis contributions

The goal of this thesis is to provide innovative, efficient representations of 3D scenes using networks of omnidirectional vision sensors. Our main contributions are:

- the construction of a novel overcomplete dictionary of oriented and anisotropic atoms on the sphere,
- a new lossy coding method for omnidirectional images based on Matching Pursuit on the sphere,
- a new method for compressing 3D objects coding based on Matching Pursuit on the sphere.
- a new multi-view correlation model, based on sparse image decompositions with geometric dictionaries. The proposed model relates image features by diverse local transforms, which makes it highly advantageous compared to state of the art models,
- a new low rate distributed coding scheme for multi-view images. The encoder and decoder designs are proposed and the method is implemented on omnidirectional images for efficient 3D scene representation,

- a novel camera pose and coarse depth estimation method based on the proposed multi-view correlation model. The method is able to estimate the camera pose using very low bit rate approximations of multi-view omnidirectional images,
- a new method for learning stereo dictionaries based on the proposed correlation model and multi-view geometry. The novel learning method yields dictionaries optimized for both image approximation and scene geometry estimation,
- theoretical derivation of the recovery conditions for sparse signals correlated by local transforms, using distributed thresholding.

Chapter 2

Single- and Multi-View Image Representation

This chapter gives an overview of the state of the art approaches in single-view and multi-view image representation methods. The main focus is put on multi-view correlation modeling methods and their application in multi-view compression. A more detailed description of some particular models and multi-view coding applications is given in corresponding chapters. Moreover, we present the characteristics of omnidirectional imaging.

2.1 Image representation and compression

2.1.1 Transform coding

Pixel-based representations of digital images require many bits, which is very inefficient for image transmission and storage. Natural images possess high spatial redundancy within image structures such as objects or background. For decades, image compression methods have been trying to exploit this redundancy by transforming images into a different domain. The role of the transform is to represent an image using elements whose statistical dependency is greatly reduced compared to the dependency between pixels. Two most successful transforms, the Discrete Cosine Transform (DCT) and the Discrete Wavelet Transform (DWT) have been implemented in image coding standards JPEG [Ada01] (with DCT) and JPEG 2000 [Tau01] (with DWT). In both transforms, image y is represented as a linear combination of basis vectors $\{\phi_i\}, i = 1, ..., n$, i.e.,

$$y = \sum_{i=1}^{n} a_i \phi_i = \mathbf{\Phi} a, \tag{2.1}$$

where n is the dimension of the image, $\mathbf{\Phi}$ is a transform matrix that contains basis functions as columns, and a is a vector of transform coefficients. DCT and DWT are orthogonal transforms and their transform matrix satisfies $\mathbf{\Phi}^{-1} = \mathbf{\Phi}^{\mathsf{T}}$. Therefore, transform coefficients can be simply computed as $a = \mathbf{\Phi}^{\mathsf{T}} y$. The compression efficiency of transform coding lies in the fact that it results in a set of independent coefficients that can be efficiently quantized in order to save bitrate, while keeping acceptable image quality. The approximated image is then reconstructed from a quantized vector of transform coefficients $\hat{a}$ as $\hat{y} = \mathbf{\Phi}\hat{a}$. The quantizer also thresholds the insignificant coefficients to zero to reduce the bitrate. Obviously, the compression gain will be greater when the number of these insignificant coefficients is much higher than the number of significant coefficients that need to be transmitted or stored, i.e., when the vector of transform coefficients is *sparse*. However, to achieve image representations with a sparse vector of coefficients, the transform needs to efficiently de-correlate 2D components, i.e., discontinuities such as edges and curvatures. In other words, image representation methods need to include the orientation and anisotropy of the basis vectors to boost the approximation performance [Do 01, Fig02]. Understanding this important observation led to image representation methods that include directionality and anisotropy [Can99, Do 05, Vel06]. An interesting alternative to these approaches is image representation using overcomplete dictionaries [Fig06]. The only requirement on the dictionary is to be overcomplete, while its design is quite flexible and allows inclusion of directionality and anisotropy of dictionary vectors. Such overcomplete dictionaries can lead to higher sparsity of coefficient vectors, using optimization methods for nonlinear signal approximation.

2.1.2 Sparse image approximation

We begin the overview of sparse approximations by giving few fundamental definitions.

Definition 2.1.1. *A dictionary $\mathcal{D}$ in $\mathbb{R}^n$ is a set $\{\phi_k\}_{k=1}^N \subset \mathbb{R}^n$ of unit norm functions (i.e., $\|\phi_k\|_2 = 1 \ \forall\ k$).*

Definition 2.1.2. *Elements of the dictionary are called atoms.*

Definition 2.1.3. *The dictionary is complete when span$\{\phi_k\} = \mathbb{R}^n$.*

Definition 2.1.4. *When atoms are linearly dependent, the dictionary is redundant.*

If $\mathcal{D}$ is complete or redundant every image y can be represented as:

$$y = \mathbf{\Phi} a = \sum_{k=1}^{N} a_k \phi_k. \tag{2.2}$$

However, when the dictionary is over-complete, a is not unique. In order to find a compact image approximation one has to search for a sparse vector a that contains a small number of significant coefficients, while the rest of the coefficients are close or equal to zero. In other words, we say that y has a *sparse* representation in $\mathcal{D}$ if it can be represented as a linear combination of a small number of atoms in $\mathcal{D}$, up to an approximation error η, i.e.:

$$y = \mathbf{\Phi}_I c + \eta = \sum_{k \in I} a_k \phi_k + \eta, \tag{2.3}$$

where c is the vector of significant elements of a, I labels the set of atoms $\{\phi_k\}_{k \in I}$ participating in the representation, and $\mathbf{\Phi}_I$ is a matrix containing the participating atoms as columns. One is generally not interested in finding the exact representation, but rather in finding a sparse expansion with a small approximation error. In order to find the sparsest approximation of y with a bounded error norm $||\eta|| \leqslant \varepsilon$, the following optimization problem needs to be solved:

$$\min_c ||c||_0 \quad \textit{subject to} \quad ||y - \mathbf{\Phi} c||_2 \leqslant \varepsilon, \tag{2.4}$$

where $||\cdot||_0$ denotes the l_0 norm. This problem involves searching for the shortest vector of significant coefficients in a. Unfortunately, this problem is NP-hard [Nat95]. However, there exist polynomial time approximation algorithms that search for a suboptimal solution to the problem (2.4). They can be classified in two main groups: greedy algorithms (Matching Pursuit (MP) [Mal93], Orthogonal MP (OMP) [Tro04], Weak OMP [Tem00]) that iteratively select locally optimal basis vectors; and algorithms based on convex relaxation methods (e.g., Basis Pursuit [Che99], Basis Pursuit Denoising [Sta03]) that solve a slightly different problem where the l_0 norm in Eq. (2.4) is replaced by an l_1 norm. Besides pursuit algorithms, there exist other sparse approximation algorithms such as FOCUSS [Gor97], Sparse Bayesian Learning [Wip04] and Locally competitive algorithms [Roz08]. The conditions under which a given sparse approximation algorithm finds the sparsest solution for a given signal (i.e., finds the solution of Eq. (2.4)) have been established for most of the above mentioned algorithms [Tro04, Tro06, Gri08a, Fuc04]. All these recoverability conditions have a common property: they guarantee the selection of the correct sparse support when the sparsity of the signal is high enough (i.e., $||c||_0 \ll N$) or when the coherence[1] of the dictionary is low enough. However, the dictionaries in which signals have sparse representations are in general characterized by high coherence. Unfortunately, none of the above mentioned algorithms guarantees a correct sparsity recovery in highly coherent dictionaries. In these cases, it is more reliable to use the Matching Pursuit algorithm for sparse approximation since it guarantees the exponential decay of the approximation error with the increase of the number of selected sparse components. Matching Pursuit is described in detail in Chapter 3. For details on other sparse approximation algorithms we refer the reader to [Tro04, Tro06, Gri08a, Fuc04, Mal08].

2.1.3 Structured parametric dictionaries for image representation

State of the art image compression has been obtained using sparse approximations with a *structured* dictionary of 2D atoms that represent the image structure with only few basis vectors [Fig06]. Atoms in the structured dictionary are derived from a generating function that undergoes a transform combining rotation, translation and scaling. This dictionary has a couple of advantages for image representation. First, there is a flexibility in choosing a generating function, thus one can choose an anisotropic function that can efficiently represent edges and contours present in natural images. For example, a 2D function that represents a Gaussian envelope in one direction and its second derivative in the orthogonal direction has shown better approximation properties than other waveforms, such as 2D Gabor and B-spline functions [Fro00]. The anisotropic scaling of the generating function allows its refinement in order to reach even higher approximation rates. Furthermore, by introducing a high number of different rotations of the generating function, the dictionary reaches good directionality properties that are important for approximating image discontinuities.

Given a generating function g defined in a Hilbert space H, the dictionary $\mathcal{D} = \{\phi_k\} = \{g_\gamma\}_{\gamma \in \Gamma}$ is constructed by changing the atom index $\gamma \in \Gamma$ that defines parameters of the transformation applied to the generating function g. This is equivalent to applying a unitary operator $U(\gamma)$ to the generating function g, i.e., $g_\gamma = U(\gamma)g$. The atom transform is achieved by evaluating the generating function g on a linearly transformed coordinate system in H. Formally, if $\mathbf{v}$ denotes the unit vector of coordinates, and $\mathbf{u} = Q_\gamma(\mathbf{v})$ its linear transform with translation, rotation and scaling parameters given by γ, then:

$$g_\gamma(\mathbf{v}) = \frac{1}{K} g(Q_\gamma(\mathbf{v})) = \frac{1}{K} g(\mathbf{u}), \tag{2.5}$$

where K is a normalization factor. For example, consider planar images which are defined in Cartesian coordinates $\mathbf{v} = (x, y)$, and the atom given by the generating function equal to the 2-dimensional Gaussian function $g(x, y) = 1/\sqrt{\pi} \exp -(x^2 + y^2)$. This atom is shown in Figure 2.1(a), where the image center is at $(0, 0)$. By a simple linear transform of coordinates [Fig06]:

$$\hat{x} = \frac{\cos(\psi)(x - b_x) + \sin(\psi)(y - b_y)}{s_x} \tag{2.6}$$

$$\hat{y} = \frac{\cos(\psi)(y - b_y) - \sin(\psi)(x - b_x)}{s_y}, \tag{2.7}$$

we get a transformed atom $g_\gamma(x, y) = g(\hat{x}, \hat{y})$, which is translated by (b_x, b_y), rotated by ψ and scaled by s_x, s_y along the x and y axis respectively. Therefore, $\mathbf{u} = (\hat{x}, \hat{y})$. An example of a transformed Gaussian atom is shown in Figure 2.1(b).

An additional advantage of using such parametric dictionaries is that atom parameters inform the geometric properties of the image feature approximated by that atom, where image features are usually multidimensional discontinuities like edges. The size and redundancy of the dictionary is directly driven by the number of distinct geometric parameters. Structured parametric

---

[1] The coherence of a given dictionary is defined as the maximum inner product between any two different atoms in that dictionary [Tro04].

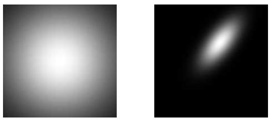


Figure 2.1: *Gaussian atoms on the plane: a) centered; b) translated by (20,40) pixels, rotated by* $\pi/3$ *and scaled by* $s_x = 0.5, s_y = 0.2$.

dictionaries are usually characterized by very high coherence (close to 1). Because of this, image compression with parametric dictionaries employs the Matching Pursuit algorithm for sparse approximation [Fig06]. The exponential decay rate of the approximation error contributes to progressive image representation, which is an important advantage of MP coders. Sparse approximations have been also successfully applied to video [Nef97, Fro00, Rah06]. Besides the good mathematical properties of sparse approximations, another motivation for this approach is that sparse image approximations with redundant dictionaries of localized and oriented features represent a plausible encoding strategy in the primary visual cortex of the human brain [Ols97].

2.2 Multi-view imaging

2.2.1 Main multi-view correlation models

The most important part of efficient 3D scene representation using a camera network is without doubt the multi-view correlation modeling. As they represent the same 3D scene, multi-view images contain a high amount of inter-view redundancy, in addition to the spatial intra-view redundancy within a single image. To compress such images, one should remove the redundancy inherent to the multi-view geometry relations that are defined by the epipolar geometry constraints.

Epipolar geometry relates multiple images of the observed environment to the 3-dimensional structure of that environment. It captures the geometric relationship between 3D points and their image projections, which enables 3-dimensional reconstruction of the scene using multiple images taken from different viewpoints. Epipolar geometry was first formulated for the pinhole camera model, leading to the well-known epipolar constraint given in the following theorem [Ma 04]:

Theorem 2.2.1. *Consider a point* p *in* $\mathbb{R}^3$*, given by its spatial coordinates* $\mathbf{X}$*, and two images of this point given by their homogeneous coordinates* $\mathbf{x}_1 = [x_1 \ \ y_1 \ \ 1]^\mathsf{T}$ *and* $\mathbf{x}_2 = [x_2 \ \ y_2 \ \ 1]^\mathsf{T}$ *in two camera frames. Let the two camera positions have relative pose* $(\mathbf{R}, \mathbf{T})$*, where* $\mathbf{R} \in SO(3)$ *is the matrix of the relative rotation and* $\mathbf{T} \in \mathbb{R}^3$ *is the vector of the relative translation between cameras. Then* $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *satisfy:*

$$\langle \mathbf{x}_2, \mathbf{T} \times \mathbf{R}\mathbf{x}_1 \rangle = 0, \qquad \text{or} \qquad \mathbf{x}_2{}^\mathsf{T} \hat{\mathbf{T}} \mathbf{R} \mathbf{x}_1 = 0. \tag{2.8}$$

The matrix $\hat{\mathbf{T}}$ is obtained by representing the cross product of $\mathbf{T}$ with $\mathbf{R}\mathbf{x}_1$ as a matrix multiplication, i.e., $\hat{\mathbf{T}}\mathbf{R}\mathbf{x}_1 = \mathbf{T} \times \mathbf{R}\mathbf{x}_1$. Given $\mathbf{T} = [t_1 \ \ t_2 \ \ t_3]^\mathsf{T}$, $\hat{\mathbf{T}}$ can be expressed as:

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

The matrix $\mathbf{E} = \hat{\mathbf{T}}\mathbf{R} \in \mathbb{R}^{3\times 3}$ is called the **essential matrix**. The epipolar geometry constraint is derived from the coplanarity of the vectors $\mathbf{x}_2$, $\mathbf{T}$ and $\mathbf{R}\mathbf{x}_1$, as shown in Figure 2.2. The difference between pixel coordinate vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, which satisfy the epipolar geometry constraint, is usually called the *disparity* between these two pixels.

Obviously, the first step that needs to be taken to remove the redundancy in multi-view images, is to define proper multi-view correlation models that include the epipolar geometry constraint. We

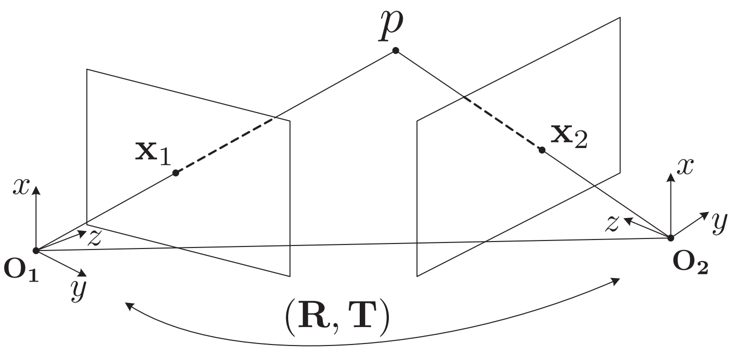


Figure 2.2: *Epipolar geometry for the pinhole camera model.*

present here the main state of the art models of multi-view correlation, while a more comprehensive overview is given in Chapter 4. The existing methods for multi-view correlation modeling can be classified into three main categories:

- pixel-based models,
- block-based models, and
- perspective and affine models.

Pixel-based methods are most frequently used in computer vision in applications such as structure from motion, motion tracking and depth estimation. Disparity and depth estimation algorithms [Kol02, Sun03] find disparity vectors that relate each pixel in one view to a pixel in another view, under epipolar and pixel intensity similarity constraints. Thus, for each pixel pair there is one disparity vector. Although pixel-wise mapping leads to good depth estimates, the use of this correlation model is usually impractical in multi-view compression due to the high number of disparity vectors that need to be encoded.

Block-based methods relate corresponding blocks in multi-view images, i.e., blocks that satisfy the epipolar constraint and are similar with respect to the mean Euclidean distance between corresponding pixel values. This is shown in Figure 2.3. Each block in one view is then related to its corresponding block in the other view using only one disparity vector. Basically, these methods try to extend the pixel-wise correlation model to a model that has smaller number of disparity vectors and hence leads to more efficient multi-view coding. This approach is analogous to hybrid video coding employed in H.264 video coding standard [Mar06a]. However, the big disadvantage of this approach is the assumption that all pixels within a block have the same epipolar matching (i.e., the same disparity), which limits the use of block-based methods to camera networks where cameras are arranged in a line and have the same orientation.

Finally, perspective and affine models are based on the assumption that one view can be represented as a perspective or affine transformation of another view. As such, these models assume a global disparity where all pixels in an image follow the same transformation model based on the epipolar geometry. An example of a perspective transform is shown in Figure 2.4. Compared to block-based methods, perspective models can deal with more diverse types of transforms and hence different camera arrangements. This approach has been successfully used in the domain of multi-view pattern recognition, where the observed pattern is assumed to be planar. However, 3D scenes are far away from planar, and global disparity methods can work only for very narrow camera field of view.

To summarize, the state of the art methods lack a multi-view representation approach that can deal with a wide variety of transforms of image features between views, and more importantly, which can deal with the local nature of those transforms.

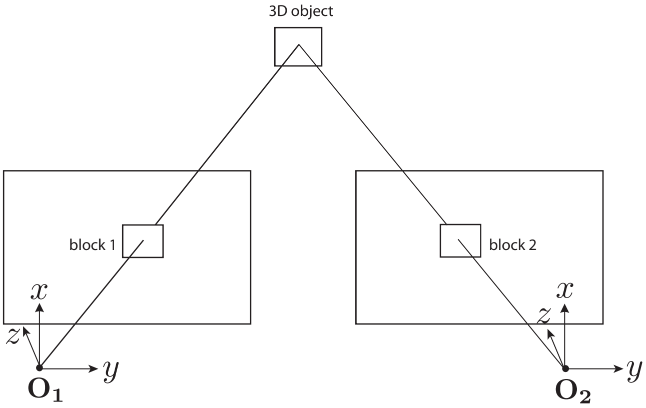


Figure 2.3: *Block-based epipolar matching. Cameras are aligned such that image projections of an object in the 3D scene result in two displaced blocks.*

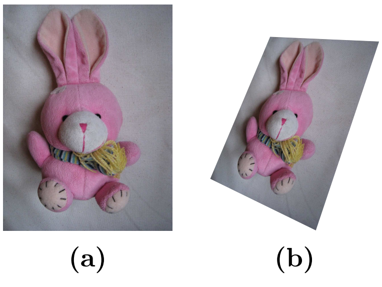

(a) (b)

Figure 2.4: *Example of the perspective correlation model: a) image of the rabbit Zeka; b) perspective transform of the image in a).*

2.2.2 Compression of multi-view images

Depending on the existence of communication links between cameras in a network, lossy coding of multi-view images can be achieved in two ways: joint and distributed coding. In the joint coding scenario the central encoder collects images or video streams from all cameras and uses a multi-view correlation model to remove the inter-view redundancy (see Figure 2.5). For example, joint coding based on the block-based multi-view correlation model has been used in the design of the multi-view coding (MVC) extension of the video coding standard H.264/AVC [Mer06, Mer07, Mar06b]. Unfortunately, due to the inefficiency of the block-based correlation model for inter-view coding, the compression gains due to the inter-view correlation are marginal (up to 10%).

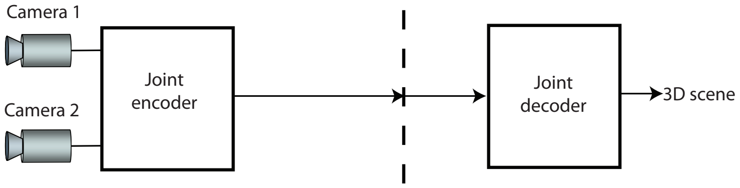


Figure 2.5: *Joint coding scheme for a camera network. Two cameras are used for illustration, but the network can include more cameras.*

Distributed source coding (DSC) permits efficient coding of correlated sources without requiring the communication of sources at the encoder side [Sle73, Wyn76]. The redundancy due to the correlation of sources is exploited at the decoder side, unlike in joint encoding where it is exploited at the encoder side. Distributed coding seems to be an attractive approach for compression in camera networks, since it removes the need for a central encoder or for inter-camera communication, as shown in Figure 2.6. However, designing DSC schemes for camera networks is much more difficult than designing joint coders of multi-view images. Efficient distributed coding relies on the knowledge of the source's correlation model at the encoder, and its performance is extremely sensitive to incorrect correlation modeling. As we have seen above, we still lack good multi-view correlation models, and this has significantly slowed down the development of novel distributed coding methods for camera networks.

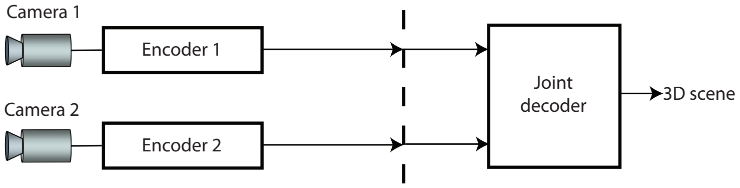


Figure 2.6: *Distributed coding scheme for a camera network. Two cameras are used for illustration, but the network can include more cameras.*

We describe here a few of the most relevant works that address the problem of DSC in camera networks, using the multi-view correlation models overviewed in Section 2.2.1. Gehrig and Dragotti proposed a method where the multi-view correlation is modeled by relating the locations of image discontinuities, thus this method uses pixel-based correlation [Geh09]. Their efficient image compression is achieved by a quadtree decomposition, assuming a geometrical image model in which 2D linear regions are separated by a 1D linear boundary. The block-based correlation model has been applied to multi-view DSC by Yeo and Ramchandran [Yeo07]. They extend the PRISM [Pur07] architecture, initially designed for distributed video coding, to multi-view compression. Finally, the affine and perspective correlation models have been used for multi-view DSC [Guo06, Oua06a]. A more extensive survey of distributed coding methods for multi-view compression is given in Section 5.5, and also in the overview paper [Gui07].

2.3 Spherical imaging

In the previous part of this chapter we gave an overview of state of the art methods for representation and compression of single-view and multi-view planar images. Planar image projections are currently the most usual way to represent visual content, giving to the observer only a windowed view of the world. Since the creation of the first paintings our minds have been strongly bound to this idea. However, planar projections have important limitations for building accurate models of 3D environments since light has a naturally radial form. The radial organization of photo-receptors in the human fovea also suggests that we should reconsider the way we acquire and sample visual information, and depart from the classical planar imaging with rectangular sampling. The fundamental object underlying dynamic vision and providing a firm mathematical foundation is the Plenoptic Function (PF) [Ade91], which simply measures the light intensity at all positions in a scene and for all directions. In static scenes, the function becomes independent of time, and it is convenient to define it as a function on the product manifold $\mathbb{R}^3 \times S^2$, where S^2 is the 2D sphere (we drop the chromaticity components for the sake of simplicity). We can consider the plenoptic function as the model for a perfect vision sensor and hence processing of visual information on the sphere becomes an interesting problem. Therefore, the rest of this chapter first describes the geometry of omnidirectional vision sensors whose output images can be uniquely mapped to spherical images. We then present image processing methods that can be applied to spherical imaging. Finally, we discuss the calibration issue and define the epipolar geometry constraint for the spherical camera model.

2.3.1 Omnidirectional vision sensors

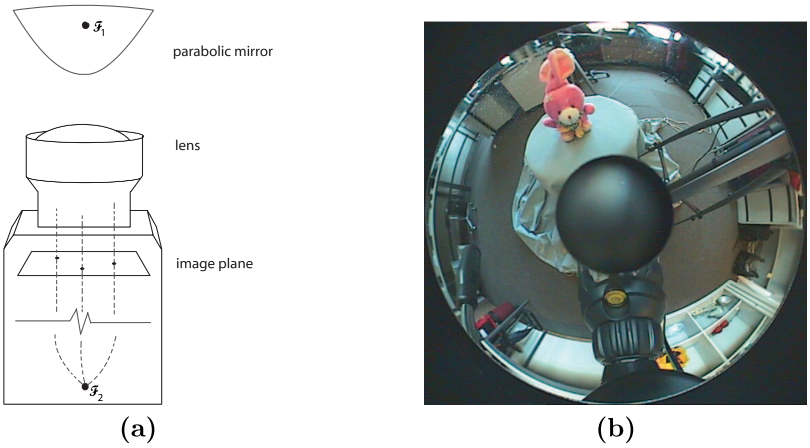
Omnidirectional vision sensors are devices that can capture a 360-degree view of the surrounding scene. According to their construction, these devices can be classified into three types [Yag99]: systems that use multiple images (i.e., image mosaics), devices that use special lenses, and catadioptric devices that employ a combination of convex mirrors and lenses.

The images acquired by traditional perspective cameras can be used to construct an omnidirectional image, either by rotating a single camera or by construction of a multi-camera system. Obtained images are then aligned and stitched together to form a 360-degree view. However, a rotating camera system is limited to capturing static scenes because of the long acquisition time for all images. It may also suffer from mechanical problems that lead to maintenance issues. On the other hand, multiple camera systems can be used for real-time applications, but they suffer from difficulties in alignment and camera calibration.

In this context, true omnidirectional sensors are interesting since each image provides a wide field of view of the scene of interest. Special lenses (e.g. fish-eye) probably represent the most popular classes of systems that can capture omnidirectional images. However, such cameras present the important disadvantage that they do not have a single center of projection, which makes omnidirectional image analysis extremely complicated. Alternatively, omnidirectional images can also be generated by catadioptric devices. These systems use a convex mirror placed above a perspective camera, where the optical axis of the lens is aligned with the mirror's axis. The catadioptric cameras with quadric mirrors are of particular interest, since they represent omnidirectional cameras with a single center of projection. Moreover, the images obtained with such cameras can be uniquely mapped on the sphere [Gey01]. We focus now on the class of catadioptric systems with quadric mirrors.

Catadioptric cameras achieve an almost hemispherical field of view with a perspective camera and a catadioptric system [Nay97a], which represents a combination of reflective (catoptric) and refractive (dioptric) elements [Hec97,Gey01]. Such a system is schematically shown in Figure 2.7(a) for the case of a parabolic mirror. Figure 2.7(b) illustrates one omnidirectional image captured by the parabolic catadioptric camera.

Catadioptric image formation has been extensively studied [Nay97b, Bak99, Gey01]. Central catadioptric systems are of special interest because they have a single effective viewpoint. This property is important not only for easier image analysis but also for multi-view 3D reconstruction.



(a) **(b)**

Figure 2.7: *(a) Omnidirectional system with a parabolic mirror: the parabolic mirror is placed at the parabolic focus* $\mathcal{F}_1$*; the other focus* $\mathcal{F}_2$ *is at infinity [Gey01]. (b) Omnidirectional image "Zeka" captured by the parabolic catadioptric camera in (a).*

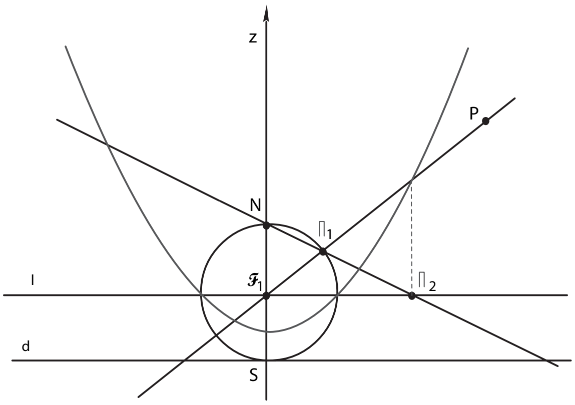


Figure 2.8: *Cross-section of mapping an omnidirectional image on the sphere [Gey01].*

Unfortunately, images captured by the perspective camera from the light reflected by the catadioptric system are not straightforward to analyze because the lines in the 3D space project onto conic sections on the image [Gey01,Svo98,Nen98]. However, a unifying model for catadioptric projective geometry of central catadioptric cameras established by Geyer and Daniilidis permits us to map omnidirectional images from such sensors to spherical images [Gey01]. According to their model, any central catadioptric projection is equivalent to a composition of two mappings on the sphere. The first mapping represents a central spherical projection, with the center of the sphere incident to the focal point of the mirror and independent of the mirror shape. The second mapping is a projection from a point on the principal axis of the sphere to the plane perpendicular to that axis. However, the position of the projection point on the axis of the sphere depends on the shape of the mirror. The model of Geyer and Daniilidis includes perspective, parabolic, hyperbolic and elliptic projections. The catadioptric projective framework permits us to derive efficient and simple scene analysis from spherical images captured by catadioptric cameras. We describe now in more detail the projective model for the parabolic mirror, where the second mapping represents the projection from the North Pole to the plane that includes the equator. In particular, the second mapping is known as the stereographic projection, and it is conformal.

We consider a cross-section of the paraboloid in a catadioptric system with a parabolic mirror. This is shown in Figure 2.8. All points on the parabola are equidistant to the focus $\mathcal{F}_1$ and the directrix d. Let l pass through $\mathcal{F}_1$ and be perpendicular to the parabolic axis. If a circle has a center $\mathcal{F}_1$ and a radius equal to the double of the focal length of the paraboloid, then the circle and parabola intersect twice the line l and the directrix is tangent to the circle. The North Pole N of the circle is the point diametrically opposite to the intersection of the circle and the directrix. Point P is projected on the circle from its center, which gives Π_1. This is equivalent to a projective representation, where the projective space (set of rays) is represented as a circle here. It can be seen that Π_2 is the stereographic projection of the point Π_1 to the line l from the North Pole N, where Π_1 is the intersection of the ray $\mathcal{F}_1 P$ and the circle. We can thus conclude that the parabolic projection of a point P yields point Π_2, which is collinear with Π_1 and N. Extending this reasoning to three dimensions, the projection by a parabolic mirror is equivalent to a projection on the sphere (Π_1) followed by a stereographic projection (Π_2). Formal proof of the equivalence between the parabolic catadioptric projection and the composite mapping through the sphere can be found in [Gey01]. A direct corollary of this equivalence is that the parabolic catadioptric projection is conformal since it represents a composition of two conformal mappings.

For the other types of mirrors in catadioptric systems, the position of the point of projection in the second mapping is a function of the eccentricity ϵ of the conic (see Theorem 1 in [Gey01]). For hyperbolic mirrors with $\epsilon > 1$ and elliptic mirrors with $0 < \epsilon < 1$, the projection point lies on the principal axis of the sphere, between the center of the sphere and the North Pole. A perspective camera can be also considered as a degenerative case of a catadioptric system with a conic of eccentricity $\epsilon = \infty$. In this case, the point of projection for the second mapping coincides with the center of the sphere.

2.3.2 Spherical camera model

We exploit the equivalence between the catadioptric projection and the composite mapping through the sphere in order to map an omnidirectional image through the inverse stereographic projection to the surface of a sphere whose center coincides with the focal point of the mirror. This leads to the definition of the Spherical camera model [Tor05], which consists of a camera center and a surface of a unit sphere whose center is the camera center.

The projection of light rays on the surface of the unit sphere is usually referred to as the **spherical image**. The spherical image is formed by a central spherical projection from a point $\mathbf{X} \in \mathbb{R}^3$ to the unit sphere with the center $\mathbf{O} \in \mathbb{R}^3$, as shown in Figure 2.9(a). Point $\mathbf{X}$ is projected into a point $\mathbf{x}$ on the unit sphere S^2, where the projection is given by the following relation:

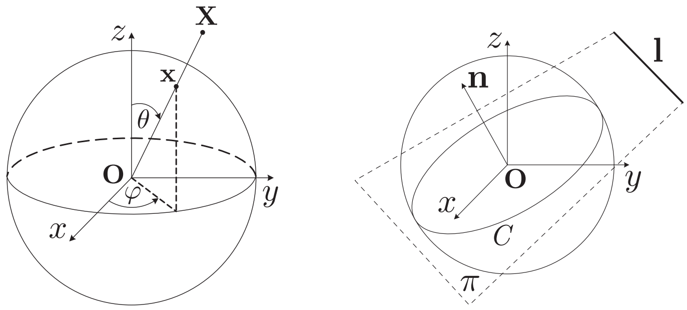
$$\mathbf{x} = \frac{1}{|\mathbf{X}|}\mathbf{X}. \tag{2.9}$$

The point $\mathbf{x}$ on the unit sphere can be expressed in spherical coordinates:

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{pmatrix}$$

where $\theta \in [0, \pi]$ is the zenith angle, and $\varphi \in [0, 2\pi)$ is the azimuth angle. The spherical image is then represented by a function defined on S^2, that is $y(\theta, \varphi) \in S^2$. Figure 2.10 shows an example of a spherical image obtained by mapping an omnidirectional image "Zeka" (Figure 2.7) to the sphere, using the projective geometry explained previously.

We now briefly discuss the consequences of mapping the three-dimensional information on the 2D sphere. The spherical projection of a line $\mathbf{l}$ in $\mathbb{R}^3$ is a great circle on the sphere S^2, as illustrated in Figure 2.9(b). This great circle, denoted C, is obtained as the intersection of the unit sphere and the plane π that passes through the line $\mathbf{l}$ and the camera center $\mathbf{O}$. Since C is completely defined by the normal vector $\mathbf{n} \in S^2$ of the plane π, there is a duality between a great circle and a point on the sphere, as pointed out by Torii et al. [Tor05]. Moreover, they have shown that there exists a duality between a line $\mathbf{l}$ on a projective plane $P^2 = \{(x, y, z)|z = 1\}$ and a point on the sphere. These two duality relations play an important role in defining the trifocal tensor for spherical cameras, which is usually used to express the three-view geometry constraints [Tor05].



(a) (b)

Figure 2.9: *Spherical projection model: (a) The projection of a point* $\mathbf{X} \in \mathbb{R}^3$ *to a point* $\mathbf{x}$ *on the spherical image. (b) The projection of a line* $\mathbf{l}$ *on the plane* $\pi \in \mathbb{R}^3$ *to a great circle on the spherical image [Tor05].*

Figure 2.10: *Spherical image obtained from omnidirectional image "Zeka".*

2.3.3 Image processing on the sphere

Because images captured by catadioptric systems can be uniquely mapped on the sphere, it becomes interesting to process the visual information directly in the spherical coordinate system. Similar to the Euclidean framework, harmonic analysis and multi-resolution decomposition represent efficient tools for processing data on the sphere. When mapped to spherical coordinates, the omnidirectional images are typically re-sampled on an equi-angular grid on the sphere:

$$\mathcal{G}_j = \{(\theta_{jp}, \varphi_{jq}) \in S^2 : \theta_{jp} = \tfrac{(2p+1)\pi}{4B_j}, \varphi_{jq} = \tfrac{q\pi}{B_j}\}, \tag{2.10}$$

$p, q \in \mathcal{N}_j \equiv \{n \in \mathbb{N} : n < 2B_j\}$ and for some range of bandwidth $B = \{B_j \in 2\mathbb{N}, j \in \mathbb{Z}\}$. These grids permit to perfectly sample any band-limited function $y \in L^2(S^2)$ of bandwidth B_j.

The omnidirectional images can then be represented as spherical signals, modeled by elements of the Hilbert space of square-integrable functions on the two-dimensional sphere $L^2(S^2, \mathrm{d}\mu)$, where $\mathrm{d}\mu(\theta, \varphi) = \mathrm{d}\cos\theta d\varphi$ is the rotation invariant Lebesgue measure on the sphere. These functions are characterized by their Fourier coefficients $\hat{y}(m, n)$, defined through the spherical harmonics expansion:

$$\hat{y}(m, n) = \int_{S^2} \mathrm{d}\mu(\theta, \varphi)\, Y^*_{m,n}(\theta, \varphi) y(\theta, \varphi),$$

where $Y^*_{m,n}$ is the complex conjugate of the spherical harmonic of order (m, n) [Dri94]. It can be noted here that this class of sampling grids is associated with a Fast Spherical Fourier Transform [Hea03].

Multi-resolution representations are particularly interesting in applications such as image analysis and image coding. The two most successful embodiments of this paradigm, the various wavelet decompositions [Mal08] and the Laplacian Pyramid (LP) [Bur83], have been extended to spherical manifolds.

The spherical continuous wavelet transform (SCWT) has been introduced by Antoine and Vandergheynst [Ant98]. It is based on affine transformations on the sphere, namely: rotations, defined by the element ρ of the group $SO(3)$, and dilations D_α, parameterized by the scale $\alpha \in \mathbb{R}_+^*$ [Ant99]. Interestingly, it can be proved that any admissible 2-D wavelet in $\mathbb{R}^2$ yields an admissible spherical wavelet by inverse stereographic projection.

To process images given on discrete spherical grids, the SCWT can be replaced by *frames* of spherical wavelets [Bog04, Bog05]. One of the most appealing features of frames is their ability to expand any spherical map into a finite multi-resolution hierarchy of wavelet coefficients. The scales are discretized in a monotonic way, and the positions are taken in an equi-angular grid $\mathcal{G}_j$, as described previously.

Another simple way of dealing with discrete spherical data in a multi-resolution fashion is to extend the Laplacian Pyramid [Bur83] to spherical coordinates. This can be simply done considering the recent advances in harmonic analysis on S^2, in particular the work of Driscol and Healy [Dri94]. Indeed, based on the notations introduced earlier, one can define a series of downsampled grids $\mathcal{G}_j$ with $B_j = 2^{-j} B_0$. A series of multi-resolution spherical images S_j can be generated by recursively applying convolution with a low-pass filter h and downsampling. The filter h could, for example, take the form of an axisymmetric low-pass filter defined by its Fourier coefficients:

$$\hat{h}_{\sigma_0}(m) = e^{-\sigma_0^2 m^2}. \tag{2.11}$$

Suppose, then, that the original data y_0 is bandlimited (i.e., $\hat{y_0}(m,n) = 0, \forall m > B_0$) and sampled on $\mathcal{G}_0$. The bandwidth parameter σ_0 is chosen so that the filter is numerically close to a perfect half-band filter $\hat{H}_{\sigma_0}(m) = 0, \forall m > B_0/2$. The low-pass filtered data is then downsampled on the nested sub-grid $\mathcal{G}_1$, which gives the low-pass channel of the pyramid y_1. The high-pass channel of the pyramid is computed as usual, that is, by first upsampling y_1 on the finer grid $\mathcal{G}_0$, low-pass filtering it with H_{σ_0} and taking the difference with y_0. Coarser resolutions are computed by iterating this algorithm on the low-pass channel y_l and scaling the filter bandwidth accordingly (i.e., $\sigma_l = 2^l \sigma_0$).

Because of, for example, their multi-resolution and local nature, frames of spherical wavelets or the Spherical Laplacian Pyramid can be advantageous over the spherical harmonics. This is also thanks to their covariance under rigid rotations [Mak03].

2.3.4 Calibration of catadioptric cameras

We have assumed up to this point that the camera parameters are perfectly known and that the cameras are calibrated. However, camera calibration is generally not given in practical systems; one usually has to estimate the intrinsic parameters of the projection, which enable mapping the pixels on the image to corresponding light rays in the space. For catadioptric cameras these parameters include the focal length of the catadioptric system (the combined focal lengths of the mirror and the camera); the image center; and the aspect ratio and skew factor of the imaging sensor. For the catadioptric camera with a parabolic mirror, the boundary of the mirror is projected to a circle on the image, which can be exploited for calibration. Namely, one can fit a circle to the image of the mirror's boundary and calculate the image center as the circle's center. Then, knowing the field of view of the catadioptric camera with respect to the zenith angle, the focal length can be determined by simple calculations. This simple strategy for calibration is very advantageous since it does not require the capturing and analysis of the calibration pattern. It can be performed on any image where the mirror boundary is sufficiently visible.

Another approach for calibration of catadioptric cameras uses the image projections of lines to estimate the intrinsic camera parameters [Gey01, Gey02]. Whereas in perspective cameras calibration from lines is not possible without any metric information, Geyer and Daniilidis showed that it is possible to calibrate central catadioptric cameras only from a set of line images. They first considered the parabolic case and assumed that the aspect ratio is one and the skew is zero, so the total number of unknown intrinsic parameters is three (focal length and two coordinates of the image center). In the parabolic case, a line in the space projects into a circle on the image plane. Since a circle is defined by three points, each line gives a set of three constraints but also introduces two unknowns that specify the orientation of the plane containing the line. Therefore, each line contributes with one additional constraint, leading us to the conclusion that three lines are sufficient to perform calibration. In the hyperbolic case, there is one additional intrinsic parameter to be estimated, the eccentricity, so in total there are four unknowns. The line projects into a conic, which is defined by five points and thus gives five constraints. Therefore, for the calibration of a hyperbolic catadioptric camera, only two line images suffice. Based on this reasoning, Geyer and Daniilidis have proposed an algorithm for calibration of parabolic catadioptric cameras from three line images. The algorithm details can be found in [Gey02] and [Tos09].

Besides calibration using the projection of lines, camera calibration can also be performed by the projection of spheres [Yin04], which actually offers improved robustness. Researchers also investigated the calibration of cameras with different types of mirrors, such as hyperbolic and elliptical [Bar05]. For example, Scaramuzza et al. proposed a calibration method that uses a generalized parametric model of the single viewpoint omnidirectional sensor and can be applied to any type of mirror in the catadioptric system [Sca06]. The calibration requires two or more images of the planar pattern at different orientations. Calibration of non-central catadioptric cameras was proposed by Mičušík and Pajdla [Mic04], based on epipolar correspondence matching from two catadioptric images. Epipolar geometry was also exploited for calibration of paracatadioptric cameras [Kan00].

2.4 Omnidirectional multi-camera systems

With the development of panoramic cameras, epipolar or multi-view geometry has been recently formalized for general camera models [Stu05]. However, we focus here on the case of calibrated paracatadioptric cameras, where the omnidirectional image can be uniquely mapped through the inverse stereographic projection to the surface of the sphere whose center coincides with the focal point of the mirror. Two- and three- view geometry for spherical cameras has been introduced by Torii et al. [Tor05], and we overview the two-view case in this section. We refer the interested reader to [Tor05] for the three-view geometry framework.

2.4.1 Epipolar geometry for paracatadioptric cameras

Epipolar geometry can be used to describe the geometrical constraints in systems with two spherical cameras. Let $\mathbf{O}_1$ and $\mathbf{O}_2$ be the centers of two spherical cameras, and let the world coordinate frame be placed at the center of the first camera, i.e., $\mathbf{O}_1 = [0 \ \ 0 \ \ 0]^\mathsf{T}$. A point $p \in \mathbb{R}^3$ is projected to unit spheres corresponding to the cameras, giving projection points $\mathbf{x}_1, \mathbf{x}_2 \in S^2$, as illustrated in Figure 2.11. Let $\mathbf{X}_1$ be the coordinates of the point p in the coordinate system of camera centered at $\mathbf{O}_1$. The spherical projection of a point p to the camera centered at $\mathbf{O}_1$ is given as:

$$\lambda_1 \mathbf{x}_1 = \mathbf{X}_1, \quad \lambda_1 \in \mathbb{R}. \tag{2.12}$$

If we further denote the transform of the coordinate system between two spherical cameras with $\mathbf{R}$ and $\mathbf{T}$, where $\mathbf{R}$ denotes the relative rotation and $\mathbf{T}$ denotes the relative translation, the coordinates of the point p can be expressed in the coordinate system of the camera at $\mathbf{O}_2$ as $\mathbf{X}_2 = \mathbf{R}\mathbf{X}_1 + \mathbf{T}$. The projection of the point p to the camera centered at $\mathbf{O}_2$ is then given by:

$$\lambda_2 \mathbf{x}_2 = \mathbf{X}_2 = \mathbf{R}\mathbf{X}_1 + \mathbf{T}, \quad \lambda_2 \in \mathbb{R}. \tag{2.13}$$

As for the pinhole camera model, vectors $\mathbf{x}_2$, $\mathbf{R}\mathbf{x}_1$, and $\mathbf{T}$ are coplanar, and the epipolar geometry constraint is formalized as:

$$\mathbf{x}_2{}^\mathsf{T} \hat{\mathbf{T}} \mathbf{R} \mathbf{x}_1 = \mathbf{x}_2{}^\mathsf{T} \mathbf{E} \mathbf{x}_1 = 0. \tag{2.14}$$

The epipolar constraint is one of the fundamental relations in the multi-view geometry because it allows the estimation of the 3D coordinates of point p from its images $\mathbf{x}_1$ and $\mathbf{x}_2$, given $\mathbf{R}$ and $\mathbf{T}$. That is, it allows scene geometry reconstruction. However, when point p lies on vector $\mathbf{T}$, which connects camera centers $\mathbf{O}_1$ and $\mathbf{O}_2$, it leads to a degenerative case of the epipolar constraint because vectors $\mathbf{x}_2$, $\mathbf{R}\mathbf{x}_1$, and $\mathbf{T}$ are collinear and the coordinates of point p cannot be determined.

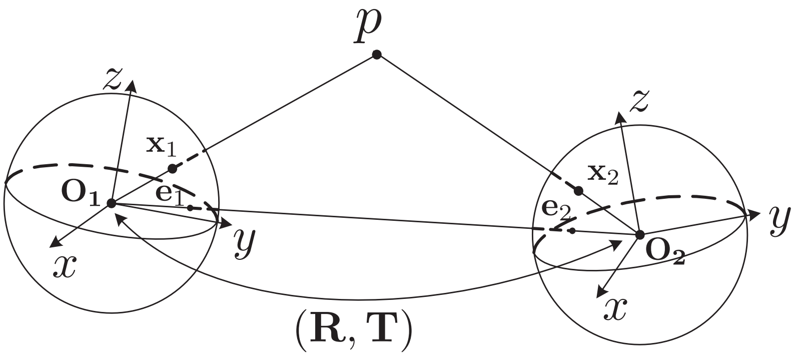


Figure 2.11: *Epipolar geometry for the spherical camera model.*

The intersection points of unit spheres of both cameras and vector $\mathbf{T}$ are the **epipoles**, denoted $\mathbf{e}_1$ and $\mathbf{e}_2$ in Figure 2.11. In other words, when point p is projected to the epipoles of two cameras, its reconstruction is not possible from these cameras.

2.4.2 Estimation of extrinsic parameters

In multi-camera systems, the calibration process includes the estimation of extrinsic parameters in addition to the intrinsic ones discussed in Section 2.3.4. Extrinsic parameters include the relative rotation and translation between cameras, which are necessary in applications like depth estimation and structure from motion. Antone and Teller [Ant02b] consider the calibration of extrinsic parameters for omnidirectional camera networks, while the intrinsic parameters are assumed to be known. Their approach decouples the rotation and translation estimation in order to obtain a linear time calibration algorithm. The Expectation maximization (EM) algorithm recovers the rotation matrix from the vanishing points, followed by the position recovery using feature correspondence coupling and the Hough transform with Monte Carlo EM refinement. Robustness of extrinsic parameters recovery can be improved by avoiding commitment to point correspondences, as presented by Makadia and Daniilidis in [Mak07]. For a general spherical image model, they introduce a correspondenceless method for camera rotation and translation recovery based on the Radon transform. They define the Epipolar Delta Filter (EDF) which embeds the epipolar geometry constraint for all pairs of features with a series of Diracs on the sphere. Moreover, they define the similarity function on all feature pairs. The main result of this work is the demonstration that the Radon transform is actually a correlation on the SO(3) group of rotations of the EDF and a similarity function; it can be efficiently evaluated by the fast Fourier transform on the SO(3). The extrinsic parameters are therefore in the maximum of the five-dimensional space given by the Radon transform.

2.5 Conclusion

This chapter has revisited state of the art approaches in representation of single-view and multi-view images. In particular, we have described the main multi-view correlation models used for 3D scene representation and discussed their advantages and disadvantages for different applications, mainly focussing on multi-view compression. Our conclusion is that state of the art multi-view correlation models are quite limited, mainly because they do not provide an effective way to combine the spatial correlation within each image and geometry-based correlation between different multi-view images. This thesis overcomes the limitation of existing models by proposing an elegant framework to merge the spatial and inter-view correlation in a single geometry-based multi-view correlation model. This approach leads to efficient distributed multi-view compression and coarse scene geometry estimation, as we will show in the rest of the thesis. Since finding a multi-view correlation model is currently the main challenge in distributed coding for camera networks, the novel correlation model and the distributed coding scheme proposed in this thesis represent an important contribution to this field of research. Moreover, we will show that the new model can be used to infer the statistically important stereo image components.

Furthermore, we have described the construction and imaging geometry of omnidirectional sensors, and particularly for catadioptric cameras whose images can be mapped directly on the sphere. Therefore, the spherical camera model and the epipolar geometry constraints for multiple spherical cameras have been discussed. The presented overview of state of the art methods in spherical image processing suggests that there still exists a need for methods that include the notion of anisotropy and directionality on the sphere. In this direction, this thesis contributes by constructing an overcomplete dictionary of anisotropic atoms on the sphere and showing its efficiency in spherical image compression.

Chapter 3

Sparse Approximations on the 2D Sphere

3.1 Introduction

The projective geometry theory for catadioptric cameras presented in Section 2.3 demonstrates that we can capture the natural radial light field using catadioptric cameras and the appropriate spherical mapping. Since spherical mapping projects lines in the 3D space into circles on the sphere, it is clear that the analysis and processing of light field needs to be performed on the sphere. This outlines the importance and need for appropriate spherical image representation and processing methods. Besides in omnidirectional imaging, processing of signals defined on the spherical manifold has applications in many other research areas. In astrophysics, for example, signal processing on the sphere allows efficient analysis of the cosmic microwave background data [Vie06]. The advantages of spherical representations for 3D object modeling and coding have been also demonstrated in computer graphics applications [Hop03, Kho00, She06].

We have described in Section 2.3.3 two main approaches for representing signals on the sphere: the spherical harmonic transform [Dri94] and the wavelet transform on the sphere [Ant99, Ant98, Sch95, Ant02a]. The drawback of spherical harmonics is their global spatial support, which makes them unsuitable for local analysis on the sphere. On the other hand, spherical wavelets are spatially localized and of multi-resolution nature, but they are not optimal in representing contours in spherical images because they do not include directionality and anisotropy. Discrete spherical wavelets have been developed in the last few years by Schröder et al. [Sch95] and by Wiaux et al. [Wia08]. The latter work presents a method for constructing directional spherical wavelets with isotropic scales. Besides spherical harmonics and spherical wavelets, a useful tool for multi-resolution analysis of spherical signals is the Laplacian Pyramid on the sphere [Tos05].

This chapter proposes to represent a spherical signal as a series of oriented and anisotropically refined functions taken from a redundant dictionary of atoms. We construct the dictionary from atoms that are edge-like functions living on the 2D sphere, and which can take arbitrary positions, shapes and orientations. Since the number of atoms in the redundant dictionary is usually much higher than the signal's dimension, there is a high probability that a given signal is well approximated with only a small number of atoms. This leads to sparse signal approximations and potentially to high signal compression ratios. The proposed sparse signal representation using the geometric dictionary on the sphere is thus exploited in this thesis for two applications:

- compression of omnidirectional images, and
- 3D objects compression.

We first propose a novel omnidirectional image coder, which uses the iterative Matching Pursuit (MP) algorithm to find a sparse approximation of the spherical image obtained by mapping an omnidirectional image on the sphere. Matching Pursuit inherently produces a progressive stream of spatially localized atoms, which is advantageously used in the design of scalable representations. Atom coefficients are appropriately quantized in an adaptive manner. Experimental results demonstrate that the new coder gives better rate-distortion performance at low rates than the standard JPEG2000 coder on unfolded images. The new coder also outperforms the SPIHT-encoded Laplacian Pyramid on the sphere.

We then present a progressive coding scheme for 3D objects based on sparse representations on the 2D sphere, thus showing the generic nature of this signal processing method. We propose to map simple 3D models on 2D spheres and then to decompose the spherical signal over a redundant dictionary of atoms on the sphere. These atoms are adapted to the characteristics of 3D objects. Our 3D coder is similar to the omnidirectional image coder, and has an additional entropy coding step of atom indexes. The entropy coding block is beneficial for 3D object compression where higher number of atoms is usually needed for good approximation. The proposed 3D coder outperforms state of the art progressive coders in terms of distortion while offering an increased flexibility for easy stream manipulations (e.g. view-dependent transmission).

The 3D object representation with sparse approximations on the sphere has been also successfully applied to 3D face recognition. This is, however, not covered in the thesis and the interested reader is referred to [Llo08a, Llo08b] for more details.

3.2 Redundant dictionary on the 2-D sphere

3.2.1 Preliminaries

Since a spherical signal is in general composed of multi-dimensional features, we propose to decompose it as a series of atoms taken from a redundant dictionary of functions defined on the 2D sphere. While there is a priori no restriction on the dictionary construction, it is in general constructed as a set of different waveforms, where each waveform is defined by a generating function. Each generating function can serve as a basis for building an overcomplete dictionary, simply by changing the function parameters (e.g., position or scale indexes). The usage of generating functions advantageously leads to structured parametric dictionaries, whose indexes directly correspond to atom characteristics. Furthermore, the storage or transmission of the dictionary become unnecessary, since atoms can be reconstructed only from their indexes.

Dictionary construction is certainly the most important step towards efficient approximation algorithms. Increasing the number of functions generally increases the redundancy of the dictionary, and thus the approximation performance: there is indeed an increasingly high probability that prominent signal features can be efficiently captured by few atoms. At the same time, it also increases the size of the dictionary, and most probably augments the coding rate and the search complexity. We now discuss in more detail the construction of the overcomplete dictionary that we propose for expansions of signals on the 2D sphere. It involves the three following steps:

- definition of the generating function(s) on the sphere,
- definition of the motion of atoms on the sphere, and their rotation around their axis,
- implementation of the anisotropic scaling of atoms.

Let g denote a square-integrable generating function on the unit 2-sphere S^2 (i.e., $g(\theta, \varphi) \in L^2(S^2)$). By combining motion, rotation and scaling, we form an overcomplete set of atoms g_γ, where $\gamma = (\tau, \nu, \psi, \alpha, \beta) \in \Gamma$ is the atom index. This index is described by five parameters that respectively represent: the position of the atom on the sphere, τ along zenith θ angle, and ν along azimuth φ angle; its orientation ψ; and the scaling parameters (α, β). In order to finally map the atoms on the sphere, we use an inverse stereographic projection from the complex plane $\mathbb{C}$, to the 2D sphere. There are few different definitions of the stereographic projection, depending on the

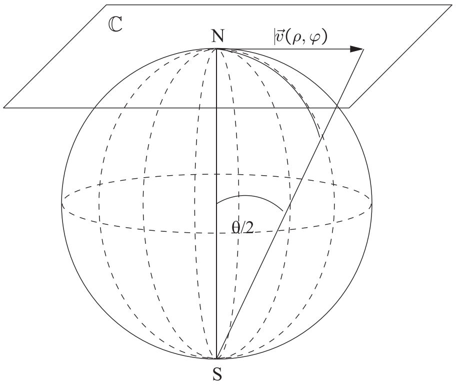


Figure 3.1: *Stereographic Projection. A point on the 2-D sphere can be uniquely mapped on the plane tangent to the North Pole.*

position of the plane ℂ. We will use here the stereographic projection at the North Pole shown in Figure 3.1, which is different from the one in Section 2.3.1 where ℂ contains the equator.

The stereographic projection at the North Pole can be expressed as $\Omega : S^2 \rightarrow \mathbb{C}$ and written as:

$$\Omega(\omega) = \mathbf{v} = \rho e^{j\varphi} = 2\tan\left(\tfrac{\theta}{2}\right) e^{j\varphi} \ , \tag{3.1}$$

with $\omega \equiv (\theta, \varphi)$ and $0 \leqslant \theta \leqslant \pi, -\pi \leqslant \varphi < \pi$. Since the stereographic projection is bijective, any point with polar coordinates (ρ, φ) and represented by a vector $\mathbf{v} = (\rho\ \cos\varphi, \rho\ \sin\varphi)$ on the tangent plane, can be uniquely mapped back onto the 2D sphere. We use that property in the design of the dictionary, as presented below.

3.2.2 Generating functions

Under the assumption that spherical images are mostly composed of smooth surfaces and singularities aligned on pieces of great circles, we propose to build the dictionary on two generating functions. First, in order to efficiently capture the singularities, we use a generating function that resembles to a piece of contour on the sphere. We define a contour-like function on the plane tangent to the North Pole, i.e., in the space $L^2(\mathbb{R}^2)$. Our choice is a function that is a Gaussian function in one direction and its second derivative in the orthogonal direction:

$$g_{image}(\vec{v}) = \frac{1}{K_1}\left(4x^2 - 2\right)\exp\left(-\left(x^2 + y^2\right)\right), \tag{3.2}$$

where $\mathbf{v} = (x, y)$ is a vector in $\mathbb{R}^2$, and K_1 is a normalization factor. Note that this function has been efficiently used for image coding [Van01, Fig06].

The motivation for the choice of a Gaussian kernel lies in its optimal joint spatial and frequency localization. On the other hand, the second derivative in the orthogonal direction is used to filter out the smooth polynomial parts of the signal and capture the signal discontinuities. The generating function from Eq. (3.2) can be further expressed in polar coordinates, as:

$$g_{rect}(\rho, \varphi) = -\frac{1}{K_1}\left(4\rho^2 cos^2\varphi - 2\right)\exp\left(-\rho^2\right). \tag{3.3}$$

We have taken $g_{rect} = -g_{image}$ to have the positive central value of the atom. By inverse stereographic projection $\Omega^{-1} : \mathbb{R}^2 \rightarrow S^2$, the generating function is mapped on the sphere, and can be written as:

$$g_{HN}(\theta, \varphi) = -\frac{1}{K_1}\left(16\tan^2\frac{\theta}{2} cos^2\varphi - 2\right)\exp\left(-4\tan^2\frac{\theta}{2}\right). \tag{3.4}$$

The generating function g_{HN} defines an edge-like atom that is centered exactly on the North Pole.

In order to efficiently represent the smooth areas in spherical signals, corresponding to low-frequency (LF) components, we propose to use a second generating function for the construction of the dictionary. This function is a two-dimensional Gaussian function in $L^2(S^2)$:

$$g_{LN}(\theta, \varphi) = \frac{1}{K_2} \exp\left(-4 \tan^2 \frac{\theta}{2}\right), \tag{3.5}$$

where K_2 is a normalization constant. Eq. (3.5) represents an isotropic function, centered at the North Pole. Using the dictionary built on two generating functions actually improves the approximation rate, but does not increase the search complexity. In our implementation, the dictionary is indeed divided into two distinct parts, one with LF atoms (LF part) and the other of oscillating or high-frequency atoms (main part), which are used successively to form the signal expansion.

Now that the generating functions have been defined, we form the redundant dictionary by applying geometrical transformations to these functions. In other words, the dictionary is constructed by moving the generating functions on the sphere, by rotation of the functions around their axis, and by anisotropic scaling.

3.2.3 Motion on the sphere

Motion and rotation belong to the group of affine transformations of the unit 2D sphere S^2. They are both realized by a single rotation $\varrho \in SO(3)$, where $SO(3)$ is the rotation group in $\mathbb{R}^3$. This is equivalent to applying a unitary operator $\mathbf{\Pi}_\varrho$ on the matrix of Cartesian coordinates (x, y, z) of the unit sphere, denoted as $\mathbf{P}$:

$$\mathbf{P}_r = \mathbf{\Pi}_\varrho \mathbf{P} = \mathbf{R}(\psi)\mathbf{U}(\tau)\mathbf{R}(\nu)\mathbf{P}, \qquad \varrho \in SO(3), \tag{3.6}$$

where $[\mathbf{P}]_{3\times N}$ is the matrix of (x, y, z) coordinates of the non-transformed unit sphere, and $[\mathbf{P}_r]_{3\times N}$ is the matrix of (x, y, z) coordinates of the transformed unit sphere. Three rotation matrices $\mathbf{R}(\nu)$, $\mathbf{U}(\tau)$ and $\mathbf{R}(\psi)$ realize the rotation given by Euler angles (ν, τ, ψ), which respectively describe the motion of the atom on the sphere by angles ν and τ, and the rotation of the atom around its axis by an angle ψ. These rotation matrices are given by:

$$\mathbf{R}(\nu) = \begin{pmatrix} \cos\nu & \sin\nu & 0 \\ -\sin\nu & \cos\nu & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{U}(\tau) = \begin{pmatrix} \cos\tau & 0 & \sin\tau \\ 0 & 1 & 0 \\ -\sin\tau & 0 & \cos\tau \end{pmatrix},$$

$$\mathbf{R}(\psi) = \begin{pmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The generating function, as defined in Eq. (3.4), is therefore transformed into an atom that can be moved to a particular point (τ,ν) on the sphere, and rotated.

3.2.4 Anisotropic refinement of atoms on the sphere

In order to efficiently approximate the elongated characteristics of spherical signals, we further deform atoms by anisotropic refinement that scales the generating function differently in each orthogonal direction, with scale factors α and β. We perform the scaling operation on the plane tangent to the North pole and then map the resulting atom on the sphere S^2 by inverse stereographic projection. Let $\mathbf{v} = (x, y)$ denote a vector on the tangent plane; the anisotropic scaling operator is then expressed as:

$$D(\alpha, \beta)g(\mathbf{v}) = Cg(\alpha x, \beta y), \tag{3.7}$$

where the constant C is a normalization factor. The coordinates of the vector after scaling, $\mathbf{v_s}$, become:

$$\begin{aligned} x_s &= \alpha x = \alpha \rho \cos\varphi, \\ y_s &= \beta y = \beta \rho \sin\varphi. \end{aligned} \tag{3.8}$$

In polar coordinates, it translates to:

$$\begin{aligned} \rho_s &= \sqrt{x_s^2 + y_s^2} = \rho\sqrt{\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi}, \\ \varphi_s &= \arctan\frac{y_s}{x_s} = \arctan\frac{\beta\sin\varphi}{\alpha\cos\varphi}. \end{aligned} \tag{3.9}$$

Anisotropic refinement of high frequency atoms, as given in Eq. (3.3), is obtained by replacing the polar coordinates with the ones obtained after scaling. They can be written as:

$$g_{ra}(\rho, \varphi) = -\frac{1}{K_A}\left(4\alpha^2\rho^2\cos^2\varphi - 2\right)\exp\left(-\rho^2\left(\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi\right)\right), \tag{3.10}$$

where K_A is a normalization factor. By inverse stereographic projection $\Omega^{-1} : \mathbb{R}^2 \rightarrow S^2$, the reshaped atom is mapped on the sphere, and can be written as:

$$g_{HF}(\theta, \varphi) = -\frac{1}{K_A}\left(16\alpha^2\tan^2\frac{\theta}{2}\cos^2\varphi - 2\right)\exp\left(-4\tan^2\frac{\theta}{2}\left(\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi\right)\right). \tag{3.11}$$

On the other hand, the low-frequency atoms after anisotropic refinement, can be written as:

$$g_{LF}(\theta, \varphi) = \frac{1}{K_G}\exp\left(-4\tan^2\frac{\theta}{2}\left(\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi\right)\right), \tag{3.12}$$

where K_G is a normalization factor.

By a proper choice of the transformation parameters, one finally obtains an overcomplete parametric dictionary of functions, which is used to represent spherical signals. Sample atoms are illustrated in Figure 3.2.

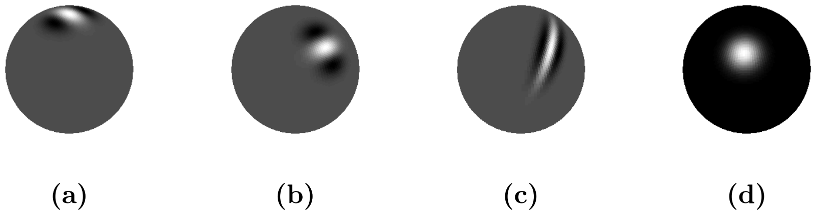

(a) **(b)** **(c)** **(d)**

Figure 3.2: *Anisotropic atoms: a) on the North pole ($\tau = 0$, $\nu = 0$), $\psi = 0$, $\alpha = 4$, $\beta = 4$; b) $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{2}$, $\psi = 0$, $\alpha = 4$, $\beta = 4$; c) $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{2}$, $\psi = \frac{\pi}{4}$, $\alpha = 8$, $\beta = 2$; d) Low frequency atom: $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{4}$, $\psi = \frac{\pi}{4}$, $\alpha = 4$, $\beta = 4$.*

3.3 Matching Pursuit algorithm for sparse approximation

Finding the sparsest representation of a spherical signal with functions taken from a redundant dictionary is an NP-hard problem [Nat95]. When we have parametric dictionaries characterized by high coherence, as the one described above, none of the existing sparse approximation algorithms guarantees the optimal solution. As discussed in Section 2.1.2, in the case of coherent dictionaries it is more reliable to use the Matching Pursuit (MP) algorithm for sparse approximation, since it offers the exponential decay of the approximation error.

Under its generic form, MP is an algorithm that iteratively decomposes a signal into a linear combination of waveforms, or atoms. Interestingly, very few restrictions are imposed on the dictionary construction, besides the fact that it should at least span the space of the signal to be represented. In other words, the dictionary is defined as a set of unit norm vectors $\mathcal{D} = \{g_\gamma\}_{\gamma\in\Gamma}$ in a Hilbert space H. In order to be able to represent each vector in H as a linear combination of unit norm vectors in $\mathcal{D}$, the dictionary must satisfy the completeness property (i.e., $\mathcal{D}$ spans H).

Let $y \in H$ denote a function that we want to approximate with a linear expansion over $\mathcal{D}$. With MP, an M-term linear expansion is obtained by successive approximations of $R^m y$ through orthogonal projections on dictionary vectors:

$$y = \sum_{m=0}^{M} \langle R^m y, g_{\gamma_m} \rangle \, g_{\gamma_m} + R^{M+1} y, \tag{3.13}$$

where $R^m y$ is the residue after $m-1$ iterations of the algorithm ($R^0 y = y$). One must choose, at each iteration, the atom that best approximates $R^m y$, with the maximal projection $|\langle R^m y, g_{\gamma_m} \rangle|$ over the dictionary:

$$|\langle R^m y, g_{\gamma_m} \rangle| = \sup_{\gamma\in\Gamma} |\langle R^m y, g_\gamma \rangle|. \tag{3.14}$$

When $M \to \infty$ it can be shown that:

$$y = \sum_{m=0}^{M} \langle R^m y, g_{\gamma_m} \rangle \, g_{\gamma_m}. \tag{3.15}$$

It has been proven that the residue decays exponentially in a finite dimensional space [Mal93]. The decay rate depends on the correlation between the residue and the dictionary elements. Hence, the construction of an efficient dictionary, adapted to the structure of the signal y, represents a crucial step.

Overall, MP offers a sub-optimal solution to the optimal (sparsest) signal representation problem, since it iteratively approximates the signal in a greedy manner. However, it allows for an efficient approximation of the signal by rapidly capturing its most important components. Therefore, MP results in a progressive approximation of a given signal, which is an interesting property in the design of scalable coders [Fro00]. At the same time, MP does not impose any condition on the dictionary design. Moreover, the complexity of the signal reconstruction is linear and therefore low.

MP is a general algorithm for sparse approximation, and can be applied to functions in any Hilbert space. We will use MP and the redundant dictionary on the sphere constructed in the Section 3.2 to approximate spherical signals. Throughout the rest of this chapter, we will refer to MP on the sphere as the Spherical Matching Pursuit (SMP) and we will apply it to compression of omnidirectional images and 3D objects.

3.4 Omnidirectional image compression

Although omnidirectional cameras have found wide range of applications lately, there has been little work done on compression of the acquired images. The specific nature of the projective geometry for catadioptric systems leads to a clear conclusion that applying standard 2D planar image compression techniques to omnidirectional images is suboptimal. To the best of our knowledge, the only work that considered the geometry of catadioptric systems for omnidirectional image and video compression was presented by Bauermann et al. [Bau06]. They proposed a preprocessing step that maps each image in an omnidirectional video into a panoramic image and then compresses that video using the standard H.264 coder. However, panoramic mapping cannot fully exploit the correlation statistics of the radial light field captured by an omnidirectional camera. This is because the sampling density of light rays is non-uniform on a panoramic cylinder, while the pixel distribution on the cylinder is uniform. We propose here to use the novel spherical signal representation method with overcomplete expansions to reach high compression gains for omnidirectional images.

3.4.1 Omnidirectional image coder

Our objective is to build on the effective approximation properties offered by redundant expansions to obtain compressed versions of omnidirectional images on the sphere. The block diagram of the proposed coder, Omni-SMP (omnidirectional image coder based on the Spherical Matching Pursuit), is presented in Figure 3.3. At the encoder side, an omnidirectional image is first mapped to a spherical image $y(\theta, \varphi)$ by inverse stereographic projection, as explained in Section 2.3.1. Matching Pursuit is then performed on the spherical image, in order to select from the dictionary a series of M atoms indexed by $\{\gamma_n\}$,$n = 1, ..., M$, with their relative coefficients $\{c_n\}$. Atoms are first sorted in decreasing order of their coefficient magnitudes. The coefficients are then uniformly quantized with a decaying quantization range, based on the method proposed by Frossard et al. [Fro04]. This method takes advantage of the property that energy of the Matching Pursuit coefficients is limited by an exponentially decaying upper-bound. Let $\{q_n\}$ denote the set of codewords for quantized coefficients. The decoder first performs the inverse quantization of MP coefficients to obtain reconstructed coefficients $\{\hat{c}_n\}$. It then reconstructs the approximated spherical image $\hat{y}(\theta, \varphi)$ by linear combination of atoms whose relative weights are given by the MP coefficients. This reconstruction step on the decoder side has a low computational complexity, roughly proportional to the number of atoms.

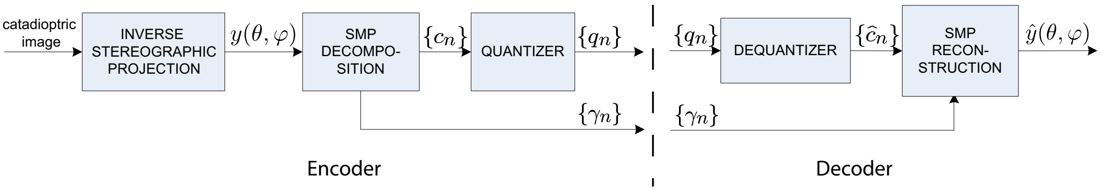


Figure 3.3: *Omni-SMP coding scheme.*

3.4.2 Implementation

In the dictionary presented in Section 3.2, atom indexes are defined by the parameters of the generating functions, and they obviously take discrete values. In general, a fine granularity of atom indexes leads to high redundancy, and likely to high approximation rate. At the same time, it leads to a large dictionary, with possibly high coding cost. The design of an optimal dictionary is still an open problem, and not targeted by this chapter. Here, we propose to use a dictionary mostly built on empirical choices for atom parameter values. First, we use the equiangular spherical grid to drive the values of the position parameters, τ and ν; both parameters are uniformly distributed on the interval $[0, \pi]$, and $[-\pi, \pi)$, respectively, with a resolution N that is identical to the input signal. The resolution of a spherical signal is twice as large as its bandwidth, i.e., $N = 2B$. The rotation parameter ψ is uniformly sampled on the interval $[0, \pi]$, with $N_\psi = 16$ orientations. Therefore, dictionary parameters uniformly cover the space of motions on the sphere. Finally, the scaling parameters α and β are distributed in a logarithmic manner, from 1 to half of the signal's resolution, with a granularity of one third of octave. Due to the definition of atoms (Eq. (3.11)), scaling parameters are inversely proportional to the atoms's size. The largest atom has a scale 1 and it covers half of the sphere. To reduce the dictionary size and hence the complexity, we limit the scale β to be $\beta \leqslant \alpha$. This constraint gives a dictionary with atoms elongated along edges, which efficiently represent image structures. For low-frequency atoms, the maximal value of scaling parameters is chosen to be 1/16 of the signal's resolution. Motion and rotation parameters are discretized in the same way as for anisotropic atoms. This choice is mostly due to the use of fast correlation on the SO(3) group in the Matching Pursuit algorithm, as explained below.

In our implementation, the full dictionary is divided into low-frequency atoms g_{LF}, and high-frequency ones. During first iterations, MP uses the low-frequency sub-dictionary, and later switches to the anisotropic sub-dictionary when energy of the coefficients starts to saturate, or more precisely when :

$$\frac{|C^m|}{\|R^m\|_2} \to const, \tag{3.16}$$

where C^m denotes a projection after $m-1$ iterations. In each of these sub-dictionaries, MP performs a full search to determine the highest energy atom. Our discretization of the motion parameters (τ, ν, ψ) in the dictionary permits us to use the Fourier transform on the SO(3) group [Kos08] and compute the correlation of signals on the SO(3). In particular, we have used in our implementation the *SOFT* library[1], which is a part of the *YAW toolbox*[2]. This transform allows us to identify the position and the rotation of the atom on the sphere that has the best correlation with the signal. One SO(3) correlation allows to determine the parameters (τ, ν, ψ) for each atom with given scale parameters. Therefore, our algorithm iterates over the scale parameters: for each couple (α, β), it computes the correlation on the SO(3) between the residual signal $R^m f$ and the atom with parameters $(\tau = 0, \nu = 0, \psi = 0, \alpha, \beta)$. The convolution coefficient with the largest magnitude corresponds to the position and rotation parameters (τ, ν, ψ) of the best matching atom, for that pair of scales. The coefficient of the corresponding atom is computed by the inner product defined on the sphere, given as:

$$\langle R^m f, g\rangle = \int_\theta \int_\varphi R^m f(\theta, \varphi) g(\theta, \varphi) \sin\theta \mathrm{d}\theta \mathrm{d}\varphi. \tag{3.17}$$

Finally, the algorithm selects among all pairs of scales the atom with the largest coefficient, removes its contribution from the residual signal, and repeats the whole procedure until a stopping criteria is met (e.g., a pre-defined number of atoms or an energy threshold). The search algorithm is described in Algorithm 1.

Algorithm 1 Full search in the dictionary

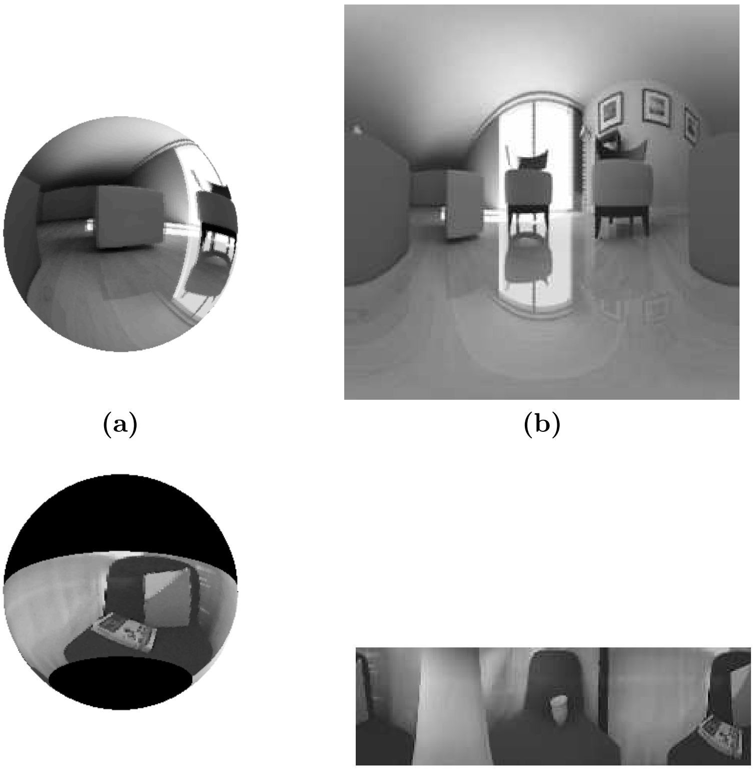
```
for all scale couples $(\alpha(j), \beta(k))$ do
    $C = \mathrm{softcorr}(R^m y, g(0, 0, 0, \alpha(j), \beta(k)))$
    $C_{max} = \max_{1<t<N, 1<l<N, 1<p<N_\psi} C(t, l, p)$
    $\tau = \tau(t); \nu = \nu(l); \psi = \psi(p)$
    $P(j, k) = \langle R^m y, g(\tau, \nu, \psi, \alpha(j), \beta(k))\rangle$
end for
$P_{max} = \max_{j,k}(P(j, k))$
$\alpha = \alpha(j); \beta = \beta(k)$
```

Note that MP with a full search of the dictionary can become quite complex when the dictionary size is large. However, the search complexity can be significantly reduced by efficient arrangement of atoms [Jos06a], or by parallelization [Rah06], possibly at the expense of a small decrease in the

---

[1] http://www.cs.dartmouth.edu/ geelong/sphere/

[2] http://fyma.fyma.ucl.ac.be/projects/yawtb/



(a) (b)

(c) (d)

Figure 3.4: *Original images, resolution* $2B = 256$*: (a) spherical image Room; (b) unfolded spherical image Room; (c) omnidirectional image Lab mapped to a spherical image; (d) unfolded omnidirectional image Lab.*

approximation rate. The decoding complexity is however very low, and it is performed in linear time.

3.4.3 Experimental compression performance evaluation

Compression performance of the proposed Omni-SMP coder has been evaluated on synthetic spherical and natural omnidirectional images. The original spherical image of the synthetic scene Room, rendered by the ray tracer engine[3], is shown in Figure 3.4(a). We show in Figure 3.4(b) its unfolded version in order to display all image features. Natural Lab image has been captured by a paracatadioptric camera with a mirror from the Remote Reality Corporation[4]. The camera has a view range of 360-degrees with respect to the azimuth angle φ and 35 to 92.5 degrees with respect to the zenith angle θ. Camera calibration with respect to the intrinsic parameters has been performed by circle fitting (see Section 2.3.4). Omnidirectional Lab image has been mapped to the spherical image in Figure 3.4(c), whose unfolded version is shown in Figure 3.4(d). Both test spherical images have resolution $2B = 256$ pixels on the equiangular spherical grid.

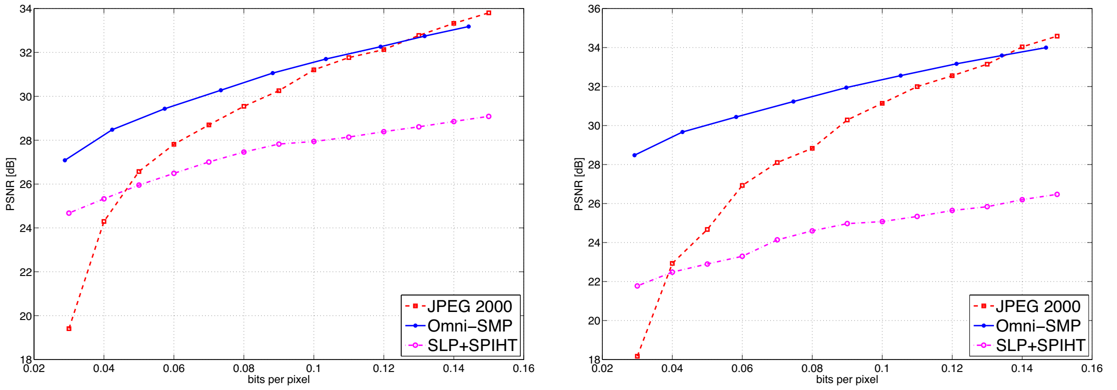
The rate-distortion (RD) performance of Omni-SMP has been evaluated. It reports the number of bits per pixel of the compressed image versus the quality of the reconstructed image, measured by the PSNR (Peak Signal to Noise Ratio). When the range of a grayscale image is from zero to one, PSNR is given as:

$$PSNR = 20 \log_{10} \frac{1}{D}, \tag{3.18}$$

---

[3] http://yafray.org/

[4] http://www.remotereality.com/



(a) **(b)**

Figure 3.5: *Rate-distortion performance for the (a) Room image, (b) Lab image.*

where D is the distortion evaluated as the root mean square error between the decoded and original image:

$$D = \sqrt{\frac{1}{N^2}\langle f - \hat{f}, f - \hat{f}\rangle}. \tag{3.19}$$

The inner product $\langle f - \hat{f}, f - \hat{f}\rangle$ is evaluated on the sphere as given by Eq. (3.17). For the natural Lab image, PSNR has been evaluated only on the non-black (informative) part of the sphere.

Since there are no spherical wavelet-based methods adapted to the compression of omnidirectional images, but only to shape compression, we compare Omni-SMP to JPEG2000, which is a wavelet based coder for planar images and currently state of the art method in image coding. Performing compression on unfolded spherical images using the planar image coder JPEG2000 is very similar to projecting omnidirectional images to panoramic images and then applying JPEG2000. Omnidirectional image coding with panoramic representation outperforms direct coding of captured omnidirectional images, as pointed out by Bauermann et al. [Bau06]. This advantage relies on the fact that the visual information captured by omnidirectional cameras is usually displayed by projection on planar perspective views. Therefore, we compare Omni-SMP with JPEG2000 applied on unfolded (panoramic) omnidirectional images. In order to have a correct comparison, the distortion introduced by JPEG2000 coding has been evaluated on the sphere, in the same manner as done for Omni-SMP. We have also compared Omni-SMP to a multiresolutional method that employs the Spherical Laplacian pyramid (SLP) [Tos05], followed by the SPIHT [Sai96] coding of LP coefficients. The solid lines in Figures 3.5(a) and (b) present the RD performance of the Omni-SMP coder for Room and Lab images, respectively. On the same figures, we plot the RD performance of JPEG2000 on unfolded images, with the dashed line. We can see that the proposed Omni-SMP coder outperforms JPEG2000 for up to 6dB at low rates. However, Omni-SMP is not efficient at high rates since the designed dictionary is not optimized to approximate texture information that is dominant at high rates, but rather for structural image information. RD curves for the SLP+SPIHT method are shown with dash-dotted lines. Due to the redundancy of the Laplacian Pyramid, RD performance of this method is much worse than the performance of Omni-SMP and JPEG2000.

Finally, we observe the visual image quality of the proposed Omni-SMP coder and compare it to JPEG2000. Figures 3.6 (a) and (b) show respectively the decoded images using Omni-SMP and JPEG2000 coder, at the same bit rate 0.057bpp. We can see that the image decoded using Omni-SMP is more visually pleasing, with sharper edges and smoother uniform regions. Moreover, we can see how the atoms in the image decomposition fit to 3D lines that are projected to curvatures on the sphere. On the other hand, JPEG2000 introduces artifacts that degrade the structure of the image. At higher rates, images decoded using Omni-SMP and JPEG2000 become closer with respect to the PSNR value, but the coding artifacts are less annoying for Omni-SMP. This is shown in Figures 3.6 (c) and (d), which display the decoded Room images at bit rate 0.088bpp. Similar observations can be made for the Lab image (see Figure 3.7).

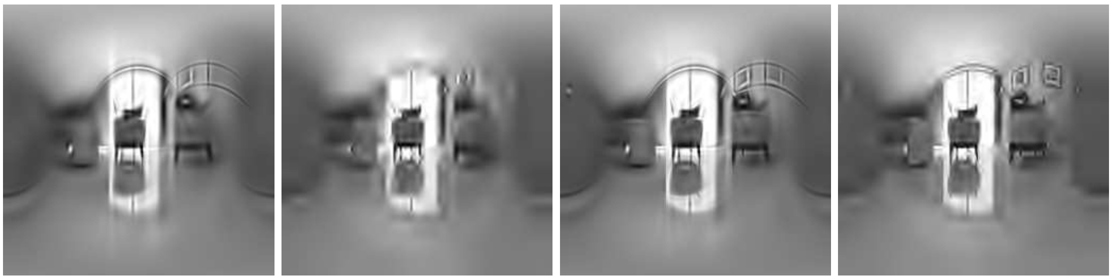
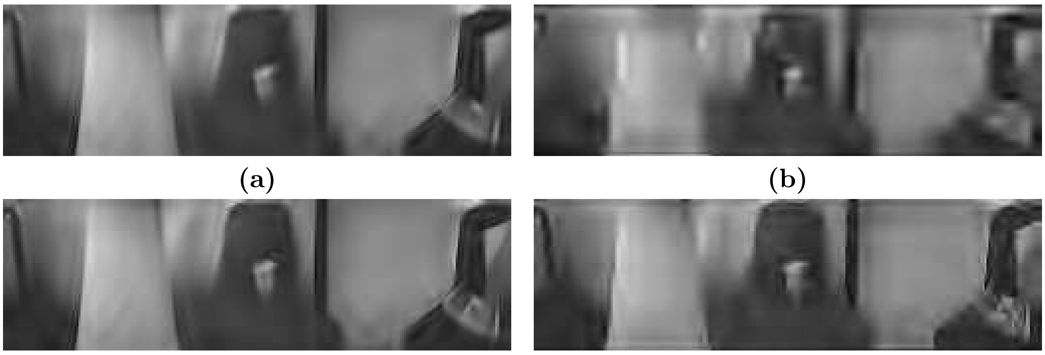

(a) (b) (c) (d)

Figure 3.6: *Decoded Room image: (a) Omni-SMP coder at 0.057 bpp, PSNR=29.43 dB; (b) JPEG coder at 0.057 bpp, PSNR=27.56 dB; (c) Omni-SMP coder at 0.088 bpp, PSNR= 31.06 dB; (b) JPEG coder at 0.088 bpp, PSNR=30.25 dB.*


(a) (b)

(c) (d)

Figure 3.7: *Decoded Lab image: (a) Omni-SMP coder at 0.058 bpp, PSNR=30.44 dB; (b) JPEG coder at 0.058 bpp, PSNR=25.50 dB; (c) Omni-SMP coder at 0.089 bpp, PSNR=31.94 dB; (d) JPEG coder at 0.089 bpp, PSNR=30.29 dB.*

3.5 3D object compression

The widespread use of 3D data in many areas like gaming or entertainment, architecture, robotics, or medical imaging, has created an essential need for efficient compression of 3D models. Simultaneously, the increasingly large variety of decoding engines, with heterogeneous capabilities and connectivity, imposes a need for multi-resolution representations, as well as low-complexity decoders, without the need for dedicated hardware. The most common approaches for 3D data representation are based on polygonal meshes, which are described by both geometry (i.e., the position of vertices in space) and connectivity information, as well as optional information about normals, colors and textures. It generally results in models built on arbitrarily defined and non-uniform grids that lead to efficient decoding performance on dedicated hardware. These forms of representations stay however quite voluminous, and do not provide a lot of flexibility for adaptation to the requirements of specific applications, or to the constraints imposed by the decoding engine.

Numerous works have addressed the coding of 3D models, and we just mention here the most relevant ones in the context of the present work. The first mesh geometry compression scheme, introduced by Deering [Dee95], was based on *triangle strips* and *triangle fans*, and implemented in GL [Sil91] and OpenGL [Nei97]. In GL, triangles are ordered to form strips, whose connectivity is defined with a *marching bit* per triangle; it specifies to which of the two free edges of the current triangle the next triangle has to be attached. In OpenGL, triangles are attached alternatively on left and right edges, and no connectivity information is transmitted. The drawback of this technique is that most meshes have twice as many faces as vertices: each vertex has to be transmitted twice, in average. Taubin and Rossignac later introduced the Topological Surgery (TS) scheme, which is a single-resolution triangular mesh compression scheme that preserves the connectivity [Tau98a]. After extensions to arbitrary manifold meshes, TS has become a part of the MPEG-4 standard. One of the first progressive transmission schemes for multi-resolution triangular manifold meshes has been introduced by Hoppe [Hop96]. A triangular manifold mesh is represented by a base mesh followed by a sequence of successive vertex split refinements. Taubin has introduced the Progressive Forest Split (PFS) scheme, which highly reduces the number of levels of detail, and thus unnecessary information [Tau98b]. Along with the Topological Surgery (TS), PFS represents the core of 3D mesh coding in the MPEG-4 standard. Alternatively, Karni and Gotsman proposed a 3D mesh compression method based on spectral decompositions [Kar00].

A common characteristic of multi-resolution mesh-based compression schemes mentioned above is that most of the geometry information of a coarse mesh is embedded within a finer mesh, except for a set of vertices or edges that result from vertex or edge split operations. This kind of surface sampling does not necessarily lead to the best approximation at a given resolution. On the other hand, by representing a 3D model as a continuous function on a 2D surface, positions of vertices are determined by uniform sampling of this function so they are different from one resolution to another. This results in equal approximation enhancement over the 3D object surface, which is an important advantage of 2D surface methods versus mesh-based methods. Moreover, the mapping of a 3D object in the continuous space enables the use of various signal transforms towards building fully progressive representations. Schröder and Sweldens proposed one of the earliest works that represent 3D models as functions defined on the surface of a sphere, as an alternative to mesh-based approaches [Sch95]. They introduced a lifting scheme to construct bi-orthogonal spherical wavelets with customized properties. Shape compression using spherical wavelets has become recently an active area of research. The progressive coding scheme introduced by Khodakovsky et al. uses the wavelet transform, a zerotree coding and a subdivision-based reconstruction to improve the compression ratio [Kho00]. Hoppe and Praun have described a shape compression technique using spherical geometry images [Hop03]. In comparison to ordinary image wavelets, spherical wavelets are shown to provide better compression performance for surfaces that can be nicely parametrized on the sphere. However, the related compression techniques suffer from rippling artifacts for surfaces with long extremities.

This section proposes a novel coding scheme for 3D objects, built on the proposed sparse approximation on the sphere. The new coder provides a progressive representation with flexibility in the stream manipulation, whilst achieving good compression performance. A progressive representation enables the decoder to construct a model at different resolutions, simply by proper stream truncation to meet a well-chosen rate-distortion trade-off. At the same time, a flexible representation provides the possibility to manipulate the model in the compressed domain, to decode the model at different sizes, or from different viewpoints, for example. We first propose to resample 3D data on a regular spherical grid, thus reducing the dimension of the input data into a 2D data set. A 3D surface, which can be represented as a function on a 2D sphere is a genus-zero[5] surface that has only one intersection point with each radial line from the center of the point cloud, and thus does not contain any folds. We will reference to these models that can be mapped on the 2D sphere, as *simple genus-zero*, or *star-shape* models. We eventually show that the representation of more complex models is feasible by splitting it into several spherical mappings.

---

[5] A mesh has a genus g, iff one can cut the mesh along 2g closed loops without disconnecting the mesh.

3.5.1 3D-SMP coder

The initial block of the proposed 3D-SMP encoder shown in Figure 3.8 maps 3D objects into spherical data. It first extracts a set of vertices $p_i = (x_i, y_i, z_i)$, which represents a point cloud of a 3D model. This point cloud generally describes a set of non-uniformly spaced samples on the 2D sphere S^2 that define a function $f : S^2 \rightarrow \mathbb{R}$ as $R_i = f(\theta_i, \varphi_i)$. Since the proposed coding scheme requires the spherical data that is sampled on an equiangular grid of resolution $N = 2B$, an interpolation step may be needed. Because of its low complexity, we have chosen a simple nearest neighbor interpolation method, where each value on the equiangular spherical grid R_{int} is interpolated as an average of its four nearest neighbors. In addition to enabling the use of processing algorithms like the Fast Spherical Fourier Transform, the equiangular spherical grid has a regular structure that can be exploited at the decoder for regenerating the mesh connectivity removed by the encoding process.

After the SMP decomposition of the spherical signal $f(\theta, \varphi)$, MP coefficients $\{c_n\}, n = 1, ..., M$ are sorted in decreasing order of magnitude and they are quantized. The quantizer is inspired from the scheme proposed by Frossard et al. [Fro04], and uses the piecewise linear approximation of the exponential upper-bound for quantization. Codewords $\{q_n\}$ of quantized coefficients, and discrete atom indexes $\{\gamma_n\}$ are finally encoded with an arithmetic coder [Wit87], in order to obtain a compact representation. Interested readers are referred to [Fig06] for more details about quantization and entropy coding of MP atoms.

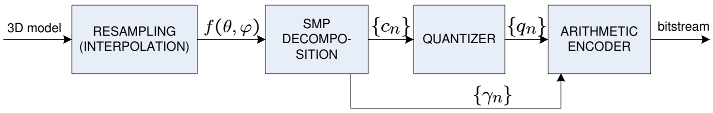


Figure 3.8: *3D-SMP encoding scheme.*

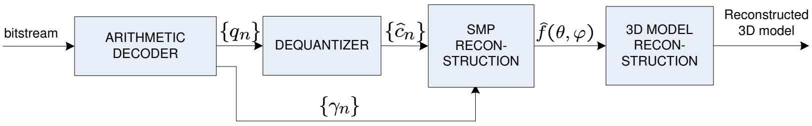
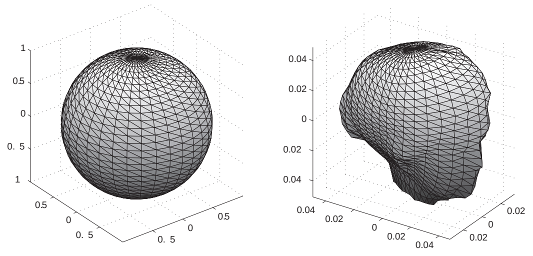


Figure 3.9: *3D-SMP decoding scheme.*

The decoder, as represented in Figure 3.9, first performs the entropy decoding and inverse quantization to obtain the quantized coefficients $\{\hat{c}_n\}$. This is followed by SMP reconstruction of the spherical signal $\hat{f}(\theta, \varphi)$ that represents the decoded spherical signal. The decoder then generates the decoded 3D object in the form of a standard polygonal mesh, as accepted by all modern computer graphics applications and hardware. Since the encoder has completely discarded the mesh connectivity information of the original 3D model, the decoder has to generate new connectivity. This problem can be formulated as a surface reconstruction problem from an unorganized point cloud, which is still an active area of research, and many surface reconstruction algorithms already exist (e.g., [Hop92]). Since we are primarily dealing with simple models parameterized as one spherical function, we can use the a priori knowledge of (θ, φ) coordinates for each vertex on the spherical grid and construct a *semi-regular* connectivity structure. A mesh with semi-regular connectivity has almost all vertices of valence six (i.e., six incident edges), except for a few isolated extraordinary vertices. The connectivity matrix is defined with indexes of three incident vertices for each face. In order to obtain a semi-regular triangular mesh, we can divide the spherical grid into rings limited by two successive values of θ, and then triangulate each ring to produce a triangular strip. Such mesh construction is illustrated in Figure 3.10 (a), which shows the triangular subdivision of the sphere. Figure 3.10 (b) represents the same grid, but applied to the Venus model. All vertices are of valence six, except the two poles, thus the resulting mesh is indeed semi-regular.

For more complex models whose representation requires multiple spheres, the method explained above is not directly applicable, since the boundary between two neighboring spheres does not necessarily coincide with a great circle on the sphere. In these cases, a simpler solution would be to use a more generic surface reconstruction algorithm. The proposed 3D-SMP scheme uses the algorithm of Cohen-Steiner and Da [Coh01][6].



(a) **(b)**

Figure 3.10: *Generating the connectivity matrix: a) Sphere connectivity; b) Connectivity on the Venus model.*

3.5.2 Implementation

Spherical signals that represent 3D objects have somewhat different characteristics than omnidirectional images, usually with a smaller dynamic range. We therefore propose to slightly modify the generating functions in Eq. (3.11) and Eq. (3.12) in order to account for these characteristics, and we use the following functions:

$$g_{HO}(\theta,\varphi) = -\frac{1}{K_{ho}}\left(16a_1^2\tan^2\frac{\theta}{2}\cos^2\varphi-2\right)\exp\left(-\tan^2\frac{\theta}{2}\left(a_1^2\cos^2\varphi+a_2^2\sin^2\varphi\right)\right), \quad (3.20)$$

$$g_{LO}(\theta,\varphi) = \frac{1}{K_{lo}}\exp\left(-\tan^2\frac{\theta}{2}\left(a_1^2\cos^2\varphi+a_2^2\sin^2\varphi\right)\right), \quad (3.21)$$

where K_{ho} and K_{lo} are normalization factors. The function defined in Eq. (3.20) differs from Eq. (3.11), in the sense that it generates longer atoms (slower decay) in the direction of Gaussian, but keeps the same sharp decay in the direction of its derivative. This leads to improved approximation of singularities of 3D objects. Samples of 3D atoms generated from the functions in Eq. (3.20) and Eq. (3.21) by motion, rotation and anisotropic scaling are given in Figure 3.11.

The discretization of dictionary parameters $(\tau,\nu,\psi,\alpha,\beta)$ is done similarly as for omnidirectional images. The position parameters τ and ν are uniformly distributed on their respective ranges, with resolution that is equal to the resolution of the signal. Sampling of the rotation parameter ψ is uniform on the interval $[-\pi,\pi)$, with the resolution equal to the signal's resolution. Higher resolution of orientations with respect to the dictionary used for omnidirectional images does not increase the complexity of MP since the correlation on the SO(3) group gives inner products for rotation samples at the signal's resolution. However, such a dictionary gives more freedom in approximating diverse curvatures present in 3D surfaces. Scaling parameters are discretized in the

---

[6] Reconstruction server is available at http://cgal.inria.fr/Reconstruction/submit.html.

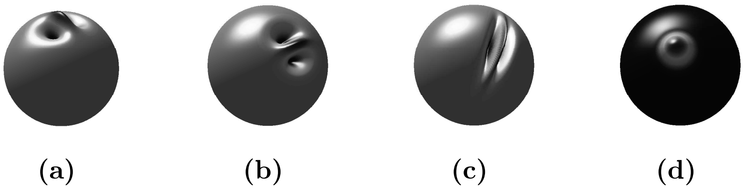


Figure 3.11: *Anisotropic atoms: a) on the North pole* $\tau = 0$, $\nu = 0$, $\psi = 0$, $\alpha = 8$, $\beta = 8$; *b)* $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{2}$, $\psi = 0$, $\alpha = 8$, $\beta = 8$; *c)* $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{2}$, $\psi = \frac{\pi}{4}$, $\alpha = 16$, $\beta = 4$; *d) Low frequency atom:* $\tau = \frac{\pi}{4}$, $\nu = \frac{\pi}{4}$, $\psi = \frac{\pi}{4}$, $\alpha = 8$, $\beta = 8$.

same manner as for the omnidirectional dictionary, except that we use all possible pairs of scales without constraining the β parameter, to achieve better approximation of 3D objects.

Two models are used in our experiments: Venus and Rabbit[7]. Venus satisfies the assumption of a simple genus-zero model, thus it is represented via one spherical function. Rabbit is not a simple model and we decompose it into three spheres separated by two parallel planes, one below the head and the other below the arms of the Rabbit. Each spherical function is obtained by interpolation of points in the point cloud corresponding to that part of the model. Afterwards, SMP is independently run on each of these three spheres and finally gathered into a single decomposition by ordering the atoms in decreasing order of their coefficient values. One additional atom parameter is introduced to denote the sphere that the atom belongs to. The quantization and entropy coding steps are the same as for the one-sphere decompositions. Finally, three spherical functions are reconstructed at the decoder, and point clouds are merged using a surface reconstruction algorithm, as explained in Section 3.5.1. The original and interpolated models are shown in Figure 3.12. It should be noted that the PSNR values obtained after interpolation are actually the upper bounds for the overall decoding quality, since the interpolated models are given as inputs to the 3D-SMP compression scheme. Interpolation error is expressed by the relative L^2 error and by PSNR [dB], as computed with the MESH algorithm [Asp02][8]. The relative L^2 error is a ratio of RMS (Root Mean Square Error), which measures the squared symmetric Hausdorff distance between two surfaces averaged over the first surface, relative to a bounding box diagonal d. PSNR (Peak Signal To Noise Ratio) for 3D meshes is thus expressed as:

$$PSNR[dB] = 20 \log \left(\frac{d}{RMS} \right) = 20 \log \left(\frac{1}{L^2} \right). \tag{3.22}$$

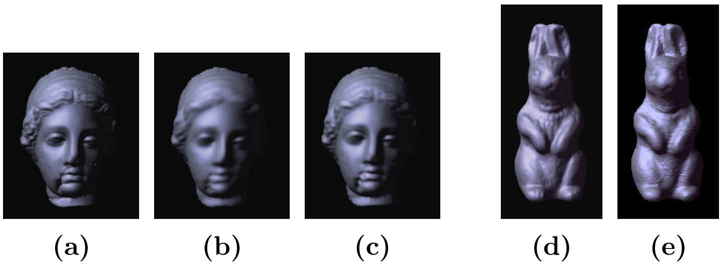


Figure 3.12: *a) Original Venus model; b) Interpolated Venus model at resolution* $2B = 128$, *PSNR=65.7983dB,* $L^2 = 5.1296 \cdot 10^{-4}$; *c) Interpolated Venus model at resolution* $2B = 256$, *PSNR=70.1930dB,* $L^2 = 3.09278 \cdot 10^{-4}$; *d) Original Rabbit model; e) Interpolated Rabbit model on three spheres at resolution* $2B = 128$, *PSNR=64.1790dB,* $L^2 = 6.18091 \cdot 10^{-4}$.

---

[7] The models have been downloaded from http://www.cyberware.com

[8] MESH is available at http://mesh.epfl.ch

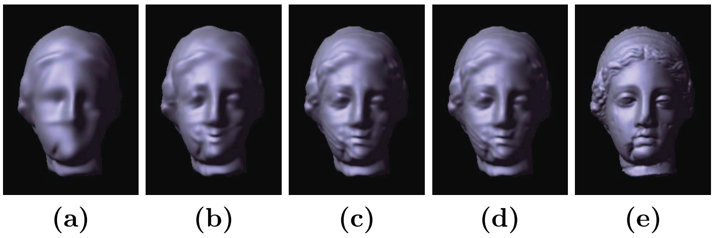


Figure 3.13: *Venus reconstructed after decoding (resolution $2B = 256$): a) 100 coefficients (0.58 KB); b) 200 coefficients (1.03 KB); c) 300 coefficients (1.5 KB); d) 400 coefficients (1.94 KB); e) Input model.*

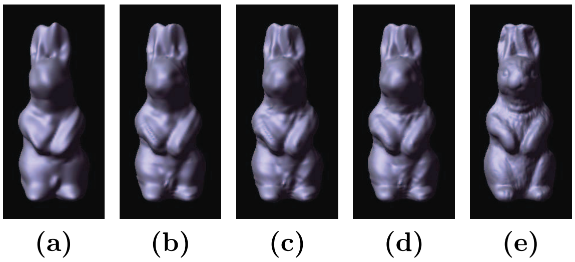


Figure 3.14: *Rabbit reconstructed after decoding (three spheres, resolution $2B = 128$): a) 100 coefficients (0.64 KB); b) 200 coefficients (1.05 KB); c) 300 coefficients (1.45 KB); d) 400 coefficients (1.84 KB); e) Input model.*

3.5.3 Compression performance evaluation

Venus and Rabbit models, as reconstructed by the decoder, are represented in Figure 3.13 and Figure 3.14, respectively, for different numbers of atoms. It can be seen that SMP rapidly captures the most important features of 3D models and progressively refines the representation with finer details. The type of coding artifacts is quite different than the degradations observed in mesh-based coders, and visually less annoying at low rate. It can be also seen that the gain in representation accuracy is less important at high rate. As expected, SMP is mostly efficient for low bit rate compression.

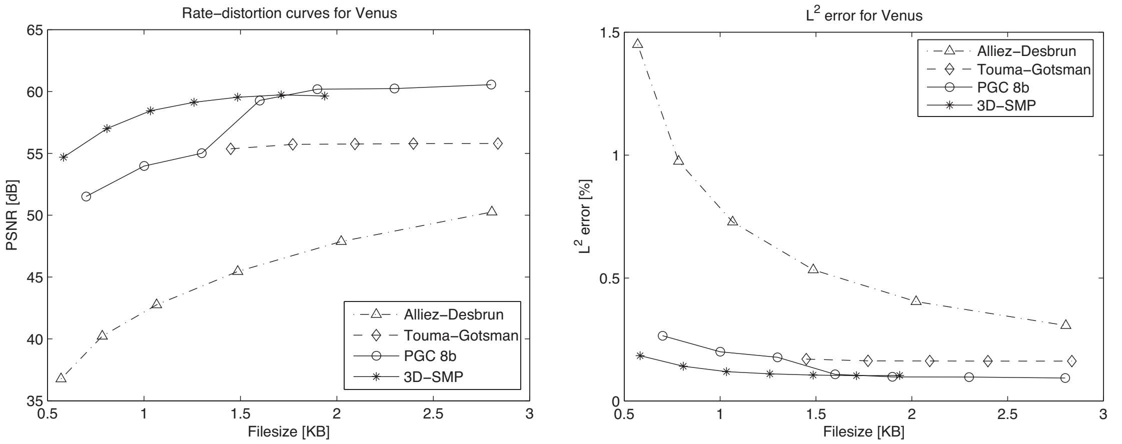
Figures 3.15 and 3.16 present the RD performance of the proposed 3D-SMP algorithm for Venus and Rabbit models, in terms of both PSNR (a) and L^2 error (b). Distortion values are evaluated with respect to the original models, meaning that they take into account both the interpolation and the decoding error. These figures compare the 3D-SMP encoder performance with the following state of the art encoders: (i) TG: Touma-Gotsman non-progressive coding [Tou98], (ii) Alliez-Desbrun progressive coding [All01], and (iii) PGC: Progressive coding scheme by Khodakovsky et al. [Kho00]. Due to the differences in input formats and coding approaches, we use the following approach to obtain fair performance comparisons between these four different methods. As PGC uses its own mesh format, input models are downloaded from the PGC website[9]. The base mesh for PGC is encoded using TG with 8 bits per vertex. Input models for TG and Alliez-Desbrun methods are those that have been used to obtain the interpolated models for 3D-SMP encoder[10]. However, these models have been decimated to 1400 faces (using the Qslim software[11]), in order to have a comparison in the same rate region. Different rates for the TG algorithm are obtained by changing the number of bits per vertex for encoding. Note that rate is actually given by filesize (the total number of bits) rather than in bits per vertex, since the proposed 3D-SMP coding

---

[9] http://www.multires.caltech.edu/software/pgc/

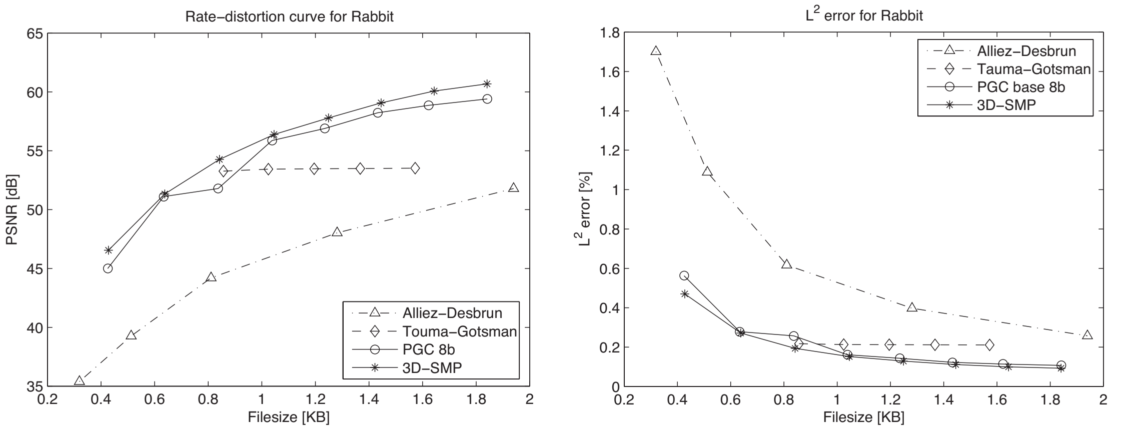
[10] http://www.cyberware.com/samples/index.html

[11] http://graphics.cs.uiuc.edu/ garland/software/qslim.html



(a) **(b)**

Figure 3.15: *Rate-distortion performance for the Venus model (resolution* $2B = 256$*) a) PSNR, b)* L^2 *error.*



(a) **(b)**

Figure 3.16: *Rate-distortion performance for the Rabbit model (resolution* $2B = 128$*) a) PSNR, b)* L^2 *error.*

scheme uses one single mesh (256 × 256 or 128 × 128 vertices, in the current implementation).

It can be seen that 3D-SMP significantly outperforms the state of the art compression methods TG and Alliez-Desbrun, as well as the PGC wavelet-based coder at low bit rate. 3D-SMP then tends to saturate towards high bite rate, as observed earlier. For the Venus model the performance is slightly better than for the Rabbit model. This behavior is actually expected since the resolution of the employed SMP for Rabbit is smaller. Therefore, we can certainly expect better performance for the Rabbit model at higher resolutions. It has to be noted also that the input model for SMP is an interpolated version of the original model, and this introduces a distortion that is independent of the coding method. Figure 3.17 shows a visual comparison of Venus encoded with 3D-SMP using 250 coefficients and resolution $2B = 256$, and encoded with PGC, for the same filesize of 1287B. It can be seen that both coders offer similar performance, but the coding artifacts are quite different. 3D-SMP coder generally provides a smoother approximation of the model, but fails in capturing highly textured regions like the hair, for example. The proposed encoder offers an interesting alternative to classical approaches, with excellent compression performance at low bit rate, and at the same time, an inherently progressive representation. Additionally, it offers a great flexibility in the stream construction, which can be advantageously exploited in adaptive applications, like view-dependent rendering [Tos06].

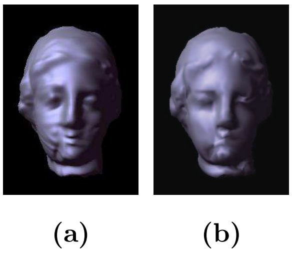


Figure 3.17: *Venus with a filesize of 1287B: a) MP; b) PGC.*

3.6 Conclusion

This chapter has presented a novel approach for representation and compression of spherical signals based on sparse signal approximations. Spherical signals are decomposed over a redundant dictionary of multi-dimensional atoms on the 2D sphere, which is built in order to efficiently capture the most prominent signal features. The first contribution of this chapter is the construction of a structured and geometric dictionary on the sphere. This dictionary represents one of the most important parts in the multi-view correlation modeling, which will be presented in the next chapter. The sparse representation of spherical signals has been effectively used within two additional contributions: development of a new compression method for omnidirectional images; and design of a new coder for 3D models. The proposed encoders have been shown to outperform state of the art progressive coders, especially at low bit rate. At the same time, they offer a truly progressive representation. Since redundant decompositions are mostly beneficial at low rate, the proposed scheme can offer an efficient coding solution for a base layer in scalable applications. Moreover, the MP-based compression of omnidirectional images can be exploited in the design of the distributed coder for multi-view images, as we will show in Chapter 5.

Chapter 4

Correlation in Multi-View Images

4.1 Introduction

The previous chapter has presented a method for representation and compression of single-view images, for the particular case of omnidirectional images mapped to the 2D sphere. From this single-view case we now move on to the problem of representing multi-view images captured by a visual sensor network. We have seen in Section 2.2.1 that state of the art multi-view technologies still need correlation models that lead to multi-view image representations suitable for simultaneous image compression and scene geometry extraction. These kind of image representations are a prerequisite for efficient representation of 3D scenes. To achieve this goal, a multi-view correlation model needs to relate image features that describe the same 3D objects in the scene across different views. Therefore, the model has to include multi-view geometry information defined by the epipolar geometry constraint.

This chapter proposes a novel geometry-based correlation model for the representation of scenes with distributed cameras. The main features of a 3D scene are likely to be present in the multiple correlated views of the scene, possibly transformed due to the geometry of the scene. We propose to capture these features by sparse image expansions with geometric atoms taken from a redundant dictionary of functions. The correlation model is then built on local geometric transforms between corresponding features taken from different views, where correspondences are defined based on shape and epipolar geometry constraints. Successful pairing of correlated atoms relies on the use of a structured dictionary that is invariant to a discrete set of local transforms like translation, rotation and scaling, or any combination of those. We apply this new correlation model to omnidirectional images that are mapped and processed on spherical manifolds. We then compute sparse image approximations on the sphere as presented in Chapter 3.

This novel correlation model represents the core concept of the proposed multi-view image modeling and we build upon it for different applications and problems. With respect to the state of the art correlation models, our model is local, can capture a variety of object transforms, and exploits epipolar geometry relations. This chapter defines the new correlation model and verifies its validity on multi-view spherical images.

4.2 State of the art

In Section 2.2.1 we have briefly discussed the main families of multi-view correlation models, which are pixel-based, block-based and perspective models. We have ordered these families according to the size of image components used for epipolar matching. In other words, these families are sorted from the most local (pixel-based) to global disparity models. In the following, we give a more detailed and comprehensive overview of existing multi-view correlation models.

Dense disparity estimation algorithms usually match pixels in two views that satisfy the epipolar geometry constraint, under intensity similarity and local depth smoothness constraints [Kol02]. Epipolar matching uses image pixels as features and ensures the consistency of depth in the local pixel's neighborhood by a regularization smoothness term. Therefore, these methods assume a pixel-based correlation model. Features that are more informative with respect to the local neighborhood of an observed point, such as scale-invariant SIFT features, can be also used for scene geometry estimation [Kos04, Liu06]. Although the similarity of SIFT features is used for feature matching, epipolar matching is performed on feature locations (pixels). Hence, we classify this method as pixel-based. Depth and geometry estimation from stereo or multi-view images has been an active research topic for many years, and existing algorithms give good estimates of the scene geometry from images captured by a camera network, usually without any compression. Still, images nowadays are rarely given in uncompressed form. Whether they are transmitted through a bandwidth constrained link, or they are stored on a limited capacity medium, images are almost always represented in a compressed form obtained by transform coding. However, disparity estimation from compressed images can be problematic. As mentioned in Section 2.1.1, transforms like DCT or orthogonal wavelet transform do not efficiently decompose an image into independent local geometric features, like edges or curvatures. Therefore, quantization of transform coefficients in the compressed image results in artifacts that damage the image geometry and deteriorate the performance of pixel-based disparity estimation algorithms [Thi09]. Another drawback of the pixel-based correlation model is that it is not efficient for joint multi-view compression. To achieve high compression gains, multi-view coding has to exploit both spatial correlation and correlation between multiple views. Namely, for two views the joint encoder would have to send one image compressed by transform coding, the disparity map, and the prediction residue for the second image, where prediction is obtained by applying the disparity map to the first image. However, encoding of the disparity values for all pixels would require a lot of bits.

The most common way to combine the spatial and multi-view geometry correlation is the block-based correlation model. This approach has been taken by the joint video team in effort to design the multi-view coding (MVC) extension of the video coding standard H.264/AVC [Mer06, Mer07, Mar06b]. However, block-based disparity estimation is valid only for an in-line arrangement of cameras with parallel optical axes, because of the need for image rectification. Moreover, it is implicitly assumed that all pixels in a block have the same disparity values, which is rarely the case. Although block-based prediction performs well for video coding due to the translational motion model, it is not advantageous for disparity compensation since multi-view correlation is far from being purely translational. This was confirmed by the performance analysis of the MVC [Mer07], showing that inter-view prediction brings marginal (up to 10%) bitrate savings with respect to the temporal prediction. Block-based disparity estimation has been also proposed for a wavelet-based multi-view coding scheme [Gar06]. In the case of omnidirectional images, accurately modeling the correlation with a block-based disparity estimation method is not possible. This is mainly because a viewpoint change typically leads to changes in scale of the image projections of 3D scene features, thus violating the translational model assumption.

Global disparity correlation models for multiple views have also been proposed in the context of multi-view coding. This approach involves the estimation of the homography matrix that maps one view to another, under perspective transformation [Oua06b, Guo04]. Since all pixels in the image undergo the same transformation, this model is applicable only in special cases where the scene can be approximated with a planar surface. The perspective transformation model can be even more simplified to an affine transform correlation model [Guo06]. Finally, global disparity prediction can be combined with block-based estimation to yield better performance, but the above mentioned limitations of both approaches still remain [Yan06]. In the case of an omnidirectional camera network, modeling the multi-view correlation with a global disparity model is not acceptable, since the surrounding scene cannot be approximated as planar. In particular, due to the wide field of view of omnidirectional cameras, even small baseline distances result in many transforms that differ locally.

The multi-view correlation model proposed in this chapter relates local image features across different views under various local transforms. Matching of transformed features is done by utilizing epipolar geometry and shape similarity constraints. Therefore, in our categorization of models this transform-based model lies in between the most local (pixel-based) models and global models. To the best of our knowledge, this is the first work that proposes to couple sparse image representations with geometric atoms and the epipolar geometry constraint to model multi-view correlation. It also differs from the related work in the types of local transforms that the model can represent. Namely, our model can cope with translations, rotations and anisotropic scaling of local image features.

4.3 Multi-view geometric correlation model

Correlation between multi-view images arises from geometric constraints on the objects in the scene due to viewpoint change, and can be simply described by local changes of image components that represent the objects in the scene. In other words, if we decompose each image into features, we can assume that the most prominent components are present in all images with high probability, possibly with some local transforms. However, as we mentioned in Section 4.2 image decompositions by common orthogonal transforms like the wavelet or DCT do not capture the scene objects and their geometry. On the other hand, sparse image approximations with overcomplete dictionaries of basis vectors (atoms) are capable of capturing the image structure and geometry using only few basis vectors, while offering excellent approximation performance [Fig06]. One of the most important advantages of sparse approximations is the flexibility in the design of the overcomplete dictionary. When the dictionary is built on geometric functions with local support, the sparse image decomposition results in a set of meaningful geometric features that represent the visual information of the scene. The comparison of these features in different views permits us to estimate the geometry of the scene and the correlation between views. This correlation between multi-view images is driven by local transforms of sparse image components in different views that represent the same component in the 3D scene.

As described in Section 2.1.2, given a certain basis or redundant dictionary of atoms $\mathcal{D} = \{\phi_k\}, k = 1, ..., N$, in a Hilbert space, every image y can be represented as:

$$y = \mathbf{\Phi}_I c + \eta = \sum_{k \in I} a_k \phi_k + \eta, \tag{4.1}$$

where c is the vector of significant coefficients, I labels the set of atoms $\{\phi_k\}_{k \in I}$ participating in the representation and η is the approximation error. Matrix $\mathbf{\Phi}_I$ contains the participating atoms as columns. We are now interested in defining a correlation model between sparse approximations of two correlated multi-view images[1]:

$$\begin{aligned} y_1 &= \mathbf{\Phi}_{I_1} c_1 + \eta_1, \\ y_2 &= \mathbf{\Phi}_{I_2} c_2 + \eta_2. \end{aligned} \tag{4.2}$$

Since y_1 and y_2 capture the same 3D scene, there exists a subset of atoms indexed respectively by $J_1 \in I_1$ and $J_2 \in I_2$ that represent image projections of the same object in the scene. We assume that these atoms are correlated, possibly under some local geometric transforms. Let $F(\phi)$ denote the transform of an atom ϕ between two image decompositions. This transform typically results from a change of camera viewpoint. Therefore, the correlation between images can be modeled as a set of transforms F_i between corresponding atoms in sets indexed by J_1 and J_2. The approximation of the image y_2 can be rewritten as the sum of the contributions of transformed atoms from J_1, atoms in $I_2 \setminus J_2$, and noise η_2:

$$y_2 = \sum_{i \in J_1} a_{2,i} F_i(\phi_i) + \sum_{k \in I_2 \setminus J_2} a_{2,k} \phi_k + \eta_2. \tag{4.3}$$

---

[1] We take two images for the sake of clarity, but the correlation model that we develop can be generalized to any number of images.

This model is independent of the sparse approximation algorithm used for image decomposition and generic with respect to the overcomplete dictionary selection. However, we choose a dictionary built on locally defined geometric atoms that can approximate multidimensional discontinuities like edges. These represent important information about the scene geometry.

The main challenge in the proposed model is to define the transforms F_i in Eq. (4.3) that relate corresponding atoms in sparse decompositions of multi-view images. Due to camera viewpoint changes, various types of transforms are introduced in the image projective space. Most of these transforms can be represented by the 2D similarity group elements, which include 2D translation, rotation and isotropic scaling of the image features. We also consider anisotropic scaling to further expand the space of possible transforms. In order to efficiently capture transforms between sparse image components, we propose to use a structured redundant dictionary of atoms for the image representation. Atoms in the structured dictionary are derived from a single waveform that undergoes rotation, translation and anisotropic scaling, as described in Chapter 3. Let us now observe the properties of this dictionary. We first note that the dictionary is *translation invariant.*

Definition 4.3.1. *[Mal08] Let $\mathcal{D} = \{g_\gamma\}_{\gamma\in\Gamma}$ be a dictionary of $K > N$ unit norm vectors in $\mathbb{C}^N$. A dictionary is translation invariant if for any $g_\gamma[n] \in \mathcal{D}$ then $g_\gamma[n-k] \in \mathcal{D}$ for $0 \leqslant k < N$.*

Therefore, translating an atom in the dictionary results in another atom in the dictionary. Note that the atoms here are discrete functions and that translation of one atom results in an atom shifted by an integer value of the sampling step (in definition 4.3.1 the sampling step is taken as 1). Hence, a translation invariant dictionary has to include atoms obtained by translating the generating function to all integers of the sampling step, from zero to the resolution N of the signal. Since the dictionary constructed in Chapter 3 satisfies this condition, it is translation invariant[2].

An interesting property shown by Davis et al. is that Matching Pursuit is translation invariant when calculated on a translation invariant dictionary [Dav97]. This means that Matching Pursuit coefficients will be the same for a given signal under different translations. This is crucial in practical applications where a pattern needs to be matched with its translated versions, as it is the case in multi-view matching. On the other hand, signal decompositions in orthogonal basis, like for example the orthogonal wavelet basis, do not have the translation invariance property. However, such transforms can be made translation-invariant by adding the redundancy corresponding to all translations of the generating function, leading to overcomplete dictionaries (e.g., translation-invariant dyadic wavelet dictionary).

Translation invariance can be generalized as an invariance to any group action [Dav97]. Besides translations, our dictionary is constructed by also applying a set of different rotations and anisotropic scalings to the generating function. Therefore, we extend the dictionary's invariance property to the rotation and anisotropic scaling. More formally, given a generating function g defined in the Hilbert space, the dictionary $\mathcal{D} = \{\phi_k\} = \{g_\gamma\}_{\gamma\in\Gamma}$ is constructed by changing the atom index $\gamma \in \Gamma$ that defines rotation, translation and scaling parameters applied to the generating function g. This is equivalent to applying a unitary operator $U(\gamma)$ to the generating function g, i.e.: $g_\gamma = U(\gamma)g$. When the dictionary is defined this way, the transform of one atom g_{γ_i} to another atom g_{γ_j} reduces to a transform of its parameters, i.e.,

$$g_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma')g_{\gamma_i} = U(\gamma' \circ \gamma_i)g. \tag{4.4}$$

Hence, the transformation of an atom by translation, rotation and anisotropic scaling (defined by index γ') results in another atom in the same dictionary. As we work with discrete parameters in γ, the transform group actions are limited to the set that includes integer multiples of the sampling step for each parameter. Therefore, the dictionary is invariant with respect to any transform action from this set. We say that the dictionary is *transform invariant in the discrete domain.* This property makes our dictionary more advantageous for identifying correlated image features from multiple views, compared to dictionaries that are only translation-invariant.

---

[2] The dictionary in Chapter 3 is defined on the sphere; here translation refers to the motion τ and ν along θ and φ angles, respectively.

The correlation model given in Eq. (4.3) does not put any assumption on the type of cameras used for multi-view image acquisition. It can be applied to planar or omnidirectional multi-view images by introducing epipolar geometry constraints that are defined accordingly. In the next section, we define multi-view transforms that satisfy epipolar geometry for omnidirectional images in particular, due to their advantages for compact 3D scene representation.

4.4 Geometric transforms in omnidirectional images

4.4.1 Multi-view image expansions on the sphere

We focus now on omnidirectional images, and describe in more detail the specific correlation model for images that can be mapped on spheres. We use omnidirectional cameras that are constructed by placing a parabolic mirror in front of a camera with an orthographically projecting lens as depicted in Figure 2.7(a). As these images can be precisely mapped on the sphere through inverse stereographic projection (see Section 2.3.1), we further use a dictionary of atoms on the 2D unit sphere constructed in Chapter 3. The generating function g is hence defined in the space of square-integrable functions on the 2D unit sphere S^2, $g(\theta, \varphi) \in L^2(S^2)$, while the dictionary is built by changing the atom indexes $\gamma = (\tau, \nu, \psi, \alpha, \beta) \in \Gamma$. The triplet (τ, ν, ψ) represents Euler angles that respectively describe the motion of the atom on the sphere by angles τ and ν, and the rotation of the atom around its axis with an angle ψ, while α and β represent anisotropic scaling factors.

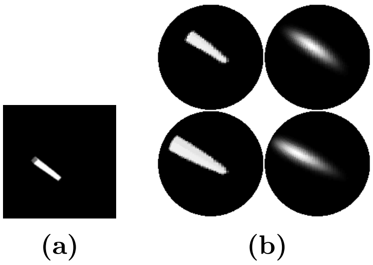


Figure 4.1: *Example of atom transforms in approximation of two views of a 3D scene: a) a simple 3D scene b) first column: two captured spherical images of the scene; second column: sparse approximations of the two views with one Gaussian atom per view. The atom in the approximation of the second view is related to the atom in the first view by the following transform:* $\tau_2 - \tau_1 = \pi/128$, $\nu_2 - \nu_1 = -12\pi/128$, $\psi_2 - \psi_1 = -\pi/16$, $\alpha_2/\alpha_1 = 1$, $\beta_2/\beta_1 = 0.8$.

We are interested in finding correspondences between atoms that represent images y_1 and y_2, generated by two omnidirectional cameras capturing the same scene. For the sake of clarity, let $\{g_\gamma\}_{\gamma \in \Gamma}$ and $\{h_\gamma\}_{\gamma \in \Gamma}$ denote the set of functions used for the expansions of images y_1 and y_2, respectively. The same dictionary is used for both images, so that two corresponding atoms g_{γ_i} and h_{γ_j} in images y_1 and y_2 are linked by a simple transform of the atom parameters, and Eq. (4.4) can be rewritten as

$$h_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma')g_{\gamma_i} = U(\gamma' \circ \gamma_i)g. \tag{4.5}$$

The subset of transforms $V_i^0 = \{\gamma' | h_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma')g_{\gamma_i}\}$ allows us to relate g_{γ_i} to the atoms h_{γ_j} in the expansion of y_2. Figure 4.1 depicts an example of a simple 3D scene captured by two omnidirectional cameras and shows how their sparse approximations can be linked with a transform of the atom parameters. However, not all transforms in V_i^0 are possible in multi-view correlated images. The set of possible transforms can be greatly reduced by identifying two constraints between corresponding atoms, namely the *shape similarity* constraint and the *epipolar geometry* constraint.

4.4.2 Shape and epipolar constraints

First, we assume that the change of viewpoint on a 3D object results in a limited difference between shapes of corresponding atoms since they represent the same object in the scene. Therefore, we can restrict the set of possible transforms in V_i^0 by the shape similarity constraints between candidate atoms. From the set of atom parameters γ, the last three parameters (ψ, α, β) describe the atom shape (its rotation and scaling), and therefore they are taken into account for the shape similarity constraint. We measure the similarity or coherence of atoms by the inner product $\mu(i,j) = |\langle g_{\gamma_i}, h_{\gamma_j} \rangle|$ between centered atoms (at the same position (τ, ν)), and we impose a minimal coherence between candidate atoms, i.e., $\mu(i,j) > s$. This defines a set of possible transforms $V_i^\mu \subseteq V_i^0$ with respect to the atom shape, as:

$$V_i^\mu = \{\gamma' | h_{\gamma_j} = U(\gamma') g_{\gamma_i}, \mu(i,j) > s\}. \tag{4.6}$$

Equivalently, the set of atoms h_{γ_j} in y_2 that are possible transformed versions of the atom g_{γ_i} is denoted as the *shape candidates set*. It is defined by the set of atoms indexes $\Gamma_i^\mu \subset \Gamma$, with

$$\Gamma_i^\mu = \{\gamma_j | h_{\gamma_j} = U(\gamma') g_{\gamma_i}, \gamma' \in V_i^\mu\}. \tag{4.7}$$

Second, pairs of atoms that correspond to the same 3D points have to satisfy epipolar geometry constraints, which represent one of the fundamental relations in multi-view analysis [Ma 04]. As explained in Section 2.2.1, the epipolar constraint defines the relation between 3D point projections $(\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^3)$ on two cameras, as:

$$\mathbf{z}_2^T \hat{\mathbf{T}} \mathbf{R} \mathbf{z}_1 = 0, \tag{4.8}$$

where $\mathbf{R}$ and $\mathbf{T}$ are the rotation and translation matrices of one camera frame with respect to the other, and $\hat{\mathbf{T}}$ is obtained by representing the cross product of $\mathbf{T}$ with $\mathbf{R}\mathbf{z}_1$ as a matrix multiplication, i.e., $\hat{\mathbf{T}}\mathbf{R}\mathbf{z}_1 = \mathbf{T} \times \mathbf{R}\mathbf{z}_1$. The set of possible transforms between atoms from different views is therefore further reduced to the transforms that respect epipolar constraints between the atom g_{γ_i} in y_1 and the candidate atoms h_{γ_j} in y_2. The constraint given in Eq. (4.8) is rarely exactly satisfied for corresponding pixels or areas in two multi-view images, and the decision on the epipolar matching of two correspondences is commonly taken when their epipolar distance is smaller than a certain threshold κ.

By imposing the epipolar constraint on atoms in V_i^0, we define the set $V_i^E \subseteq V_i^0$ of possible transforms of atom g_{γ_i} as:

$$V_i^E = \{\gamma' | h_{\gamma_j} = U(\gamma') g_{\gamma_i}, d_{EA}(g_{\gamma_i}, h_{\gamma_j}) < \kappa\}, \tag{4.9}$$

where $d_{EA}(g_{\gamma_i}, h_{\gamma_j})$ denotes the epipolar distance between atoms g_{γ_i} and h_{γ_j}. This distance measures how much atoms g_{γ_i} and h_{γ_j} deviate from the perfect epipolar matching given by Eq. (4.8) and it will be formally defined in Section 4.5. Similarly, we define a set of candidate atoms in y_2, called the *epipolar candidates set*, whose indexes belong to $\Gamma_i^E \subset \Gamma$, with:

$$\Gamma_i^E = \{\gamma_j | h_{\gamma_j} = U(\gamma') g_{\gamma_i}, \gamma' \in V_i^E\}. \tag{4.10}$$

A graphical interpretation of the epipolar constraint for spherical images is shown in Figure 4.2, where we denote as S_1 and S_2 the two unit spheres corresponding to camera projection surfaces. A given atom g_{γ_i} in y_1, on the sphere S_1, can be a projection of infinitely many different 3D objects, at different scales and distances from S_1. We show an example of several different objects whose projection on S_1 is g_{γ_i}, and whose projections on S_2 are h_{γ_j}. Due to the epipolar constraints, the atoms h_{γ_j} are positioned on the part of a great circle $\mathcal{C}_i$ obtained by projecting the ray L_i on the sphere S_2. This ray originates from the center of camera 1 and passes through the atom g_{γ_i} on the sphere S_1.

Finally, we combine the epipolar and shape similarity constraints to define the set of possible transforms for atom g_{γ_i}, as $V_i = V_i^E \cap V_i^\mu$. Similarly, we denote the set of possible parameters of the transformed atom in y_2 as $\Gamma_i = \Gamma_i^E \cap \Gamma_i^\mu$. Given the set Γ_i of possible atom parameters in y_2 corresponding to the atom g_{γ_i} in y_1, the correspondence h_{γ_j} in y_2 can be uniquely defined with high probability under the assumption that the decomposition of y_2 is sufficiently sparse.

4.5 Disparity map estimation from atom transforms

The local transforms between geometric atoms are now used to estimate the correlation between pixels in multiview images, as represented by a *disparity map*. A disparity map typically allows for view interpolation under epipolar constraints. It is defined as the point-wise correlation between multi-view images, which relates a point $\mathbf{z}_1$ on the image y_1 to a point $\mathbf{z}_2$ on y_2, such that the epipolar constraint from Eq. (4.8) is satisfied. Such dense disparity mapping is most commonly estimated based on the pixel-wise correlation between rectified stereo images, or by block-based matching. The performance of these approaches unfortunately deteriorates with the decrease of the image quality, for example when images are compressed at very low bit rates [Thi09]. The disparity map can also be estimated by identifying corresponding feature points in multi-view images and relating them with a cross-correlation similarity measure [Kon00]. However, the cross-correlation measure [Har00] is not rotationally invariant and it fails to capture rotations of patterns between views. Since our correlation model relates similar atoms under different scale and rotation transforms, it represents a feature similarity measure that is invariant with respect to rotation and scaling. Therefore, each pair of corresponding atoms can give a reliable estimate of the disparity map, obtained by the atom transform. Since atoms and their transforms are spatially localized, each corresponding atom pair gives an estimate of the local disparity map for the image area covered by the atom's spatial support. Combining local disparity maps from each correspondence pair leads to the estimation of the global disparity map for the whole 3D scene. Moreover, this estimation can be performed with images that are encoded at a very low bit rate. We describe here the estimation of the local disparity map from the atom transforms, which permits us to define a measure of the estimation error that can be used to refine the atom pairing process.

Let us consider a pair of corresponding atoms $(g_{\gamma_i}, h_{\gamma_j})$ in two images. We want to find a mapping of each point on g_{γ_i} to its corresponding point on h_{γ_j}. Since this mapping is point-wise, we need to define g_{γ_i} in the discrete space, i.e., on the spherical grid $\mathcal{G}_1$. Then, the disparity mapping translates to the grid deformation induced by the local transform between g_{γ_i} and h_{γ_j}, denoted as $\mathcal{F}\{\mathcal{G}_1\}$. Let P_1 be a point on $\mathcal{G}_1$, given in Euclidean coordinates as $\mathbf{z}_1$. Similarly, let P_2 be a point on $\mathcal{G}_2$, given in Euclidean coordinates as $\mathbf{z}_2$, which is obtained by applying the grid transform $\mathcal{F}$ to P_1. Further, let $\gamma_i = (\tau_i, \nu_i, \psi_i, \alpha_i, \beta_i)$ and $\gamma_j = (\tau_j, \nu_j, \psi_j, \alpha_j, \beta_j)$. The grid transform $\mathcal{G}_2 = \mathcal{F}\{\mathcal{G}_1\}$ includes two transforms:

1. the transform of the motion of atom g_{γ_i}, given by Euler angles (τ_i, ν_i, ψ_i), into the motion of atom h_{γ_j}, given by Euler angles (τ_j, ν_j, ψ_j),

2. the transform of anisotropic scaling of atom g_{γ_i}, given by the pair of scales (α_i, β_i), into the anisotropic scaling of atom h_{γ_j}, given by the pair of scales (α_j, β_j).

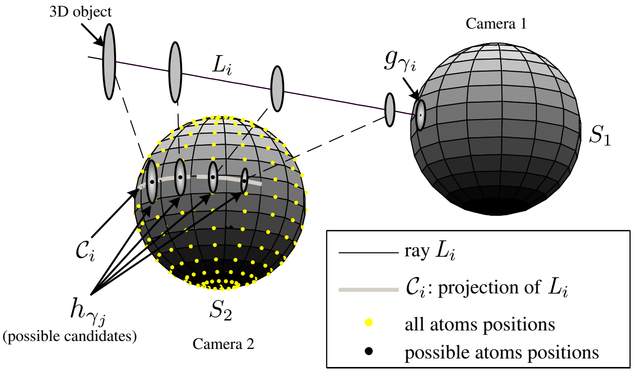


Figure 4.2: *Selection of positions of atoms that satisfy epipolar constraints.*

By combining these two transforms, the point $\mathbf{z}_2$ can be written as:

$$\mathbf{z}_2 = \mathbf{R}_{\gamma_j}^{-1} \cdot \zeta(\mathbf{R}_{\gamma_i} \cdot \mathbf{z}_1), \tag{4.11}$$

where $\mathbf{R}_{\gamma_i}$ and $\mathbf{R}_{\gamma_j}$ are rotation matrices given by Euler angles (τ_i, ν_i, ψ_i) and (τ_j, ν_j, ψ_j), respectively, and $\zeta(\cdot)$ defines the grid transform due to anisotropic scaling. Since the anisotropic scaling of atoms on the sphere is performed on the plane tangent to the North Pole by stereographically projecting the atom, the grid $\mathcal{G}_1$ is first rotated such that the North Pole is aligned with the center of atom g_{γ_i}, then deformed with respect to anisotropic scaling, and finally rotated back with the rotation matrix of atom h_{γ_j}.

As defined in Section 3.2.1, the stereographic projection at the North Pole projects a point (θ, φ) on the sphere to a point (x, y) on the plane tangent to the North pole, given by:

$$x + jy = \rho e^{j\varphi} = 2tan\left(\frac{\theta}{2}\right) e^{j\varphi}. \tag{4.12}$$

Let now (θ_1, φ_1) and (θ_2, φ_2) denote the spherical coordinates of points P_1 and P_2, respectively (the point belongs to the unit sphere and $r = 1$ is assumed). Under stereographic projection, the transform of point (θ_1, φ_1) on grid $\mathcal{G}_1$ to point (θ_2, φ_2) on grid $\mathcal{G}_2$ due to anisotropic scaling can be obtained by scaling the stereographic projection of (θ_1, φ_1) with $1/\alpha_j$ and $1/\beta_j$, in the following way:

$$\begin{aligned} x_2 &= \rho_2 \cos\varphi_2 = \frac{1}{\alpha_j}\alpha_i x_1 = \frac{\alpha_i}{\alpha_j}\rho_1 \cos\varphi_1, \\ y_2 &= \rho_2 \sin\varphi_2 = \frac{1}{\beta_j}\beta_i y_1 = \frac{\beta_i}{\beta_j}\rho_1 \sin\varphi_1, \end{aligned} \tag{4.13}$$

where $\rho_2 = 2\tan(\theta_2/2)$ and $\rho_1 = 2\tan(\theta_1/2)$. By solving the system of Eq. (4.13) for θ_2 and φ_2, we get:

$$\begin{aligned} \varphi_2 &= \zeta_p(\varphi_1) = \arctan\left(\frac{\alpha_j \beta_i \sin\varphi_1}{\alpha_i \beta_j \cos\varphi_1}\right), \\ \theta_2 &= \zeta_t(\theta_1, \varphi_1, \varphi_2) \\ &= 2\arctan\left(\tan\frac{\theta_1}{2}\sqrt{\frac{\alpha_i^2\cos^2\varphi_1 + \beta_i^2\sin^2\varphi_1}{\alpha_j^2\cos^2\varphi_2 + \beta_j^2\sin^2\varphi_2}}\right). \end{aligned} \tag{4.14}$$

We can therefore define the function $\zeta(\cdot)$ as a pair of transforms $\zeta_p(\varphi_1)$ and $\zeta_t(\theta_1, \varphi_1, \zeta_p(\varphi_1))$ followed by the transform of spherical coordinates (θ_2, φ_2) to Euclidean coordinates $\mathbf{z}_2$. The relation given in Eq. (4.11) is now completely defined, based on the parameters of corresponding atoms in two images. When the transform is applied to points lying on the spatial support of the atoms, it forms the local disparity map between the correlated views.

Finally, we define the *Symmetric epipolar atom distance* in order to quantify the mismatch between two corresponding atoms g_{γ_i} and h_{γ_j} related by the disparity map. The symmetric epipolar atom distance measures how much the atom pair g_{γ_i} and h_{γ_j} deviates from the perfect epipolar matching given in the correlation model of Eq. (4.9). It is evaluated as the weighted average of the symmetric epipolar distance of all pairs of points given by the disparity map:

$$d_{EA}(g_{\gamma_i}, h_{\gamma_j}) = \sum_{\mathbf{z}_1 \in \mathcal{G}_1} w_{\gamma_i}(\mathbf{z}_1) d_{SE}(\mathbf{z}_1, \mathbf{z}_2). \tag{4.15}$$

The symmetric epipolar distance $d_{SE}(\mathbf{z}_1, \mathbf{z}_2)$ between points $\mathbf{z}_1$ and $\mathbf{z}_2$ related by the disparity map is defined as [Har00]:

$$d_{SE}(\mathbf{z}_1, \mathbf{z}_2) = \sqrt{d(\mathbf{z}_1, \mathcal{C}_{\mathbf{z}_2}) + d(\mathbf{z}_2, \mathcal{C}_{\mathbf{z}_1})}, \tag{4.16}$$

where $d(\mathbf{z}_1, \mathcal{C}_{\mathbf{z}_2})$ denotes the Euclidean distance of the point $\mathbf{z}_1$ to the epipolar circle $\mathcal{C}_{\mathbf{z}_2}$ corresponding to point $\mathbf{z}_2$ (see Figure 4.3 for an illustration of this distance). The weight w_{γ_i} is a normalized weight function that prioritizes the points where the atom g_{γ_i} has higher value. The goal of this function is to give more importance to the disparity mismatch of points that lie closer to the geometric component captured by the atom (typically edges). One example could be a two-dimensional Gaussian weight function, anisotropically scaled and oriented, which covers the spatial support of the atom g_{γ_i}. If the overcomplete dictionary is composed of Gaussian atoms, the weight function is equal to the atom itself. We use 2D Gaussian weight function in the rest of this thesis.

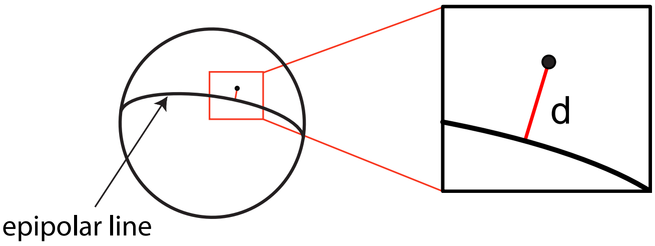
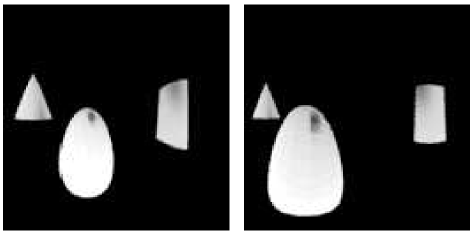


Figure 4.3: *Example of the distance* $d(\mathbf{z}_1, \mathcal{C}_{\mathbf{z}_2})$ *that is equal to the Euclidean distance of a point to the epipolar line, and denoted as d.*

4.6 Evaluation of the proposed correlation model

In order to confirm the validity of the proposed multi-view correlation model, we evaluate the shape correlation defined by the inner product, and the epipolar correlation defined by the symmetric epipolar atom distance d_{EA}, for all pairs of atoms from two views of a 3D scene. We have tested two scenes: a synthetic scene that contains three simple shapes (a cone, a sphere and a box) called the Shapes scene, and a natural Lab scene. Two spherical images of resolution $2B = 64$ have been taken of the Shapes scene from two spherical cameras with a relative rotation $\mathbf{R} = \mathbf{I}$ ($\mathbf{I}$ represents the identity matrix) and translation $\mathbf{T} = [1.2 \quad 0 \quad 0]^{\mathsf{T}}$. This has been done using the yafray tracing software[3], and the unfolded images are shown in Figure 4.4. Sparse approximations of the spherical images have been obtained using the Matching Pursuit algorithm on the sphere (see Chapter 3) with a Gaussian dictionary based on the generating function in Eq. (3.12).



(a) **(b)**

Figure 4.4: *Shapes scene, original images of resolution* $2B = 64$*: a) view 1, b) view 2.*

We first observe the sparse approximations of two views of the Shapes scene shown in Figure 4.5. Due to the geometric property of the dictionary, each atom corresponds to one of the objects in the scene, or to the background. For the first six atoms in the SMP decomposition, their semantic

---

[3] http://yafray.org/

meaning is given in Table 4.1. Since the sparse decompositions of the two views have been performed independently, the order of atoms and scene features in two views is not necessarily the same. We can see that within the first six atoms, there are three atom correspondence pairs: (1,1), (2,2) and (3,6), i.e., the pairs of atoms from two views that correspond to the same 3D scene feature.

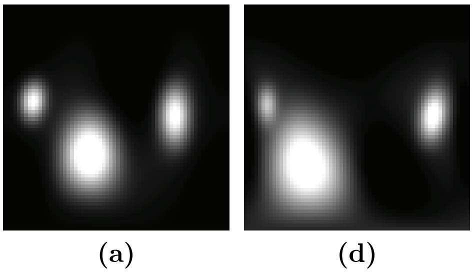

(a) (d)

Figure 4.5: *Approximated images of the Shapes scene: a) view 1 with six atoms, b) view 2 with six atoms.*

Table 4.1: *Shapes scene: Semantic meaning of atoms in sparse representations of two views.*

atom number	object, view 1	object, view 2
1	sphere	sphere
2	box	box
3	cone	background
4	background	background
5	background	background
6	background	cone

The shape similarity measured by the inner product between centered atoms is shown in Table 4.2. For the corresponding atoms the inner product is high (close to 1), thus confirming our shape correlation model. However, we can see that other atom pairs can also give high inner products, as some atom shapes can be similar although they represent different objects. This makes us conclude that the shape similarity measure is not a sufficient criterion for determining atom correspondences and should be used together with other criteria, like the epipolar correlation constraint.

Table 4.2: *Shapes Scene: Inner products for atom pairs from two views. Values of the inner products for the corresponding atom pairs are displayed in bold.*

	atoms in view 2					
atoms in view 1	1	2	3	4	5	6
1	**0.9748**	0.9045	0.8767	0.9112	0.8745	0.7555
2	0.8664	**0.9876**	0.7266	0.7734	0.8527	0.9063
3	0.7064	0.9716	0.5608	0.6053	0.7059	**0.9873**
4	0.9510	0.6688	1.0000	0.9902	0.8521	0.5108
5	0.9510	0.6688	1.0000	0.9902	0.8521	0.5108
6	0.9697	0.7692	0.9604	0.9566	0.8231	0.6000

The symmetric epipolar atom distance has been evaluated for all pairs of atoms in two views of the Shapes scene. The resulting values, averaged by the total number of pixels, are given in Table 4.3. We can see that the d_{EA} values are substantially smaller for the atom pairs that represent the same scene features, which confirms the validity of the epipolar correlation constraint. For

Table 4.3: *Shapes scene: Average symmetric epipolar atom distance for atom pairs from two views. Values of the epipolar distance for corresponding atom pairs are displayed in bold.*

	atoms in view 2					
atoms in view 1	1	2	3	4	5	6
1	**0.0006**	0.0122	0.0116	0.0088	0.0066	0.0058
2	0.0084	**0.0020**	0.0069	0.0060	0.0041	0.0074
3	0.0051	0.0058	0.0055	0.0064	0.0045	**0.0013**
4	0.0228	0.0167	0.0063	0.0206	0.0139	0.0150
5	0.0234	0.0250	0.0341	0.0222	0.0060	0.0221
6	0.0254	0.0199	0.0199	0.0270	0.0203	0.0153

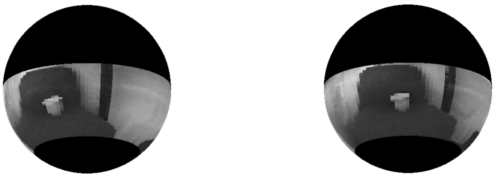

(a) (b)

Figure 4.6: *Lab Scene, spherical original images of resolution* $2B = 128$*: a) view 1; b) view 2.*

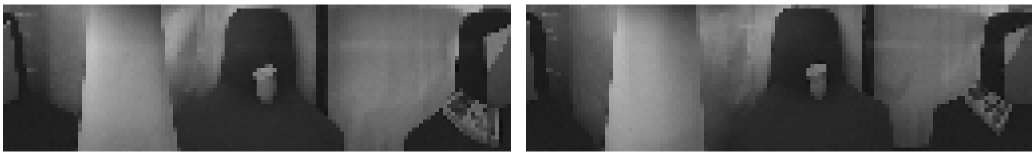

(a) (b)

Figure 4.7: *Lab Scene, unfolded original images of resolution* $2B = 128$*: a) view 1; b) view 2.*

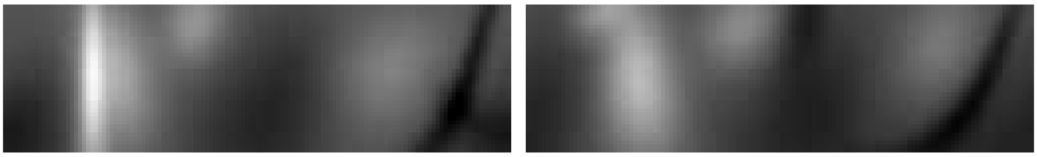

(a) (b)

Figure 4.8: *Lab Scene: Approximated images: a) view 1 with eight atoms, b) view 2 with eight atoms.*

non-corresponding atom pairs, the distance d_{EA} is at least twice as large as for the corresponding atom pairs.

Lab scene has been captured using two omnidirectional cameras, and their output has been mapped to two spherical images of resolution $2B = 128$, as explained in Section 2.3.1. For the Lab scene we have used the approximate camera pose for the specific linear camera arrangement[4]: $\mathbf{R} \approx \mathbf{I}$ and $\mathbf{T} \approx [1\ \ 0\ \ 0]^{\mathsf{T}}$. Spherical Lab images and their unfolded versions are shown in Figure 4.6 and Figure 4.7, respectively. The approximated images using eight atoms per MP decomposition are shown in Figure 4.8.

The semantic meaning of atoms, given in Table 4.4, shows four corresponding atom pairs:

---

[4] Since we do not perform depth estimation, we can use the camera translation vector which is correct up to a scale factor, hence the value 1 for the translation on x-axis.

(1,1), (2,2), (3,3) and (6,8). The results for the shape correlation constraint are shown in Table 4.5, leading us to the same conclusion as for the Shapes scene on the validity of the shape correlation constraint. Even though the extrinsic parameters between two cameras have been coarsely estimated, the symmetric epipolar atom distance is much smaller for the corresponding atoms than for most of other pairs, as shown in Table 4.6. We can see that there are some atom pairs that give smaller epipolar distance than the correct correspondences, such as (5,1), (5,2) and (6,3). Those pairs however do not give a good shape matching, as shown in Table 4.5. Therefore, we can conclude that both geometric constraints are important and should be used together for the correct matching of correlated atoms.

Table 4.4: *Lab scene: Semantic meaning of atoms in sparse representations of two views.*

atoms	object, view 1	object, view 2
1	white pillar	white pillar
2	side chair	side chair
3	central chair	central chair
4	background	background
5	background	background
6	side chair boundary	background
7	book	white pillar top
8	background	side chair boundary

Table 4.5: *Lab Scene: Inner products atom pairs from two views. Values of the inner products for the corresponding atom pairs are displayed in bold.*

	atoms in view 2							
atoms in view 1	1	2	3	4	5	6	7	8
1	**0.9964**	0.8545	0.9088	0.9073	0.9480	0.9015	0.7815	0.5918
2	0.9435	**0.9443**	0.9739	0.9359	0.9763	0.8857	0.7908	0.6900
3	0.9088	0.9903	**1.0000**	0.8473	0.9097	0.8088	0.6834	0.6779
4	0.9125	0.8904	0.9245	0.9760	0.9826	0.8803	0.8564	0.6526
5	0.9015	0.7576	0.8088	0.9880	0.9717	0.9496	0.9506	0.6703
6	0.6188	0.6426	0.6531	0.5997	0.6680	0.7639	0.5537	**0.8214**
7	0.7217	0.5352	0.5830	0.7930	0.7440	0.7298	0.8226	0.4209
8	0.7376	0.5502	0.5958	0.5884	0.6466	0.7668	0.5385	0.4385

Table 4.6: *Lab scene: Average symmetric epipolar atom distance for atom pairs from two views. Values of the epipolar distance for corresponding atom pairs are displayed in bold.*

	atoms in view 2							
atoms in view 1	1	2	3	4	5	6	7	8
1	**0.0007**	0.0030	0.0048	0.0047	0.0050	0.0053	0.0012	0.0044
2	0.0056	**0.0021**	0.0068	0.0078	0.0057	0.0081	0.0054	0.0046
3	0.0048	0.0035	**0.0009**	0.0063	0.0044	0.0068	0.0054	0.0032
4	0.0048	0.0025	0.0039	0.0049	0.0004	0.0045	0.0046	0.0010
5	0.0006	0.0020	0.0025	0.0015	0.0033	0.0029	0.0002	0.0030
6	0.0040	0.0019	0.0039	0.0045	0.0026	0.0044	0.0036	**0.0008**
7	0.0026	0.0014	0.0028	0.0032	0.0018	0.0031	0.0026	0.0012
8	0.0013	0.0024	0.0039	0.0040	0.0037	0.0044	0.0018	0.0035

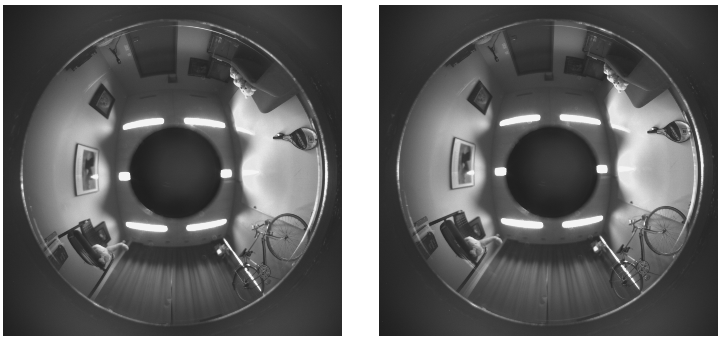

(a) **(b)**

Figure 4.9: *Two views from the "Mede" database.*

Finally, we have evaluated our correlation model on a larger set of multi-view omnidirectional images. We have used our database "Mede" that consists in two sets, each containing 27 multi-view omnidirectional images of the static indoor environment, where the camera has been moved in the $x - y$ plane (on the floor). Cameras have been moved within precise distances, obtaining the ground truth values of the translation vector. There is no change in camera rotation between different views, while the translation vector varies within the interval: $x \times y = [-40cm, 40cm] \times [-40cm, 40cm]$. Calibration with respect to intrinsic parameters has been performed using the Omnidirectional Calibration Toolbox (http://www.robots.ox.ac.uk/̃cmei/Toolbox.html)[5]. Some of the images from the "Mede" database are shown in Figure 4.9. The images have been mapped to spherical images of resolution $2B = 256$.

We have randomly chosen 10 image pairs from the database, with different translation vectors (i.e., different camera distances). For each image pair, we have computed five Gaussian atoms using MP. The epipolar distance and the inner product between centered atoms (the shape similarity measure) have been then evaluated between all possible atom correspondences in a given image pair. The ground truth that denotes the semantically correct atom correspondences has been obtained by human labeling. If two atoms correspond to the same feature in the 3D scene, their correspondence is denoted as *true*. The total of 40 true correspondences have been found in 10 image pairs. The scatter plot in Figure 4.10 presents the epipolar distance versus the inner product for all atom correspondences in all 10 image pairs. The circles denote the true correspondences, while the dots denote the atom pairs that do not represent semantic correspondences. We can see that the true correspondences are concentrated in the area of minimal epipolar distance and high inner product, which supports our correlation model. Moreover, we can see that the epipolar distance measure is more reliable than the shape similarity measure, as pointed out previously. Figure 4.11 shows the histograms of the epipolar distance and the inner product for true correspondences. Epipolar distance histogram is peaked around zero, while the inner product histogram is peaked around one. This confirms the validity of our correlation model.

4.7 Conclusion

State of the art work in multi-view image and video coding undoubtedly shows an essential need for better multi-view correlation models, beyond the classical block-based disparity models and global homography models. In this chapter, we have introduced a novel multi-view correlation model based on sparse image representations with geometric dictionaries. The new model does not

---

[5] For the Lab images the intrinsic calibration has been done by finding a circle corresponding to the mirror boundary and projecting to the sphere, as desribed in Section 2.3.4. However, this process is semi-automatic and timely for the large databases. Therefore, for the "Mede" database we have used the Omnidirectional Calibration Toolbox.

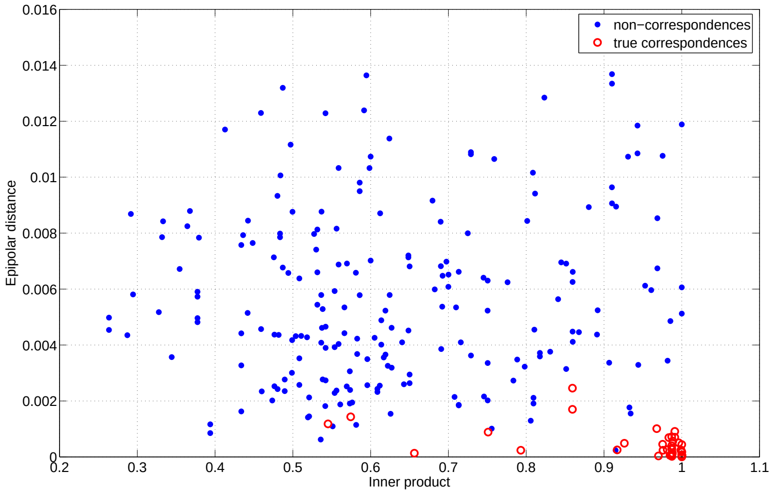
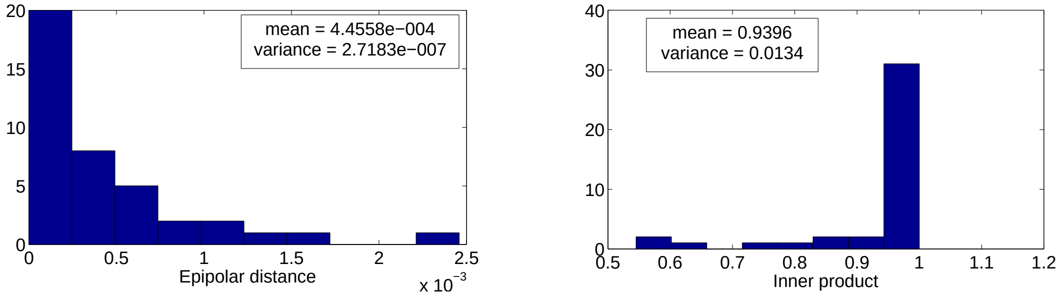


Figure 4.10: *Scatter plot of the epipolar distance and the inner product for atom correspondences, for 10 randomly selected image pairs from the "Mede" database. Circles denote the true correspondences and the dots denote the atom pairs that do not represent correspondences.*



(a) **(b)**

Figure 4.11: *"Mede" database: a) histogram of the epipolar distance for true correspondences, b) histogram of the inner product for true correspondences.*

require a specific camera arrangement nor image rectification, and is generic with respect to the projective image surface. The proposed model has been studied and validated for the special case of multi-view spherical images. In the next two chapters we will show how this new correlation model can be advantageously used in two applications: distributed multi-view image coding (in Chapter 5) and camera pose estimation from compressed images (in Chapter 6).

Chapter 5

Distributed Scene Coding

5.1 Introduction

Multi-view correlation modeling is certainly one of the most important steps in designing coding methods for multi-view image compression in camera networks. Distributed camera networks offer simple and cost effective solutions for scene acquisition, where several views of a scene can be combined to produce a complete 3D representation. Image compression in a camera network should take benefit of the correlation between multiple views to obtain rate-distortion effective scene representations, from which one can reproduce depth and visual information. Bandwidth limitations and power constraints usually do not permit communication between cameras, thus imposing a distributed coding approach for compression of visual information.

In this chapter, we consider a framework where a central decoder reconstructs the static 3D scene information based on multiples images encoded by distributed cameras (see Figure 5.1). We focus on the distributed coding of camera images, where no communication between cameras is assumed. Interestingly, results from information theory demonstrate that it is possible to exploit the correlation between sources without communication between encoders, as long as the decoding is performed jointly [Sle73,Wyn76]. Distributed coding relies on the knowledge of a good correlation model between information sources. However, correlation modeling is a quite challenging problem, especially in imaging applications. A combination of the translational motion model and the pixel-based correlation has been used in applications of distributed coding to low-complexity video

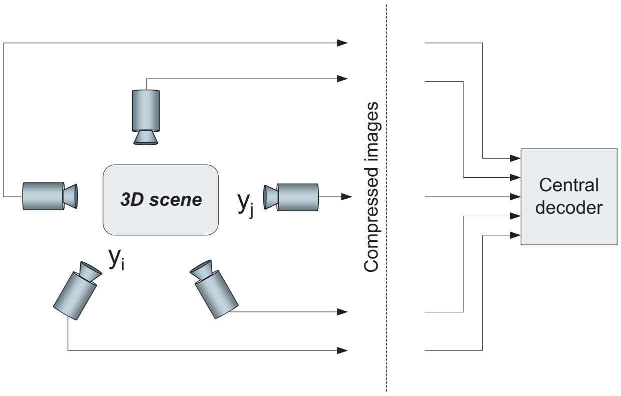


Figure 5.1: *Distributed coding of 3D scenes. Multiple correlated views $\{y_i\}$ of the scene are encoded independently, and decoded jointly by the central decoder.*

compression [Gir05, Seh04]. In case of static scene representation, images from different cameras are correlated by geometric constraints on the 3D objects in the scene. Image projections of the 3D objects from different viewpoints are correlated by local transforms such as translation, scaling or rotation. Hence, the block-translational correlation model is not sufficient to cope efficiently with all types of local transforms that happen when the viewpoint changes.

The geometric multi-view correlation model proposed in the Chapter 4 is used in this chapter to design a distributed coding method with side information for multi-view omnidirectional images. A Wyner-Ziv coder is designed by partitioning the redundant dictionary into cosets based on atom dissimilarity. The joint decoder then selects the best candidate atom within the coset with help of the side information image. The correspondences that are found during decoding between atoms in both image expansions are further used to estimate local transforms and to build a disparity map between correlated views. These transforms are used to refine the side information for decoding the rest of the atoms. Furthermore, we propose to make the Wyner-Ziv coder resilient to a certain number of occlusions or inaccuracies in the view correlation model by sending additional syndrome bits that are computed by channel coding across the atoms of the Wyner-Ziv image. Experimental results show that the proposed method successfully identifies the local geometric transforms between sparse image components in different views and implicitly provides coarse geometry information about the scene. The distributed coding scheme is shown to outperform independent coding strategies and to approach the performance of a joint coding strategy at low bit rate. Finally, we show that the occlusion-resilient coding corrects the saturation effect in the RD performance towards higher bit rates, where it performs close to the joint encoding strategy.

5.2 Distributed coding of correlated sources

5.2.1 Slepian-Wolf theorem

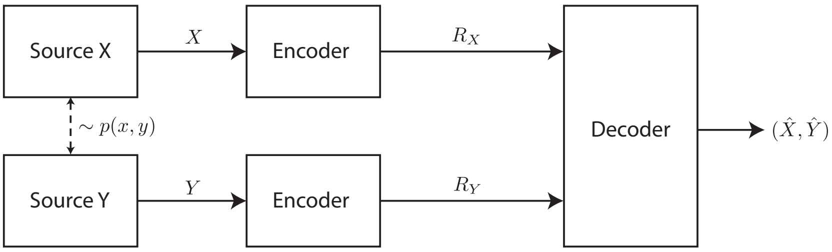


Figure 5.2: *Distributed source coding block scheme.*

Distributed source coding (DSC) refers to separate encoding and joint decoding of correlated sources. The diagram of a distributed coding scheme is shown in Figure 5.2 for the case of two correlated sources (X, Y) with a joint probability distribution $p(x, y)$. Sources X and Y are encoded at rates R_X and R_Y, respectively. Let $H(\cdot)$ denote the entropy of a given source. From information theory we know that the rates $R_X \geqslant H(X)$ and $R_Y \geqslant H(Y)$ are sufficient for lossless coding of sources X and Y, when they are encoded separately. On the other hand, when correlated sources (X, Y) are jointly encoded, the total rate needed for lossless transmission is $R = R_X + R_Y \geqslant H(X, Y)$. Surprisingly, even if the sources are encoded separately but decoded jointly, a total rate $R \geqslant H(X, Y)$ is sufficient, as it has been proven by Slepian and Wolf [Sle73]. According to the Slepian-Wolf theorem, for the distributed source coding problem of correlated sources (X, Y) drawn *i.i.d* $\sim p(x, y)$, the achievable rate region is given by:

$$\begin{aligned} R_X &\geqslant H(X|Y) && (5.1)\\ R_Y &\geqslant H(Y|X) && (5.2)\\ R_X + R_Y &\geqslant H(X,Y). && (5.3) \end{aligned}$$

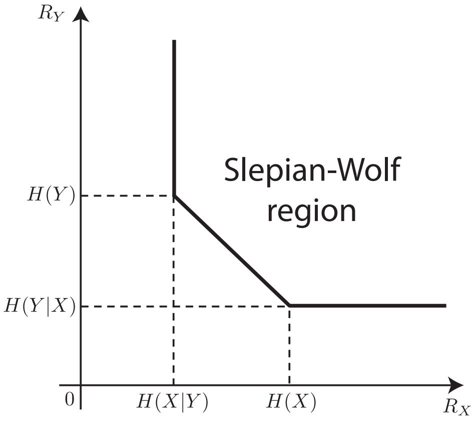


Figure 5.3: *Achievable rate region for the Slepian-Wolf problem.*

For the proof of the Slepian-Wolf theorem, please see [Sle73].

The achievable rate region, called the Slepian-Wolf region, is shown in Figure 5.3. The corner points in the Slepian-Wolf region present a special case of the distributed source coding, often referred to as coding with side information. Let us observe the corner point with rates $R_X = H(X)$ and $R_Y = H(Y|X)$. The Slepian-Wolf theorem for this special case can be interpreted with graph coloring [Cov91]. Let $\mathcal{X}$ and $\mathcal{Y}$ denote respectively the alphabets of sources X and Y, depicted by the corresponding sets of points (codewords) in Figure 5.4. Source X represents the side information and it is encoded using $R_X \geqslant H(X)$ bits with arbitrarily small error probability of error. We further want to see how many bits are needed for encoding Y. Due to the correlation of the sources, each codeword in $\mathcal{X}$ is jointly typical with a subset of codewords in $\mathcal{Y}$ that form a "typical fan". The most important condition of the Slepian-Wolf coding is that both encoder and decoder are aware of the correlation model, i.e., know the jointly typical fans of X and Y. Therefore, the encoder of Y can randomly color the points in the typical fan using 2^{R_Y} colors, and send to the decoder only the codeword index in this fan, i.e., only the color of the point. Knowing the X, the color, and the typical fans, the decoder can then decode Y. According to the Slepian-Wolf theorem, if $R_Y \geqslant H(Y|X)$ the probability of decoding error for Y is exponentially small.

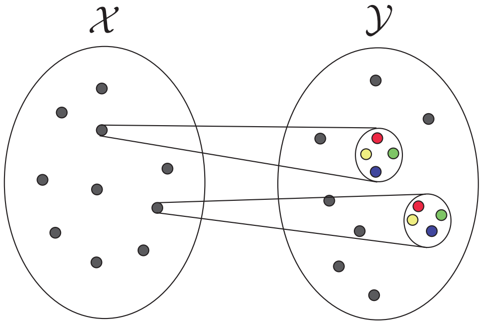


Figure 5.4: *Graph-coloring interpretation of a corner point of the Slepian-Wolf region [Cov91].*

A simple, but illustrative example of coding with side information was given by Pradhan and Ramchandran [Pra03]. Let X and Y denote two equiprobable 3-bit binary words, correlated is such a way that their Hamming distance is not greater than one. The encoder uses three bits to encode the side information X. Following the intuition behind the graph-coloring interpretation of the Slepian-Wolf theorem, a separate encoder for Y can construct four subsets of all possible

3-bit binary words such that in each subset the Hamming distance between words is three. Those four subsets are: $\{(000, 111), (001, 110), (010, 101), (011, 100)\}$, where each subset is equivalent to one of the colors in the graph-coloring interpretation. Clearly, to encode the index of the subset in which Y belongs, the encoder of Y needs two bits. Knowing that the Hamming distance between X and Y is not more than one, the decoder can then use the side information X to disambiguate between the two words in the encoded subset of Y and decode the Y. For example, if $X = 010$, and $Y = 011$, the encoder sends a full description for X using 3 bits and the information that Y belongs to the fourth subset, using 2 bits. The decoder would then have to decide which of the two words in the subset $(011, 100)$ is actually Y. Since the Hamming distance between 010 and 100 is two, the only possible decoding of Y is 011, which is the correct decoded word. As noted by Pradhan and Ramchandran, the four constructed subsets are actually cosets of a 3-bit repetition code and they cover the whole space of binary 3-tuples, i.e., the whole alphabet of Y [Pra03]. Since every coset of a linear code is associated with a unique syndrome $\mathbf{s}$, defined as $\mathbf{s}^\mathsf{T} = \mathbf{H}\mathbf{c}^\mathsf{T}$ where $\mathbf{H}$ is the parity check matrix of the linear code and $\mathbf{c}$ is a valid codeword, encoding Y is essentially equivalent to sending a syndrome of a linear channel code. Therefore, the rates in the Slepian-Wolf region can be achieved using the capacity approaching channel codes, as it was first realized by Wyner [Wyn74].

5.2.2 Wyner-Ziv theorem

Few years after the publication of the Slepian-Wolf theorem, Wyner and Ziv extended the case of coding with side information to the lossy coding scenario [Wyn76]. They have established the rate-distortion bound for coding the source X under the constraints that the average distortion $E\{d(X, \hat{X})\}$ is bounded by D_X and that the side information Y is available at the decoder. The block scheme of the Wyner-Ziv encoder is shown in Figure 5.5. Wyner and Ziv proved that for general sources, the rate-distortion bound of the Wyner-Ziv coder $R^{WZ}_{X|Y}(D_X)$, is given by:

$$R^{WZ}_{X|Y}(D_X) \geqslant R_{X|Y}(D_X), \tag{5.4}$$

where $R_{X|Y}(D_X)$ is the rate distortion bound for the joint encoding case. They also showed that the equality sign in Eq. (5.4) holds when X is Gaussian and $Y = X + U$, where U is also Gaussian and independent of X, and the distortion measure is quadratic $d(x, \hat{x}) = (x - \hat{x})^2$.

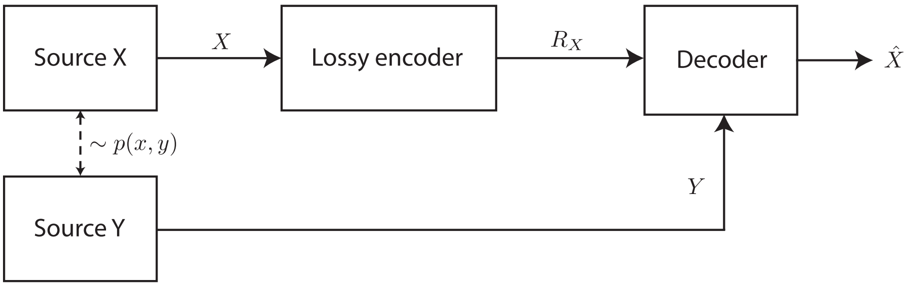


Figure 5.5: *Wyner-Ziv coding scheme.*

5.3 Wyner-Ziv coding with the geometric correlation model

We propose a scheme for coding with side information where image y_1 is independently encoded using Omni-SMP at a rate $R_{y_1} \geqslant H(y_1)$, and image y_2 is encoded by coset coding at the rate $R_{y_2} \geqslant H(y_2|y_1)$. This corresponds to an asymmetric scheme, where the rate is not balanced between the encoders. The coset design is directly based on the correlation model introduced in the Chapter 4, which explicitly relates atom parameters with scene geometry constraints.

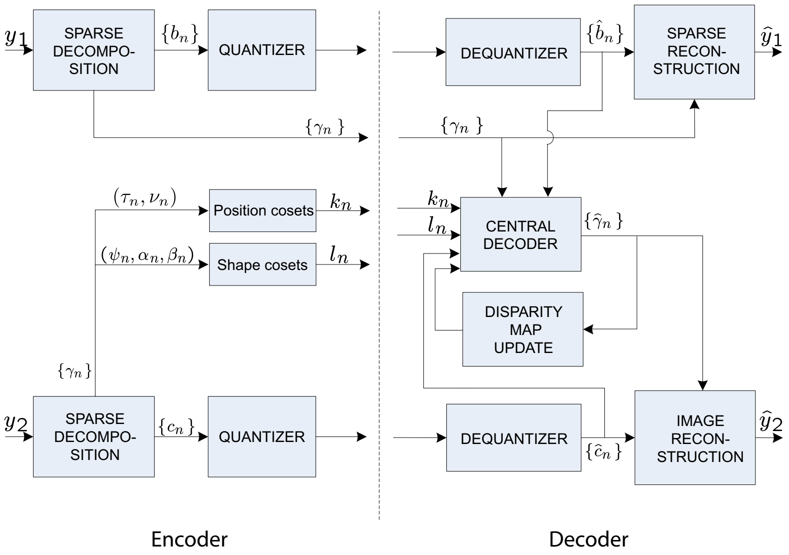


Figure 5.6: *Block diagram for the Wyner-Ziv codec.*

5.3.1 Coding with side information

Images are first independently decomposed into linear expansions of geometric atoms using the Matching Pursuit algorithm on the sphere (SMP). Nevertheless, one can use any other existing sparse approximation algorithm instead of MP. Interestingly, in certain cases when correlated signals satisfy some sparsity conditions with a chosen dictionary, simple coefficient thresholding can give their correct sparse approximation, as we show in Appendix A.1. The sparse decomposition of the reference image y_1 is independently encoded, while the decomposition of the Wyner-Ziv image y_2 is encoded by coset coding of atom indexes and quantization of their respective coefficients, as shown in Figure 5.6.

We propose to partition the set of atom indexes Γ into distinct cosets that contain dissimilar atoms with respect to their position and shape. Under the assumption that an atom h_{γ_j} in the Wyner-Ziv image decomposition given by Eq. (4.2) has its corresponding atom g_{γ_i} in the side information expansion, the Wyner-Ziv encoder does not need to code the entire γ_j. It rather transmits only the information that is necessary to identify the correct atom in the transform candidate set given by $\Gamma_i = \Gamma_i^{\mu} \cap \Gamma_i^{E}$, where Γ_i^{μ} is the shape candidates set given in Eq. (4.7) and Γ_i^{E} is the epipolar candidates set given in Eq. (4.10). The side information and the coset index are therefore sufficient to recover the atom g_{γ_j} in the Wyner-Ziv image. The achievable bit rate for encoding the atom index γ_j is thus reduced from $R_{y_2} \geqslant H(\gamma_j | \gamma_j \in \Gamma)$ to $R_{y_2} \geqslant H(\gamma_j | \gamma_j \in \Gamma_i)$.

5.3.2 Coset design

Designing cosets that allow decoding without ambiguities requires distributing in different cosets atoms whose indexes belong to the same Γ_i, for all i. This implies that we need to partition the five-dimensional space of atom parameters, which is of prohibitive computational complexity. Therefore, we simplify the coset design by considering a special case of the general epipolar constraint given in Eq. (4.9). We consider the case where only the centers of two corresponding atoms g_{γ_i} and h_{γ_j} satisfy the epipolar constraint, and not the whole atoms as in Eq. (4.9). Atom centers are defined by the coordinates of their positions (τ_i, ν_i) and (τ_j, ν_j), denoted as m_i and m_j, respectively. Therefore, the epipolar candidates set given in Eq. (4.10) reduces to:

$$\Gamma_i^{M} = \{\gamma_j | h_{\gamma_j} = U(\gamma') g_{\gamma_i}, d_{SE}(m_i, m_j) \leqslant \delta\}, \tag{5.5}$$

where δ represents a small threshold value on the symmetric epipolar distance. Formally, Eq. (5.5) is a special case of Eq. (4.9) when $\mathcal{G}_1 = m_i$, which transforms into $\mathcal{G}_2 = m_j$. For this special case the epipolar geometry constraint depends only on atom positions (τ, ν), and the epipolar and shape constraints become independent. Therefore, the cosets can be designed independently for atom shape parameters (ψ, α, β), and for atom positions (τ, ν) according to epipolar constraints. We therefore construct two types of cosets, respectively the Shape cosets: $K_l^{\mu}, l = 1, ..., N_2$ and the Position cosets: $K_k^E, k = 1, ..., N_1$. The encoder eventually sends for each atom only the indexes of the corresponding cosets (i.e., k_n and l_n in Figure 5.6).

We design Shape cosets by distributing all atoms whose parameters belong to Γ_i^{μ}, for all i, into different cosets. Samples of atoms that belong to the same Shape coset are illustrated in Figure 5.7, showing the significant difference in their shapes.

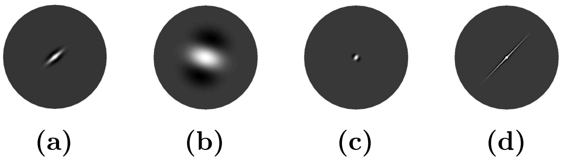


Figure 5.7: *Samples of atoms in the same Shape coset.*

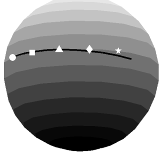


Figure 5.8: *Illustration of the epipolar coset design. The atoms with positions marked with different shapes are put in different epipolar cosets since they belong to the same epipolar line.*

Next, we propose two design methods for constructing the Position cosets, which respectively correspond to the scenarios where the camera pose $(\mathbf{R}, \mathbf{T})$ is known, or not available for coset design. We first design *Epipolar (EPI) cosets* by separating into different cosets the atoms that belong to the same set Γ_i^M for $\mathcal{G}_1 = m_i$. This is illustrated in Figure 5.8 for a random epipolar line. The parameter δ can be used in the coset design for selecting the number of cosets and adapting the encoding rate. Given the side information atom g_{γ_i}, the decoder only needs to know the coset index of h_{γ_j} for joint decoding.

As an alternative, we propose to design Position cosets based on *Vector Quantization* (VQ) [Ger92] of positions in the absence of information about the relative camera poses. The VQ cosets are constructed by 2D interleaved uniform quantization of atom positions (τ, ν) on a rectangular lattice in the $[0, \pi] \times [0, 2\pi)$ area. This coset design can be formulated similarly to the Epipolar coset design, where the set of position candidates (called the set of epipolar candidates in Eq. (5.5)) gathers the candidates positions (τ_j, ν_j) within the neighborhood of the reference atom position (τ_i, ν_i), i.e.,:

$$\Gamma_i^V = \{\gamma_j | h_{\gamma_j} = U(\gamma') g_{\gamma_i}, |\tau_i - \tau_j| < \Delta\tau, |\nu_i - \nu_j| < \Delta\nu\}. \tag{5.6}$$

The vector quantization of τ and ν distributes the pairs (τ, ν) that belong to the same Γ_i^V into different cosets. Additionally, it keeps constant the distance between coset elements, which is equal to $(\Delta\tau, \Delta\nu)$. Note that the constant intra-coset distance can not be guaranteed in the case of EPI cosets. Both coset design methods are used for the experiments, and their selection depends on the constraints of the camera network application.

5.3.3 Central decoder

The central decoder (CD) exploits the correlation model based on local atom transforms, in order to establish correspondences between atoms in the reference image and atoms within the cosets of the Wyner-Ziv image decomposition (see Figure 5.6). It also uses the information provided by the quantized coefficients, in order to improve the atom matching process. For decoding of the n^{th} atom in the Wyner-Ziv frame, the decoder has the following information: the index of the Position coset k_n, the index of the Shape coset l_n, and the coefficient $\hat{c}_n$ after inverse quantization. The decoder then selects the atom position (τ_n, ν_n) from the coset $K^E_{k_n}$ and the atom shape $(\psi_n, \alpha_n, \beta_n)$ from the coset $K^\mu_{l_n}$.

Let A_n denote the set of possible candidates for decoding the n^{th} atom in y_2, with $|A_n| = |K^E_{k_n}| \cdot |K^\mu_{l_n}|$ when $|\cdot|$ denotes the cardinality of a set. By the construction of the cosets, only a small subset of atoms in A_n have corresponding atoms in I_1, where I_1 denotes the atoms participating in the sparse expansion of the reference image y_1 (see Section 4.3). The decoder has, therefore, to identify the possible pairs of corresponding atoms between A_n and I_1.

The quantized values of the atom coefficients for the Wyner-Ziv image, denoted as $\hat{c}_n$, are available at the decoder. To reduce the set of possible corresponding atoms in I_1, the decoder selects a subset of atoms in I_1 whose inner products $\langle \hat{y}_1, g_{\gamma_i} \rangle$ are related to $\hat{c}_n$. The relation between these inner products and coefficients $\hat{c}_n$ can be established when the coefficients are obtained as projections of the image to the corresponding atom, i.e., when $c_n = \langle y_2, h_{\gamma_n} \rangle$. Under the assumption that the image approximations are sparse enough, the projections of two corresponding atoms g_{γ_i} and h_{γ_n} are related as:

$$\frac{\langle \hat{y}_1, g_{\gamma_i} \rangle}{o_i} \approx \frac{\langle y_2, h_{\gamma_n} \rangle}{o_n}, \tag{5.7}$$

where o_i and o_n denote the norms of atoms g_{γ_i} and h_{γ_n} prior to atom normalization. Therefore, the decoder can select a subset of atoms $L_n = \{\gamma_i\}$ in I_1 that satisfy:

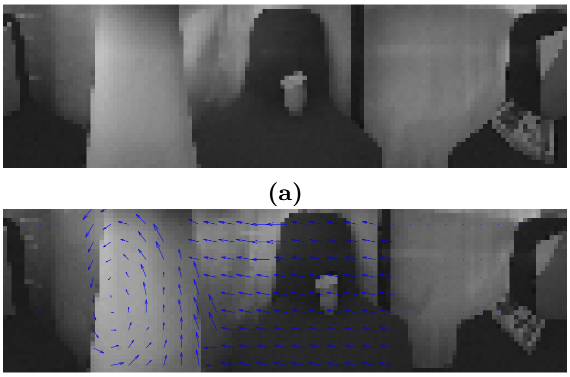
$$\Delta c = |\frac{\frac{o_n}{o_i}\langle \hat{y}_1, g_{\gamma_i} \rangle - \hat{c}_n}{\hat{c}_n}| \approx |\frac{\frac{o_n}{o_i}\langle \hat{y}_1, g_{\gamma_i} \rangle - \langle y_2, h_{\gamma_n} \rangle}{\langle y_2, h_{\gamma_n} \rangle}| < \sigma, \tag{5.8}$$

where σ is a small chosen threshold. For each $g_{\gamma_i} \in L_n$ we have a set of possible transformed atoms given by $\tilde{\Gamma}_i = \Gamma^M_i \bigcap \Gamma^\mu_i$ or $\tilde{\Gamma}_i = \Gamma^V_i \bigcap \Gamma^\mu_i$ respectively for epipolar or VQ cosets. The decoder further looks if any of the candidates in A_n belongs to $\bigcup_{\gamma_i \in L_n} \tilde{\Gamma}_i$, i.e., it looks for pair correspondences between atoms in L_n and atoms in A_n. Note that, in the general case, the parameters $\delta, \Delta\tau, \Delta\nu, s$ that define the correlation sets Γ^M_i, Γ^V_i and Γ^μ_i can have different values for the coset design and decoding. This permits us to put stricter conditions for selecting the corresponding atom pairs.

The search for atom correspondences then proceeds in two major steps. First, the decoder eliminates the candidates that do not belong to $\bigcup_{\gamma_i \in L_n} \tilde{\Gamma}_i$, as well as candidates with a large symmetric epipolar atom distance, i.e., for which $d_{EA}(g_{\gamma_i}, h_{\gamma_j}) > \kappa$. Second, the decoder selects as a correspondence the pair of atoms with the smallest symmetric epipolar atom distance $d_{EA}(g_{\gamma_i}, h_{\gamma_j})$ among the candidates that have not been eliminated in the first step. If all candidates in A_n get eliminated, the decoder decides that the n^{th} atom in y_2 does not have a corresponding atom in I_1.

Once a correspondence is identified, the decoder updates the global disparity map that relates each pixel in the Wyner-Ziv image with a pixel in the reference image, as described in Section 4.5. The global disparity map is updated by combining the local disparity map induced by the last pair of atoms with the disparity maps from correspondences that have been previously defined. The global disparity map represents the fusion of local disparity maps from multiple correspondences. Namely, for each point $\mathbf{z}_2$ in y_2, the most confident mapping point in y_1 is selected from points $\mathbf{z}_1^{(i)}, i = 1, ..., n$ that represent different mappings defined by n correspondences. The final mapping point $\mathbf{z}_1^*$ is selected as:

$$\mathbf{z}_1^* = \arg \max_{\mathbf{z}_1^{(i)}, i=\overline{1,n}} w_{\gamma_i}(\mathbf{z}_1^{(i)}), \tag{5.9}$$



(a)

(b)

Figure 5.9: *(a) Reference image; (b) Wyner Ziv image with the disparity map displayed by the disparity vectors. For better visibility, we show the disparity map that is downsampled by four. The direction and length of the vectors show how pixels from the Wyner-Ziv image map to the pixels on the reference image.*

where we have used the same weight function as for the symmetric epipolar atom distance in Eq. (4.15). Selecting the most confident mapping is advantageous over taking the weighted average of all possible mappings because it does not introduce the blurring due to averaging. Moreover, the maximum confidence based approach is consistent with existing disparity estimation algorithms, where the final mapping is chosen as the maximum a posteriori estimate [Kol02, Sun03]. Figure 5.9(b) shows an example of the disparity map obtained from few atom correspondences. It relates pixels in the Wyner-Ziv image in Figure 5.9 (b) to pixels in the reference image in Figure 5.9(a). The relation between these pixels is displayed by vectors. The disparity of the chair in the middle of the scene is clearly captured by the mapping. The white pillar is slightly rotated with respect to the reference image.

The mapping of the reference image with respect to the global disparity map provides an approximation of the Wyner-Ziv image. Namely, each pixel in the Wyner Ziv image is approximated by its corresponding pixel in the reference image, based on the disparity map. The geometry of the scene is usually captured by the most prominent atoms that are found as correspondences between the views. Therefore, the global disparity map defines the most important local transforms between views and hence leads to a good approximation of the Wyner-Ziv image, denoted as y_{tr}. However, according to the correlation model in Eq. (4.3), there are atoms in the Wyner-Ziv image that do not have corresponding atoms in the reference image. Since y_{tr} represents an approximation of the Wyner-Ziv image, it is used as a side information for decoding the atoms without correspondences. Those atoms are decoded based on the mean square error between the side information (y_{tr}) and the Wyner-Ziv image reconstructed from previously decoded atoms and the current decoding candidate from A_n. The candidate from A_n with the minimal mean square error is selected as the decoded atom.

All decoded atoms from the Wyner-Ziv image (i.e., from $\mathbf{\Phi}_{I_2}$) and their quantized coefficients are used to evaluate the decoded Wyner-Ziv image y_d. Finally, the reconstruction of the Wyner-Ziv image $\hat{y}_2$ is obtained as a linear combination of the decoded image y_d and the side information y_{tr}, i.e.,:

$$\hat{y}_2 = y_d + \lambda \mathbf{\Psi}_d \, y_{tr}. \tag{5.10}$$

The matrix $\mathbf{\Psi}_d$ denotes the orthogonal complement to the basis formed by the decoded atoms in $\mathbf{\Phi}_{I_2}$, and λ is a trade-off parameter. Namely, the reconstructed Wyner-Ziv image benefits from both the decoded information and the transformed features from the reference image. We estimate the value of λ from the energy conservation principle. Under the assumption that $\|\mathbf{\Psi}_d \, y_{tr}\| \approx \|\mathbf{\Psi}_d \, y_2\|$, we get λ from Eq. (5.10) as:

$$\lambda \approx \sqrt{1 - \frac{\|y_d\|^2}{\|y_2\|^2}}. \tag{5.11}$$

.

The pseudo-code for the decoder is given in the Algorithm 2.

Algorithm 2 Decoder

```
initialization: $DM = \mathcal{G}_1$ (disparity map DM is initialized to the uniform grid $G_1$),
$\mathbf{\Phi} = [\ ], C = [\ ]$;
input: $\hat{b}_i, g_{\gamma_i}, i = 1, ..., N_r$, evaluate $\hat{y}_1 = \sum_{i=1}^{N_r} \hat{b}_i g_{\gamma_i}$;
for n=1,...,N do
  input: $k_n, l_n, \hat{c}_n$
  initialization: $W_c = \varnothing$ (set of paired candidates)
  $A_n = \{\gamma | (\tau, \nu) \in K^E_{k_n}, (\psi, \alpha, \beta) \in K^\mu_{l_n}\}$
  for all $\gamma_p \in A_n$ do
    for all $\hat{b}_i, g_{\gamma_i}, i = 1, ..., N_r$ do
      if $(\Delta c < \sigma)$ & $(\gamma_p \in \tilde{\Gamma}_i)$ & $(d_{EA}(g_{\gamma_i}, h_{\gamma_p}) < \kappa)$ then
        add $\gamma_p$ to $W_c$
      end if
    end for
  end for
  if $W_c$ not empty then
    $\hat{\gamma}_n = \arg\min_{\gamma_p} |d_{EA}(g_{\gamma_i}, h_{\gamma_p})|$;
    announce $\hat{\gamma}_n$ decoded, update $DM$
    $\mathbf{\Phi} = [\mathbf{\Phi}; h_{\hat{\gamma}_n}], C = [C; \hat{c}_n]$
  else
    $\gamma_n$ not decoded.
  end if
end for
$y_{tr} = DM(\hat{y}_1)$
for n=1,...,N do
  if $\gamma_n$ not decoded then
    for all $\gamma_p \in A_n$ do
      $\mathbf{\Phi}_p = [\mathbf{\Phi}; h_{\hat{\gamma}_p}], C_p = [C; \hat{c}_n]$, $e_p = \|y_{tr} - \mathbf{\Phi}_p^\dagger C_p\|^2$
    end for
    $\hat{\gamma}_n = \arg\min_{\gamma_p} |e_p|$, $\mathbf{\Phi} = [\mathbf{\Phi}; h_{\hat{\gamma}_n}], C = [C; \hat{c}_n]$
  end if
end for
$y_d = \mathbf{\Phi}^\dagger C$, $\lambda \approx \sqrt{1 - \|y_d\|^2 / \|y_2\|^2}$, $\hat{y}_2 = y_d + \lambda \mathbf{\Psi}_d\ y_{tr}$
```

5.3.4 Occlusion-resilient coding

The decoding procedure described in Section 5.3.3 does not, however, give any guarantee that all atoms are correctly decoded. In a general multi-view system, occlusions are quite probable, and clearly impair the atom pairing process at the decoder. The atoms that approximate these occlusions cannot be decoded based on the geometric correlation model since they do not have any corresponding feature in the reference image. Moreover, there are also features that are not sufficiently prominent to appear in sparse approximations of both views. Hence, the encoder needs to send additional information about these atoms for correct decoding. For example, it can send the values of p_n and q_n that identify the atom g_{γ_n} within the Position coset $K^E_{k_n}$ and the Shape

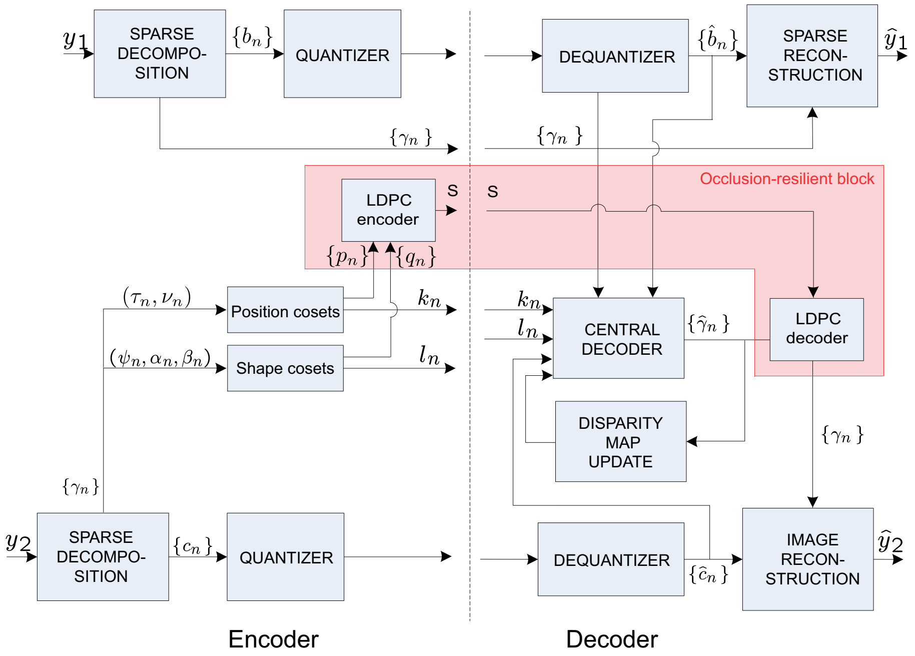


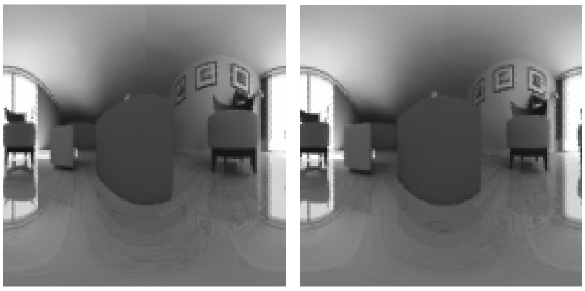
Figure 5.10: *Occlusion-resilient Wyner-Ziv coder.*

coset $K^{\mu}_{l_n}$, respectively. Along with the coset indexes k_n and l_n, indexes p_n and q_n uniquely define γ_n in the set Γ. However, the main problem in the distributed setting is that the encoder does not know which of the atoms in the decomposition are occlusions and non-prominent features since there is no communication between separate encoders. Therefore, the encoder cannot send p_n and q_n only for the problematic atoms. Still, the encoder can make the assumption that at least M out of total N atoms in the sparse decomposition represent occlusions or non-prominent features that cause decoding failures, without necessarily knowing their position in the sparse decomposition. The decoder correctly decodes $N - M$ atoms, which leads to an error probability of M/N. Following this observation, we propose to modify the described Wyner-Ziv encoder by performing channel coding on all p_n and q_n, $n = 1, ..., N$, together. Thus, the encoder sends a unique syndrome S for all atoms, which is then used by the decoder to correct the erroneously decoded atom parameters. The modification of the Wyner-Ziv coder with the occlusion-resilient block (shaded part) is shown in Figure 5.10.

5.4 Experimental Coding Results

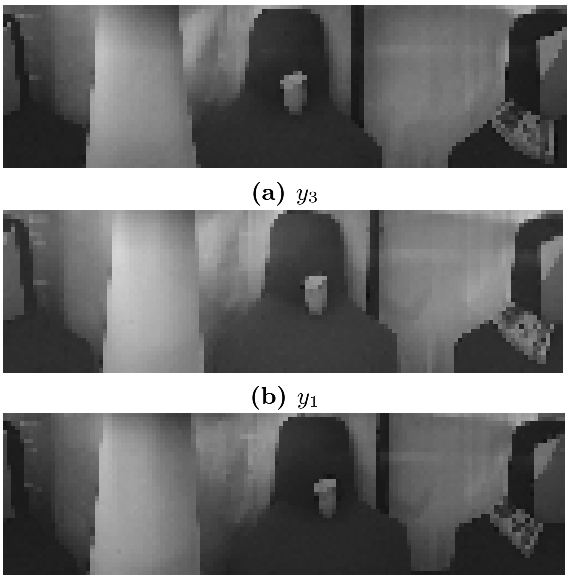
5.4.1 Experimental settings

We analyze here the performance of the above Wyner-Ziv coding method for two sets of multi-view images: synthetic spherical images of the Room scene (Figure 5.11) and natural omnidirectional images of the Lab scene (Figure 5.12). Room scene images include two 128×128 spherical images y_1 and y_2 captured from different viewpoints, where the relative camera pose is given by the rotation matrix equal to the identity matrix, i.e., $\mathbf{R} = \mathbf{I}$, and the translation vector $\mathbf{T} = [0 \;\; 0.3 \;\; 0]^{\mathsf{T}}$. The Lab scene images include three natural omnidirectional images y_3, y_1 and y_2, taken from omnidirectional cameras placed in a straight line in order 3-1-2 (the camera number corresponds to the index of the image). We have used the catadioptric sensor from the Remote Reality Corporation, as described in Section 3.4.3. The images are mapped to spherical images of resolution $2B = 128$. For the Lab scene we have used the camera pose $\mathbf{R} \approx \mathbf{I}$ and $\mathbf{T} \approx [1 \;\; 0 \;\; 0]^{\mathsf{T}}$, as in Section 4.6.



(a) y_1 **(b)** y_2

Figure 5.11: *Original Room images of resolution* $2B = 128$.



(a) y_3

(b) y_1

(b) y_2

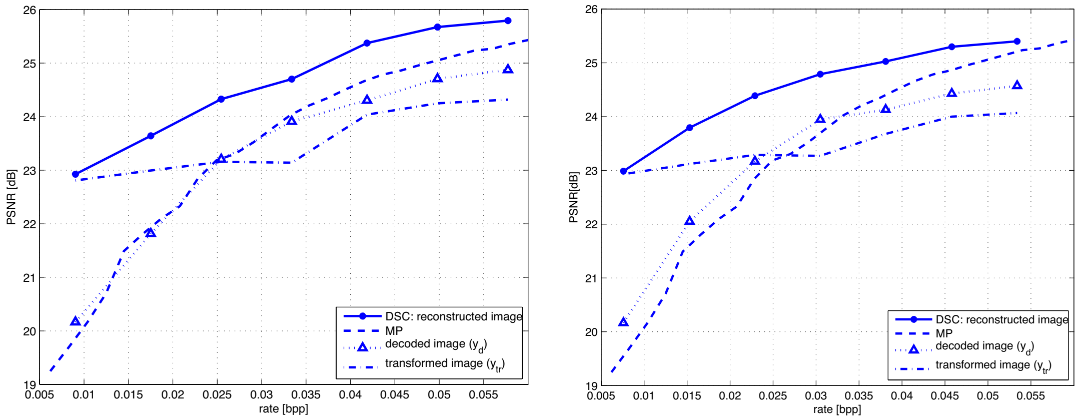
Figure 5.12: *Original Lab images. The natural omnidirectional images partially cover the sphere due to the boundaries of the mirror in the omnidirectional camera. The images are cropped to focus on the captured part of the scene.*

Sparse expansions of individual images have been constructed using the Matching Pursuit (MP) algorithm implemented on the sphere. We have used the same dictionary as for compression of omnidirectional images, which is built on low frequency and high frequency generating functions given by Eq. (3.12) and Eq. (3.11), respectively. The discretization of the dictionary is the same as in Section 3.4.2.

5.4.2 Rate-distortion analysis of the Wyner-Ziv image

We first evaluate the performance of the proposed Wyner-Ziv coding method for the Room image set by taking image y_1 as a reference image and image y_2 as Wyner-Ziv image. Image y_1 is encoded independently, with 100 MP atoms, where the coefficients are quantized by taking benefit of the energy decay properties of Matching Pursuit expansions [Fro04]. The decoded reference image is shown in Figure 5.14(a). The atom parameters for the expansion of image y_2 are coded with the proposed Wyner-Ziv scheme. The EPI cosets for position coding use a correlation parameter $\delta = \pi/5$ which gives 1024 Position cosets. Alternatively, Position cosets have also been implemented using VQ in order to generate the same number of cosets. Note that in the scheme based on EPI cosets, when the center of an atom is close to the epipoles (i.e., degenerative case of epipolar constraints) its parameters have to be encoded independently. It leads to an overhead in the coding rate for the case of EPI cosets compared to VQ cosets. The VQ coset coding with 1024 cosets enables decoding of correspondences whose translation along the zenith or the azimuth angle is strictly smaller than $\log_2(1024)/2$ pixels (see the example in Section 5.2). Since the maximum motion of objects in a 3D scene can be related to the maximum scene depth and camera position through simple geometric relations, the VQ coset size can be appropriately chosen for any given 3D scene. For the shape cosets, the correlation parameter has been set to $s_G = 0.85$ (for Gaussian atoms) and $s_A = 0.51$ (for anisotropic atoms), such that the atoms in the same coset are sufficiently different. These values lead to 128 shape cosets. The maximal change of scales and rotations for objects in a 3D scene has to result in atom correspondences with a higher inner product than s_G and s_A. This can be used as a criterion for selecting the shapes coset size for other 3D scenes. The coefficients for the Wyner-Ziv image are quantized uniformly, since they represent the inner products of the image y_2, as explained in Section 5.3.3.

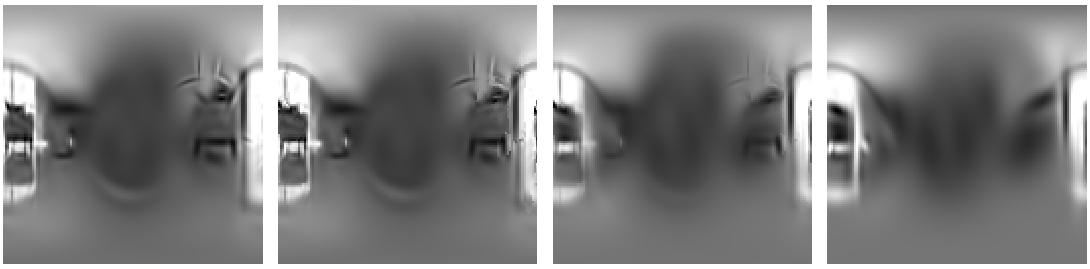
The rate-distortion (RD) performance of the proposed scheme for the Wyner-Ziv image is shown in Figure 5.13(a) and Figure 5.13(b) for EPI and VQ cosets respectively. The bit rate is changed by varying the number of received atoms, while the quantization of coefficients is kept constant. The dashed line represents the RD curve of independent coding with Matching Pursuit, while the solid line represents the proposed distributed coding scheme, given by the RD curve of the reconstructed image $\hat{y}_2$. The proposed scheme clearly outperforms the independent decoding strategy, especially at low rates. The dash-dotted line represents the RD curve of the side information image obtained by the application of the disparity map on the reference image. It shows that the disparity mapping significantly improves the side information. Moreover, it can be noted that the combination of y_d (dotted line with triangles) and y_{tr} results in a better overall PSNR of $\hat{y}_2$. The performance of the JPEG2000 has also been evaluated for the independent compression of the Wyner-Ziv image. Since JPEG2000 cannot perform at such low rates, we have evaluated the rate at which JPEG2000 reaches the PSNR performance of the proposed DSC coder. JPEG2000 gives the PSNR of 25.5 dB at 0.12 bpp, which is approximately 2.4 times greater than the rate the DSC coder needs to obtain the same image quality.



(a) EPI position cosets **(b) VQ position cosets**

Figure 5.13: *Rate-distortion performance for the Room image set.*

Images y_{tr} and $\hat{y}_2$ are presented in Figures 5.14(b) and (c). They correspond to the case of coding with VQ cosets at the rate of 0.053 bpp. We can clearly see how the disparity mapping deforms the reference image in order to compensate for different object transforms. Figure 5.14(d) shows the Wyner-Ziv image encoded independently with MP at the same rate as $\hat{y}_2$, resulting in a lower quality than the DSC coded image $\hat{y}_2$. However, the quality of image $\hat{y}_2$ is still lower than the quality of the encoded reference image $\hat{y}_1$, showing that the method results in unbalanced image qualities. On the other hand, independent coding of two images at the same overall rate could give balanced qualities, but at the price of smaller PSNR for the image considered as reference image in the DSC scheme. In order to achieve more balanced PSNR values of multi-view images in the distributed coding settings, the proposed scheme could be complemented with a more efficient method for coding the texture difference between multiple views, or by substituting the coding with side-information with a balanced distributed coding scheme. Such approaches can be envisaged in future work as extensions or applications of the correlation model proposed in this chapter. In particular, the coding of texture information is necessary for obtaining higher PSNR values than the ones obtained solely with the proposed geometric model. Therefore, our coder should be used as a first step in a hybrid (transform + texture) coding, analogously as motion estimation is used in hybrid video coders. Since the estimation of disparity and the local object transforms is the most challenging part in multi-view distributed coding, the proposed geometric model is of essential importance in this framework.



(a) $\hat{y}_1$ **(b)** y_{tr} **(c)** $\hat{y}_2$ **(d)** $\hat{y}_2^{MP}$

Figure 5.14: *DSC results for the Room images: (a) decoded reference image $\hat{y}_1$ (PSNR=30.95dB); (b) transformed reference image y_{tr}; (c) decoded Wyner-Ziv image $\hat{y}_2$ at 0.0534bpp; (d) decoded second image $\hat{y}_2^{MP}$ when encoded with MP at 0.0534bpp.*

Similar results have been obtained for the Lab image set, where the image y_1 is the reference image and image y_2 is the Wyner-Ziv image. The rate-distortion performance for the Wyner-Ziv image is shown in Figure 5.15, for the method based on the VQ coset design. Again, DSC clearly outperforms independent coding. For the Lab image set, JPEG2000 gives PSNR of 25.7 dB at 0.095 bpp, where the compression was done on cropped omnidirectional images. Therefore, the proposed DSC coder achieves this PSNR at only 1/4 of the rate that JPEG2000 needs. Images $\hat{y}_1$, y_{tr}, $\hat{y}_2$ and $\hat{y}_2^{MP}$ for the Lab Scene are presented in Figures 5.16(a), (b), (c) and (d), respectively, leading to the same conclusions as for the Room image set.

5.4.3 Overall rate-distortion performance

Figure 5.17 compares the proposed DSC method with a joint encoding algorithm, where the joint encoder finds the atom correspondences and encodes only the parameter differences for the image y_2, while the atoms without correspondences are encoded independently. Image y_1 is encoded independently at the same rate as in the DSC scheme, where the coefficients are quantized in the same manner. This joint encoding strategy is analogous to our DSC scheme, with the difference that the encoder has access to the side information. For the sake of fairness, the reconstructed image with joint encoding $\hat{y}_2^J$ is also obtained as a combination of the transformed image y_{tr} and the decoded image y_d^J, giving a better overall performance.

The DSC scheme performs very close to the joint encoding at lower rates, where the number of correspondences between views is high. However, when the number of correspondences drops, the RD performance of DSC saturates. Therefore, the proposed method should be seen as scene

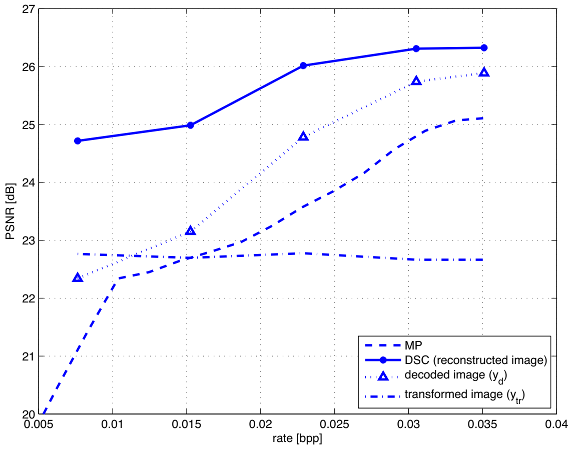
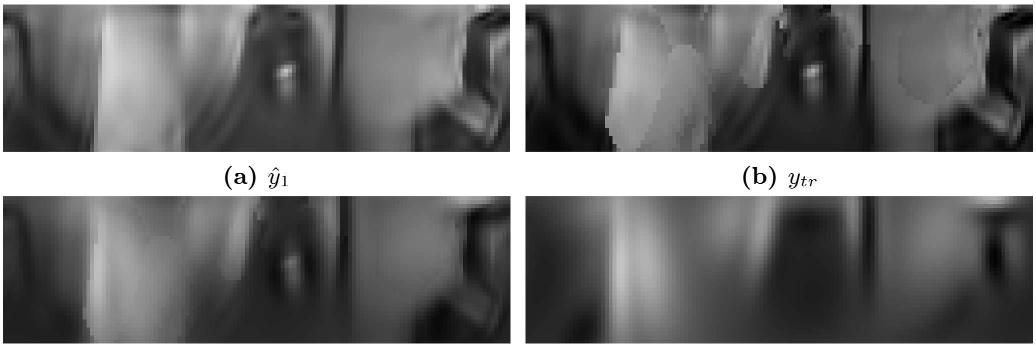


Figure 5.15: *Rate-distortion performance for the Lab image set (VQ Position cosets). Image y_2 is used as the Wyner-Ziv image and image y_1 as the reference image.*



(a) $\hat{y}_1$ **(b)** y_{tr}

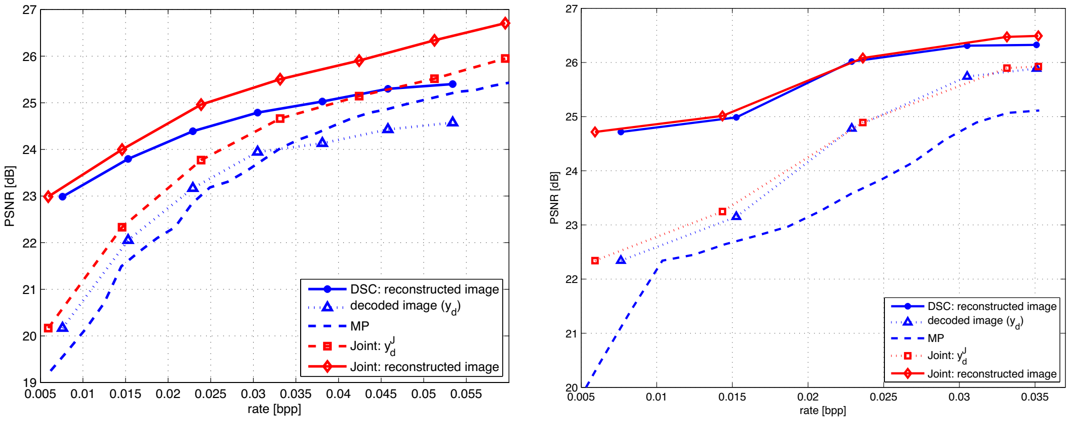
(c) $\hat{y}_2$ **(d)** $\hat{y}_2^{MP}$

Figure 5.16: *DSC results for the Lab images: (a) decoded reference image $\hat{y}_1$ (PSNR=29.4dB); (b) transformed reference image y_{tr}; (c) decoded Wyner-Ziv image $\hat{y}_2$ at 0.035 bpp; (d) decoded second image $\hat{y}_2^{MP}$ when encoded with MP at 0.035 bpp.*

geometry estimation and prediction technique that could constitute a first predictive step in a hybrid DSC coding scheme, similar to block-based disparity estimation method. Our correlation model is certainly more advantageous than the block-based disparity model since it is able to compensate rotation and scale transforms in addition to translations captured by block-based estimation.

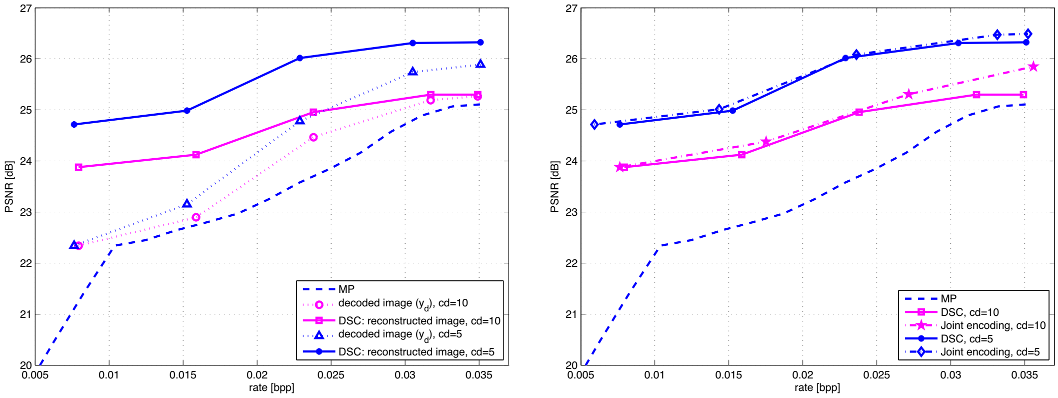
We further evaluate the influence of the camera distance on the performance of the proposed DSC scheme, for the Lab image set. We compare the RD curves for the Wyner-Ziv image y_2 in two cases. In the first case, the Wyner-Ziv image y_2 is decoded using image y_3 as a reference image, while in the second case it is decoded using image y_1 as a reference. The distance between cameras 2 and 3 is set to $cd = 10$, while the distance between cameras 1 and 2 is equal to $cd = 5$. Therefore, the disparity between the images y_2 and y_3 is certainly greater than the disparity between the images y_2 and y_1. For decoding y_2 using y_3 as reference the number of Shape cosets is increased to 256 cosets, due to smaller correlation between the Wyner-Ziv image and the reference image y_3. The number of Position cosets stays unaltered.

Figure 5.18(a) compares the RD performance of the DSC coding of image y_2 with different reference images: y_1 and y_3 corresponding to camera distances $cd = 5$ and $cd = 10$ respectively.



(a) Room image set **(b) Lab image set**

Figure 5.17: *Comparison of rate-distortion performance for distributed coding and joint encoding (VQ coset design).*

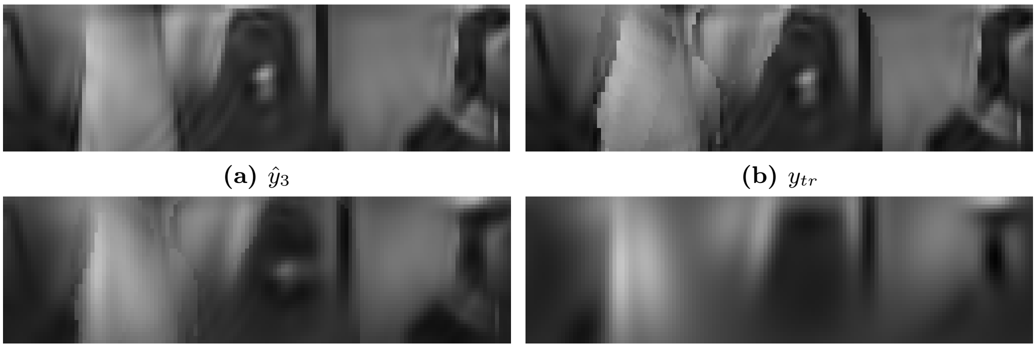


(a) **(b)**

Figure 5.18: *Influence of the camera distance on the DSC performance for the Lab image set: (a) Comparison of the rate-distortion performance for the reference camera distance* $cd = 5$ *and* $cd = 10$*. (b) Comparison of rate-distortion performance for distributed coding and joint encoding for* $cd = 5$ *and* $cd = 10$*.*

We can see that the RD curves for the decoded images y_d in these two cases are close, showing that the atom decoding process based on the proposed correlation model is not much influenced by the increase of disparity between images. The RD curve of the reconstructed image $\hat{y}_2$ for $cd = 5$ outperforms by 1 dB the RD curve of the reconstructed image $\hat{y}_2$ for $cd = 10$. This is due to the higher correlation of the Wyner-Ziv image with the reference image when the camera distance is $cd = 5$ compared to the case when the camera distance is $cd = 10$. Figure 5.18(b) presents the comparisons of the DSC method with the joint encoding method for $cd = 5$ and $cd = 10$, showing that the DSC scheme performs close to the joint encoding strategy for both cases of camera distances. Figures 5.19(a) and (b) show respectively the decoded reference image $\hat{y}_3$ and its transformed version y_{tr}, for the case of $cd = 10$. As in the case of $cd = 5$, the quality of the DSC coded Wyner-Ziv image $\hat{y}_2$, shown in Figure 5.19(c), is higher than the quality of the same image encoded at the equal rate with MP, shown in Figure 5.19(d). This demonstrates the advantage of exploiting the correlation between views in the distributed coding scheme.

We now discuss the efficiency of the geometry-based correlation model. We first analyze the



(a) $\hat{y}_3$ **(b)** y_{tr}

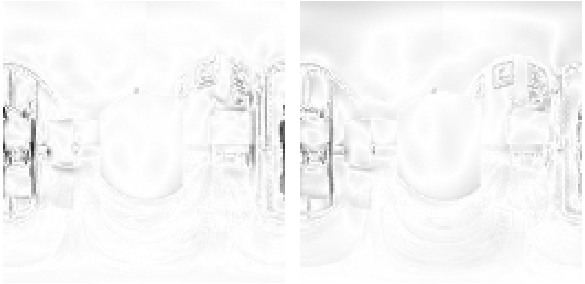
(c) $\hat{y}_2$ **(d)** $\hat{y}_2^{MP}$

Figure 5.19: *DSC results for the Lab images: (a) decoded reference image* $\hat{y}_3$ *(PSNR=31.82dB); (b) transformed reference image* y_{tr}*; (c) decoded Wyner-Ziv image* $\hat{y}_2$ *at 0.035 bpp; (d) decoded second image* $\hat{y}_2^{MP}$ *when encoded with MP at 0.035 bpp.*

residue after DSC coding, denoted with $e_2 = \hat{y}_2 - y_2$, for the Room image set. We compare e_2 with the difference between the decoded reference image and the original Wyner-Ziv image $e_1 = \hat{y}_1 - y_2$. Figures 5.20(a) and (b) show respectively the images $1 - |e_1|$ and $1 - |e_2|$. Namely, they show the residues $|e_1|$ and $|e_2|$ with the inverted pixel magnitude scale, such that the white pixels represent zero error. The energy of the difference between the decoded reference image and the original Wyner-Ziv image $|e_1|$ is 82.65, where the energy is given by the square of the norm with the inner product computed on the sphere. On the other hand, the energy of the error e_2 is much smaller and equal to 47.12. This confirms the efficiency of the model based on local geometric transforms. Whereas in the residue $|e_1|$ displacements of objects result in high error areas (dark parts), the residue after DSC decoding e_2 is almost exclusively composed of high frequencies since the object transforms have been captured efficiently. Similar results have been obtained for the Lab image set. Figures 5.21(a) and (b) show respectively the inverted residues $1 - |e_1|$ and $1 - |e_2|$ for camera distance $cd = 5$, while Figures 5.22(a) and (b) show the inverted residues $1 - |e_3| = 1 - |\hat{y}_3 - y_2|$ and $1 - |e_2|$ for $cd = 10$. Again, the energies $\|e_1\|^2 = 37.89$ and $\|e_3\|^2 = 52.9661$ are much higher than the energies of the error e_2, which is 11.0380 for the case of $cd = 5$ and 13.9376 for the case of $cd = 10$. Moreover, we show in Figure 5.23(b) that the distribution of the pixel values in the residue image after transform compensation and decoding can be well modeled with the Laplace distribution, which is not the case for the distribution of pixels in the difference between the decoded reference image and the original Wyner-Ziv image (see Figure 5.23(a)). As mentioned in Section 5.4.2, the introduced DSC scheme that exploits the geometric correlation between multi-view images can be combined with another DSC approach based on texture correlation modeling in a hybrid coding scheme. The well-fitted Laplacian distribution of the residue after the geometry-based DSC shows that applying our method in the first step of a hybrid scheme would greatly facilitate the correlation modeling of the residual texture information.

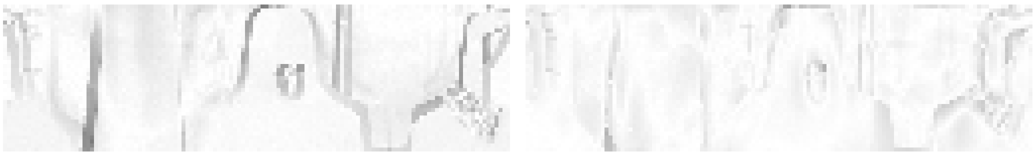
5.4.4 Performance of the occlusion-resilient Wyner-Ziv coder

Finally, we examine the performance of the occlusion-resilient Wyner-Ziv scheme where LDPC codes are used for syndrome coding [LDP]. The rate of the LDPC code is chosen to be below the BSC channel capacity with the crossover probability M/N. The RD curves of the occlusion-resilient Wyner-Ziv scheme for the Room and Lab scenes are given in Figure 5.24(a) and (b), respectively. We can clearly see that the occlusion-resilient coding corrects the saturation behavior of the previous scheme, and that it performs close to the joint encoding scheme. It can also slightly outperform the joint encoding scheme when the number of correspondences gets small. This is because the joint encoder sends the full description for atoms without a correspondence, while



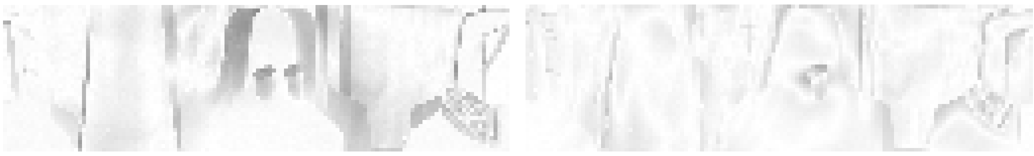
(a) $1 - |\hat{y}_1 - y_2|$ **(b)** $1 - |\hat{y}_2 - y_2|$

Figure 5.20: *Room image set: (e)* $1 - |e_1| = 1 - |\hat{y}_1 - y_2|$*, where* e_1 *is the difference between the decoded reference image and the original Wyner-Ziv image (white pixel denotes no error); (f)* $1 - |e_2| = 1 - |\hat{y}_2 - y_2|$*, where* e_2 *is the residue after DSC decoding.*



(e) $1 - |\hat{y}_1 - y_2| = 1 - |e_1|$ **(f)** $1 - |\hat{y}_2 - y_2| = 1 - |e_2|$

Figure 5.21: *Lab image set (cd* $= 5$*): (a)* $1 - |e_1| = 1 - |\hat{y}_1 - y_2|$*, where* e_1 *is the residue without transform compensation (white pixel denotes no error); (b)* $1 - |e_2| = 1 - |\hat{y}_2 - y_2|$*, where* e_2 *is the residue after DSC decoding.*



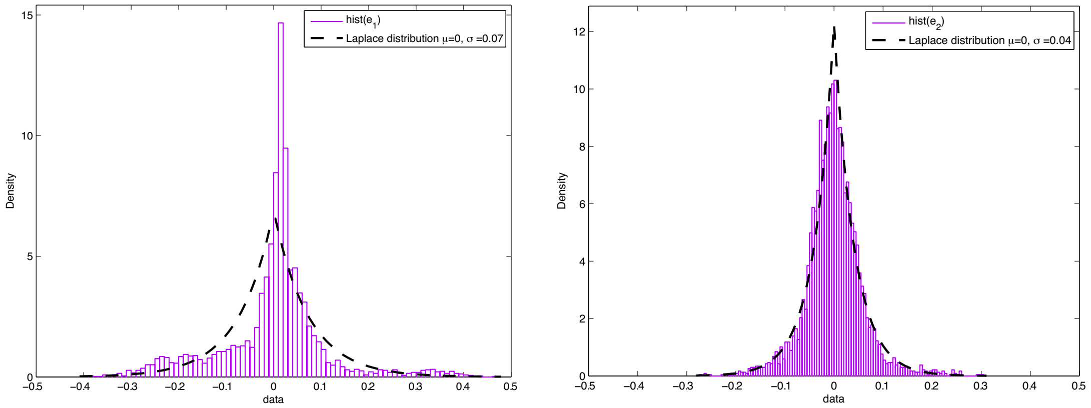
(a) $1 - |\hat{y}_3 - y_2| = 1 - |e_3|$ **(b)** $1 - |\hat{y}_2 - y_2| = 1 - |e_2|$

Figure 5.22: *Lab image set (cd* $= 10$*): (a)* $1 - |e_3| = 1 - |\hat{y}_3 - y_2|$*, where* e_3 *is the residue without transform compensation (white pixel denotes no error); (b)* $1 - |e_2| = 1 - |\hat{y}_2 - y_2|$*, where* e_2 *is the residue after DSC decoding.*

the Wyner-Ziv encoder still sends only the coset index and the overall syndrome. We have also compared the visual quality of the Wyner-Ziv image encoded by the occlusion-resilient WZ coder with the same image independently encoded using the Matching Pursuit, at very low bit rate. Figure 5.25(a) shows the image $\hat{y}_2$ decoded by the occlusion-resilient Wyner-Ziv coder, at rate $0.0527bpp$. We can see that the occlusion-resilient WZ coder gives better visual image quality compared to the MP decoded image at a close rate $0.0519bpp$, (Figure 5.25). Similar observation holds for the Lab scene, where the Wyner-Ziv encoded image at rate $0.0431bpp$ in Figure 5.26(a) contains more geometry information than the MP encoded image at a similar rate $0.0436bpp$ (Figure 5.26(b)).

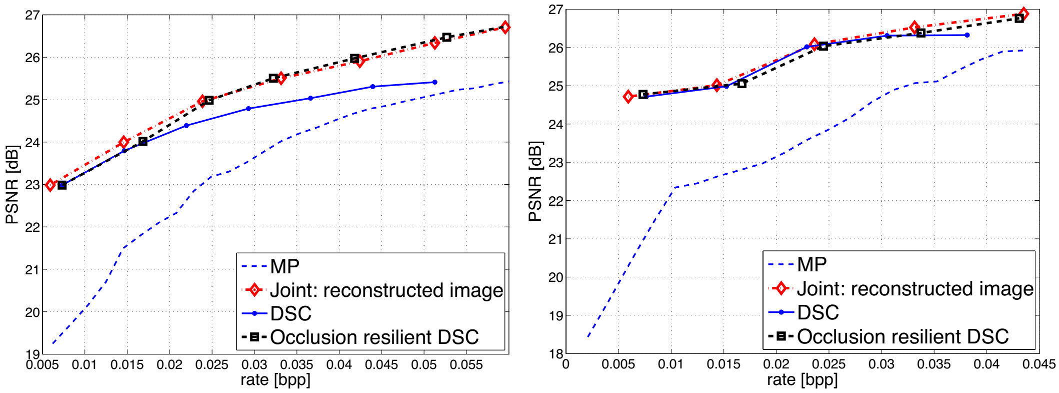
5.5 Related Work

Distributed source coding has been researched for a long time in the information theory community, but its application to imaging problems has been delayed due to the difficulty of finding good



(a) **(b)**

Figure 5.23: *(a) Laplacian distribution fitted to the histogram of the residue e_3 for the Lab scene with $cd = 10$ (b) Laplacian distribution fitted to the histogram of the residue e_2 with $cd = 10$ (fitting is performed with the Matlab statistics toolbox, using the maximum likelihood estimator).*



(a) Room scene (b) Lab scene with camera distance 5

Figure 5.24: *Rate-distortion performance of the occlusion-resilient Wyner-Ziv encoder for the image y_2.*

models of correlation between real sources. The first practical DSC schemes for jointly Gaussian sources have been proposed only recently by Pradhan and Ramchandran who constructed the algebraic trellis codes for the distributed coding scenario [Pra03]. Their work was inspired by the link of DSC with channel coding that was originally established by Wyner [Wyn74]. Most of the previous research in the DSC framework has focused on the application of DSC to low-complexity video coding [Gir05, Pur07] and error-resilient video coding [Seh04, Gir05]. Yeo et al. [Yeo07] have used the DSC principles for the error-resilient delivery of multi-view video in wireless camera networks. The DSC principles have been also exploited in a camera array to achieve video decoding that is flexible with respect to the choice of a predictor [Che07]. However, only few works have addressed the problem of distributed coding in camera networks, mainly due to the difficulty of modeling the geometric correlation among distributed cameras for 3D scene representation.

The application of DSC principles in camera networks is generally based on the disparity estimation between views, under epipolar constraints. Most of the solutions proposed in the literature are built on coding with side information, which is a special case of DSC. For example, cameras can be divided into conventional cameras that perform independent image coding, and Wyner-Ziv

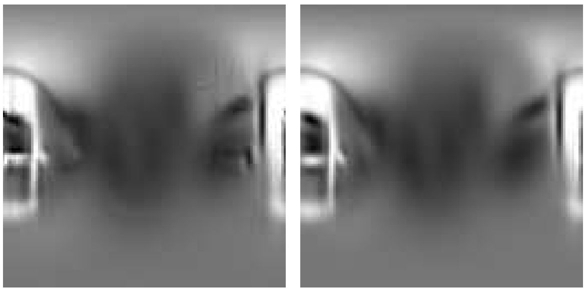

(a) $\hat{y}_2$ **(b)** $\hat{y}_2^{MP}$

Figure 5.25: *Room scene: Visual comparison at very low bit rate of the decoded image* y_2 *between the (a) occlusion-resilient Wyner-Ziv coding at* $0.0527bpp$ *and* $PSNR = 26.47dB$*, and (b) Matching Pursuit coding at* $0.0519bpp$ *and* $PSNR = 25.14dB$.

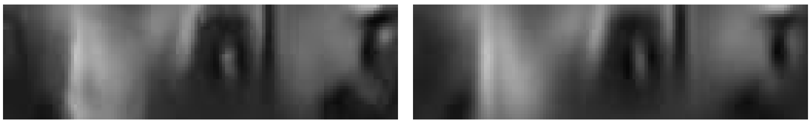

(a) $\hat{y}_2$ **(b)** $\hat{y}_2^{MP}$

Figure 5.26: *Lab scene: Visual comparison at very low bit rate of the decoded image* y_2 *between the (a) occlusion-resilient Wyner-Ziv coding at* $0.0431bpp$ *and* $PSNR = 26.76dB$*, and (b) Matching Pursuit coding at* $0.0436bpp$ *and* $PSNR = 25.92dB$.

cameras that use DSC coding [Zhu03]. The Wyner-Ziv images are decoded with the help of disparity estimation and interpolation from independent views. Shape adaptation is used to enhance the side information, where the shape information is sent by the encoders. Super-resolution techniques have been also applied to distributed coding in camera networks [Wag03]. Low-resolution images from each camera are combined after registration at the joint decoder into a high-resolution image. Image registration is performed by shape analysis and image warping with respect to the shape transforms that are limited to simple translations and rotations. Gehrig and Dragotti have proposed a distributed coding scheme for camera networks where the multi-view correlation is modeled by relating the locations of discontinuities in the polynomial representation of image scanlines [Geh05]. This scheme has been extended to the case of natural 2D images [Geh07, Geh09]. Among state of the art methods, the geometric approach for distributed coding by Gehrig and Dragotti is the closest to the work proposed in this chapter in the sense that it exploits the epipolar constraint for the design of Slepian-Wolf code and for the joint decoding. However, they consider only translations as correlation in multiple views (shifts of the discontinuities of the piecewise polynomials), while the correlation model in this chapter includes translations, rotations and anisotropic scaling in a single framework.

Disparity-based solutions have also been proposed for distributed multi-view video compression. Several works build on the advantages of distributed video coding for low complexity encoding and distributed multi-view coding for exploiting both temporal and inter-view correlation [Fli06, Oua06a, Guo06]. They take different approaches for modeling the correlation among views, like the disparity-based model [Fli06], affine model [Guo06], or homography-based model [Oua06a]. Another direction for the distributed multi-view video compression is based on classical motion compensated video encoding at each camera, while the inter-view correlation is exploited in a distributed manner [Son07, Yan07a, Yan07b]. Song et al. have presented a transform-based DSC method for multi-view video coding that tracks epipolar correspondences between macroblocks in different views [Son07]. The geometric information given by the epipolar constraint is exploited for joint multi-view video decoding, but not for the design of the Slepian-Wolf code. Moreover, the Wyner-Ziv encoder has partial access to the side information (Intra macroblocks and motion vectors), so that this scheme cannot be classified as fully distributed multi-view coding scheme. On the other hand, a completely distributed stereo-view video coding method has been proposed by Yang et al. [Yan07a]. Their method performs independent coding of I-frames and Wyner-Ziv coding of P frames, where the side information is generated by fusing the disparity map with the motion field. However, it does not exploit the correlation among I-frames and thus the achieved bit rates are still quite far from the Slepian-Wolf bound. This gap can be reduced by encoding more coarsely the I-frames [Yan07b], but a lot of geometric correlation between I-frames is still left unexploited. The work presented in this chapter is substantially different from [Son07, Yan07b] since it uses epipolar geometry information for the design of the Slepian-Wolf code and therefore exploits the correlation between multi-view images in a completely distributed manner.

The common characteristic of most state of the art disparity-based distributed coding frameworks (except [Geh07]) is the need of at least two independently encoded views in order to perform disparity estimation for DSC decoding, which leads to high encoding rates. Moreover, the disparity estimation usually requires high-resolution images, which is quite restricting in practical camera network scenarios. The work that we propose in this chapter contributes to solving these two main problems by efficiently relating the correlated data in multiple views under geometric local transforms. This enables estimation of scene geometry and successful decoding of Wyner-Ziv frames, even with a single reference frame that has been highly compressed. Finally, a very important property of the proposed method is that it does not require a special camera arrangement in a camera network, since the geometric correlation model can cope with various local transforms. This makes the applicability of our method more generic than that of distributed coding methods designed for camera arrays.

5.6 Conclusion

The contribution of this chapter is a novel geometry-based framework for the efficient representation of 3D scenes, where camera images are approximated by a sparse expansion of prominent geometric features. We have exploited the novel correlation model introduced in Chapter 4 in order to pair atoms in different images under shape and epipolar geometry constraints. The geometric correlation model leads to the design of a distributed coding algorithm in camera networks and, as we will see in Chapter 6, provides an estimation of the scene geometry. We have built on this novel framework and designed an occlusion-resilient distributed coding scheme with side information that offers an efficient rate-distortion representation of 3D scenes at low bit rate. Although we have presented the DSC results for omnidirectional images (due to the context of the thesis), the framework can be also applied for perspective images using the appropriate dictionary and the epipolar geometry for a given camera network. Since the performance of the new DSC method depends on the number of atom correspondences in different views, learning the overcomplete dictionary to contain features leading to good epipolar matching can boost the RD performance of the DSC scheme. We will show in Chapter 7 how this dictionary learning is achieved.

Chapter 6

Coarse Scene Structure Estimation

6.1 Introduction

In Chapter 4 we have presented a new correlation model for multi-view images, based on sparse image representation with geometric dictionaries. Further on, in Chapter 5 we have demonstrated how this model can be applied for efficient compression in omnidirectional camera networks. In order to achieve good rate-distortion performance in a distributed coding scheme, we have exploited the fundamental multi-view geometry constraint, i.e., the epipolar constraint. Hence, we have assumed that the relative camera pose, defined by the rotation matrix $\mathbf{R}$ and the translation vector $\mathbf{T}$, between two cameras is known. In this chapter we perform a different task: knowing the atoms that form sparse approximations of multi-view images, we want to estimate the camera pose $(\mathbf{R}, \mathbf{T})$. We show that we can use the proposed multi-view correlation model to identify corresponding pairs of features represented by sparse components, and to estimate the relative camera pose from those features. Furthermore, such correspondences and the local transforms between them can be exploited in order to construct a disparity map between images and to estimate coarse depth information. Since the main scene features are captured with only few basis vectors, camera pose and disparity map can be obtained from very coarse descriptions of multi-view images. Such benefits are quite interesting in the design of efficient multi-view coding strategies for 3D scene respresentation.

Estimation of the camera pose $(\mathbf{R}, \mathbf{T})$, or equivalently estimation of the essential matrix $\mathbf{E} = \hat{\mathbf{T}}\mathbf{R}$, is a well known problem in computer vision. Typically, the camera pose estimation consists of two phases: 1) matching of corresponding image features across two views, and 2) estimation of the pose by fitting rotation and translation parameters to the selected set of corresponding points. One of the fundamental correspondence-based algorithms for camera pose estimation is the eight-point algorithm [Ma 04, Har97]. It requires eight pairs of corresponding points to form a set of linear equations from which the essential matrix is computed. The corresponding points in multiple views are usually found by matching similar local image features, like for example the well-known SIFT features [Low04]. SIFT features have been widely employed in many problems in computer vision, among which are pose estimation and camera localization [Kos04, Se 01]. However, one of the main problems in correspondence matching is dealing with gross outliers, which can make the fitting of camera parameters very imprecise. Improvement of the pose estimation can be achieved using algorithms such as the Random Sample Consensus (RANSAC) [Fis81]. RANSAC is an iterative algorithm that selects randomly at each iteration a subset of data points to perform the model fitting, and then checks the accordance of other data points to the fitted model. The algorithm stops when most data points agree on the fitted model. RANSAC has been successfully applied to the camera pose estimation problem [Tor93, Tor97, Qia04, Kos04]. Since the pose estimation method presented in this chapter recovers the matrix $\mathbf{R}$ and the vector $\mathbf{T}$ by finding atom correspondences, it is sensitive to outliers. Therefore, we propose to complement the pose estimation method with the RANSAC algorithm to increase its robustness and reduce the estimation error.

Our goal in this chapter is to show that the multi-view correlation model presented in Chapter 4 inherently contains the information about the 3D structure of the scene with the local transforms of the sparse features. The proposed framework differs from the state of the art geometry estimation approaches because it is based on the multi-view image representation that is beneficial both for image compression and geometry estimation. We consider the atoms as low-level visual features that minimize the entropy of image representation and carry important information for the estimation of the scene geometry. In that sense, our work is closer to the work of Jones and Malik [Jon92b, Jon92a] who propose a framework for determining the stereo correspondences from a set of oriented linear filters obtained from stereo images.

6.2 Identification of corresponding features

As we have seen in Chapter 4, given sparse representations of two correlated images capturing the same 3D scene, i.e., $y_1 = \mathbf{\Phi}_{I_1} c_1$ and $y_2 = \mathbf{\Phi}_{I_2} c_2$, there exists a subset of atoms indexed respectively by $J_1 \in I_1$ and $J_2 \in I_2$ that represent image projections of the same 3D features in the scene. Due to camera displacement, these atoms are correlated by local geometric transforms F_i, as given in Eq. (4.3). Furthermore, in Section 4.4 we have defined two types of geometric constraints that narrow the set of possible local transforms that appear in correlated multi-view images. These are the shape similarity constraint (in Eq. (4.7)) and the epipolar geometry constraint (in Eq. (4.10)). Other than by geometric constraints, we can relate atoms by their inner products with the corresponding images, as given by Eq. (5.7).

The correlation model given by the local transforms that satisfy these geometric constraints relies on the use of a structured dictionary of geometric atoms for sparse image approximation. When the dictionary of geometric atoms is used, the correlation model carries significant information about the scene geometry. Namely, by identifying the transforms of the atoms in one view into the atoms in another view, we can also identify the motion of objects in the scene resulting from the viewpoint change. Therefore, to estimate the camera pose ($\mathbf{R}$, $\mathbf{T}$) we need to identify the corresponding atoms and their transforms between two views, using the geometric correlation model. However, the epipolar geometry constraint requires the knowledge of $\mathbf{R}$ and $\mathbf{T}$ and thus cannot be used in atom matching for pose estimation. Nevertheless, we can use the shape similarity constraint and the relation in Eq. (5.7) to find corresponding atoms between two views.

From the found atom correspondences, we further have to extract point correspondences between views, which are necessary for evaluating the essential matrix using standard algorithms, such as the eight-point algorithm. The easiest way to do this is to simply take atom positions (i.e., their localization) defined by their parameters τ and ν. This would lead to one point correspondence for each atom correspondence. However, we can extract more point correspondences in the local neighborhoods where the corresponding atoms have high energy. Namely, due to their geometric nature, the pairing of atoms by local transforms offers a geometrical mapping of the local neighborhood of the observed features. This mapping is essentially the estimation of the local disparity map that we have defined in Section 4.5. Local disparity map relates a point P_1 on the atom g_{γ_i} in the image y_1, with a point P_2 on h_{γ_j} in y_2, such that:

$$\mathbf{z}_2 = \mathbf{R}_{\gamma_j}^{-1} \cdot \zeta(\mathbf{R}_{\gamma_i} \cdot \mathbf{z}_1), \tag{6.1}$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ represent respectively the Euclidean coordinates of points P_1 and P_2. $\mathbf{R}_{\gamma_i}$ and $\mathbf{R}_{\gamma_j}$ are rotation matrices given by Euler angles (τ_i, ν_i, ψ_i) and (τ_j, ν_j, ψ_j), respectively, and $\zeta(\cdot)$ defines the grid transform due to anisotropic scaling, as given in Eq. (4.14). Finally, the local disparity maps obtained from all pairs of atom correspondences between images y_1 and y_2 are combined into a global disparity map by selecting the most confident mapping for each point $\mathbf{z}_2$ from different mappings $\mathbf{z}_1^{(i)}, i = 1, ..., n$, defined by n correspondences. The final mapping point $\mathbf{z}_1^*$ is selected as:

$$\mathbf{z}_1^* = \arg \max_{\mathbf{z}_1^{(i)}, i=\overline{1,n}} w_{\gamma_i}(\mathbf{z}_1^{(i)}), \tag{6.2}$$

where the confidence w_{γ_i} is defined in the same way as for the symmetric epipolar distance given in Eq. (4.15). Therefore, each point $\mathbf{z}_1^*$ and its disparity mapping point $\mathbf{R}_{\gamma_j}^{-1} \cdot \zeta(\mathbf{R}_{\gamma_i} \cdot \mathbf{z}_1^*)$ represent a point correspondence pair. Depending on the resolution of the image and the weights w_{γ_i}, we obtain different number of point correspondences for each atom correspondence.

6.3 Camera pose recovery

6.3.1 Eight-point algorithm

The eight point algorithm [Ma 04] is one of the fundamental methods for estimating camera pose $\mathbf{R}$, $\mathbf{T}$ from a set of q point correspondences $(\mathbf{z}_1^k, \mathbf{z}_2^k)$, $k = 1, ..., q$, $(q \geqslant 8)$. It consists of three basic steps:

1. **Find an approximation of the essential matrix $\mathbf{E} = \hat{\mathbf{T}}\mathbf{R}$**
Evaluate the nine-dimensional vectors $\mathbf{a}^k = \mathbf{z}_1^k \otimes \mathbf{z}_2^k$, where $\otimes$ denotes the Kronecker product). Compute the singular value decomposition of the matrix $\chi = [\mathbf{a}^1, \mathbf{a}^2, ..., \mathbf{a}^q]^\mathsf{T} \in \mathbb{R}^9$. From the singular value decomposition $\chi = \mathbf{U}_\chi \mathbf{\Sigma}_\chi \mathbf{V}_\chi^\mathsf{T}$ take $\mathbf{E}^s$ to be the ninth column of $\mathbf{V}_\chi^\mathsf{T}$, and form the approximation of the essential matrix by unstacking nine elements of $\mathbf{E}^s$ into a square 3×3 matrix $\mathbf{E}$.

2. **Project onto the essential space**
Compute the singular value decomposition of the approximated matrix $\mathbf{E}$ to be: $\mathbf{E} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\mathsf{T} = \mathbf{U} diag\{\sigma_1, \sigma_2, \sigma_3\} \mathbf{V}^\mathsf{T}$, where $\sigma_1 \geqslant \sigma_2 \geqslant \sigma_3 \geqslant 0$ and $\mathbf{U}, \mathbf{V} \in SO(3)$. In general, the matrix $\mathbf{E}$ is not in the essential space ($\sigma_1 \neq \sigma_2$ and $\sigma_3 \neq 0$), so the final approximation of the essential matrix is the projection of the matrix $\mathbf{E}$ onto the normalized essential space, evaluated as $\mathbf{E} = \mathbf{U} diag\{1, 1, 0\} \mathbf{V}^\mathsf{T}$.

3. **Recover $\mathbf{R}, \mathbf{T}$ from $\mathbf{E}$**
$\mathbf{R}$ and $\mathbf{T}$ are extracted from the essential matrix as:

$$\mathbf{R} = \mathbf{U}\mathbf{R}_Z^\mathsf{T}(\pm\frac{\pi}{2})\mathbf{V}^\mathsf{T}, \hat{\mathbf{T}} = \mathbf{U}\mathbf{R}_Z(\pm\frac{\pi}{2})\mathbf{\Sigma}\mathbf{U}^\mathsf{T}, \tag{6.3}$$

where

$$\mathbf{R}_Z^\mathsf{T}(\pm\frac{\pi}{2}) = \begin{pmatrix} 0 & \pm 1 & 0 \\ \mp 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$\hat{\mathbf{T}} = \begin{pmatrix} 0 & -\mathbf{T}(3) & \mathbf{T}(2) \\ \mathbf{T}(3) & 0 & -\mathbf{T}(1) \\ -\mathbf{T}(2) & \mathbf{T}(1) & 0 \end{pmatrix}.$$

The algorithm gives four solutions for $\mathbf{R}, \mathbf{T}$, but three of them can be eliminated by imposing the positive depth constraint.

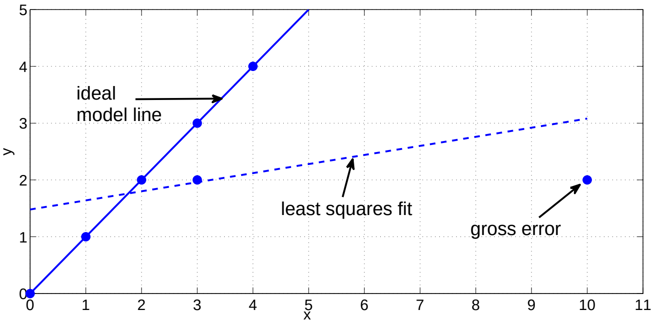
As described in Section 6.2 each atom correspondence gives a set of point correspondences by disparity mapping. Disparity map from all atom correspondence pairs, obtained by Eq. (6.2), gives a set of q point correspondences $(\mathbf{z}_1^k, \mathbf{z}_2^k)$, $k = 1, ..., q$, $(q \geqslant 8)$. Therefore, we can estimate the camera pose $(\mathbf{R}, \mathbf{T})$ from atom correspondences by applying the eight-point algorithm to all point correspondences, extracted from all atom pairs between images y_1 and y_2. Alternatively, we can estimate the camera pose by taking only the atom centers. This is a special case of the proposed method when $w_{\gamma_i}(\theta, \varphi)$ is equal to one for $\theta = \tau_i, \varphi = \nu_i$ and is zero elsewhere. However, the advantage of taking more point correspondences per atom pair is that we have a much larger number of epipolar constraints coming from the local neighborhood covered by the matching atoms, and hence possibly a better estimate.

6.3.2 Robust estimation with RANSAC

The estimation of camera parameters from a set of point correspondences is in essence a problem of model fitting from a set of observed data. However, fitting the model parameters to all available data can lead to large errors when data contains gross outliers. An example (from [Fis81]) where the least squares solution to the linear fitting of data containing one large outlier results in a completely wrong estimate of the linear model is shown in Figure 6.1.

Points	1	2	3	4	5	6	7
x	0	1	2	3	3	4	10
y	0	1	2	2	3	4	2

(a)



(b)

Figure 6.1: *Example of an erroneous least squares linear fit due to a gross outlier [Fis81]. a) Set of data points representing measurements of a linear model. Point 7 is a gross outlier. b) Least square fitting using all points is given with a dashed line, while the ideal model is given with a solid line, showing the big error in the model estimation.*

To make the model fitting more robust to gross outliers, Fischler and Bolles [Fis81] introduced the Random Sample Consensus (RANSAC) algorithm. Instead of using all available data to estimate the parameters of the model, RANSAC uses a subset of data containing a minimal number of points needed to perform initial model fitting. Afterwards, during the iterative procedure, the data set can be enlarged if the points are consistent with the current model estimate. Formally, RANSAC consists of three steps:

1. Let the whole data set P contain q points, and the model to be estimated has m free parameters ($q \gg m$). RANSAC randomly selects a subset S_1 of m data points from P and estimates the initial model M_1. Using the initial model M_1 it then determines the subset S_1^* of points in P that are within some error tolerance of M_1. The set S_1^* is called the consensus set of S_1.

2. If the number of points in S_1^* is greater than some threshold t, the algorithm computes a new model M_1^* using all points in S_1^*. The threshold t is a function of the estimate of the number of gross errors in P.

3. If the number of points in S_1^* is less than t, RANSAC randomly selects a new subset S_2 and repeats the steps 1-2. After some predetermined number of iterations, if the algorithm did not find a consensus set with t or more points, it either solves the model with the largest consensus set found, or terminates in failure.

Fischler and Bolles have suggested in [Fis81] to use the maximal number of trials l that exceed for two or three times the standard deviation of the expected number of trials required to select m good data points. They have derived this standard deviation and obtained:

$$SD(l) = \frac{\sqrt{(1-p^m)}}{p^m}, \tag{6.4}$$

where p is the probability that any selected data point is within the error tolerance of the model.

For the selection of the threshold t there is no analytic estimate. However, if the probability that any given data point is within the error tolerance of an incorrect model is less than 0.5, a value of $t = m + 5$ will provide less than 5% probability of matching with an incorrect model [Fis81].

RANSAC has been successfully used for the estimation of the camera pose and other problems that involve matching features in images. In our case of atom matching, false atom correspondences lead to point correspondences that represent gross outliers. Therefore, given a set of q identified point correspondences using atom matching, we can randomly select eight correspondences, evaluate the initial camera pose using the eight point algorithm, and then follow the RANSAC steps until a correct camera pose is estimated.

6.4 Depth estimation

In the depth estimation problem the camera pose $\mathbf{R}, \mathbf{T}$ is known. Therefore, we can use the epipolar geometry constraint in addition to the shape constraint and Eq. (5.7) to perform correspondence matching and to identify the local transforms of objects induced by the viewpoint change. From atom pairs, the global disparity map between images y_1 and y_2 is estimated as described in Section 6.2, giving a set of q point correspondences $(\mathbf{z}_1^k, \mathbf{z}_2^k), k = 1, ..., q$. The depth of the 3D point whose projections on two spherical images are given with $\mathbf{z}_1$ and $\mathbf{z}_2$, related by the disparity map, can be evaluated as:

$$\rho = \frac{|\mathbf{T} \times \mathbf{R}\mathbf{z}_2|}{|\mathbf{z}_1 \times \mathbf{R}\mathbf{z}_2|}, \tag{6.5}$$

where '×' denotes the cross product. When $\mathbf{R}, \mathbf{T}$ is given as a relative pose of camera 2 with respect to camera 1, the depth ρ is evaluated also with respect to camera 1. By computing ρ for each pair of points $(\mathbf{z}_1^k, \mathbf{z}_2^k), k = 1, ..., q$ we get a depth map of the observed 3D scene.

6.5 Experimental results

6.5.1 Setup

The performance of the proposed scene geometry estimation method has been evaluated on a simple synthetic scene and on a natural scene captured with omnidirectional cameras. The sparse image decomposition has been obtained using the Matching Pursuit algorithm on the sphere (SMP) with the same dictionary as for compression of omnidirectional images (built on low frequency and high frequency generating functions given by Eq. (3.12) and Eq. (3.11), respectively). The discretization of the dictionary is the same as in Section 3.4.2.

6.5.2 Disparity and depth estimation results for a synthetic 3D scene

For a simple and comprehensive illustration of the proposed scene geometry estimation method based on sparse image decomposition, we first observe the simple synthetic Shapes scene (see Section 4.6). The original images are shown in Figures 6.2(a) and (b) (unwrapped) and the approximated images are shown in Figures 6.2(c) and (d), using six atoms per image. Three pairs of atoms in two decompositions have been recognized as correspondences, where each of the atoms corresponds to one of the objects. Using the transforms of only these three atoms, the disparity map has been estimated and applied to the image y_1 to reconstruct an approximation

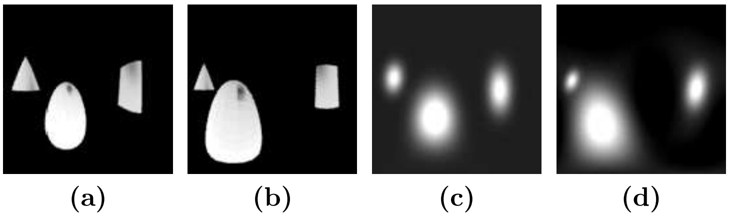


Figure 6.2: *Original synthetic images: a) y_1, b) y_2. Approximated synthetic images: c) $\tilde{y}_1$, with 3 atoms, d) $\tilde{y}_2$ with 6 atoms.*

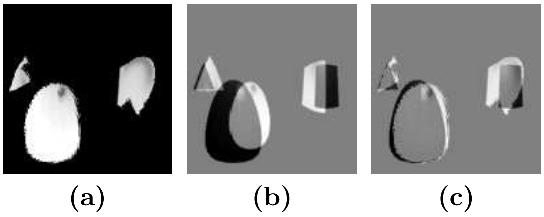


Figure 6.3: *a) Estimation of y_2 from y_1, using the disparity map: $\hat{y}_2$; b) $y_1 - y_2$; c) $\hat{y}_2 - y_2$.*

of the image y_2, as shown in Figure 6.3(a). Disparity maps from different atom pairs have been combined in a global disparity map as described in Section 6.2. This simple example shows that we can obtain a good estimate of the disparity between two images of the same scene, in using only coarse approximations of the original images (shown in Figures 6.2(c) and (d)). Figure 6.3(b) displays the difference between the two original images, while the Figure 6.3(c) shows the difference between the original and reconstructed image y_2. The range of values for both images is [-1,1]. We can observe that many high valued errors (white and black pixels) are removed or reduced by applying the disparity map as shown in Figure 6.3a). Therefore, the disparity map captures the transforms of the three objects.

In order to show that we can also extract object depths for the observed scene only from the three obtained transforms, we have evaluated the mean distance from each of the objects to the camera 1 based on the obtained disparity map, when the relative camera pose $(\mathbf{R}, \mathbf{T})$ is known. The results are shown in Table 6.1, where the estimated distances are compared with the original distances of objects. The estimated distances are measured at pixels on the spherical image that correspond to centers of the three MP atoms. We can see that the depths to the cube and the cone are well estimated, while the distance to the sphere is quite erroneous (probably due to its large size).

Table 6.1: *Comparison of estimated depths with the ground truth for the simple scene.*

Object	Ground truth distance	Mean estimated distance	Estimation error
Sphere	1.0	1.54	54%
Cube	2.0	2.29	14.5%
Cone	1.7	1.82	7%

6.5.3 Disparity and camera pose estimation results for a natural scene

The natural Lab scene images shown in Figure 6.4 have been decomposed with the SMP algorithm. The omnidirectional cameras have been placed in a line (no rotation, same altitude) where successive images are denoted y_3, y_1 and y_2. The proposed method is used to recover camera pose between camera images y_3 and y_2. The threshold s on the coherence in Eq. (4.6) is set to $s = 0.4$ to capture as many correspondences as possible and thus get more accurate results. For each

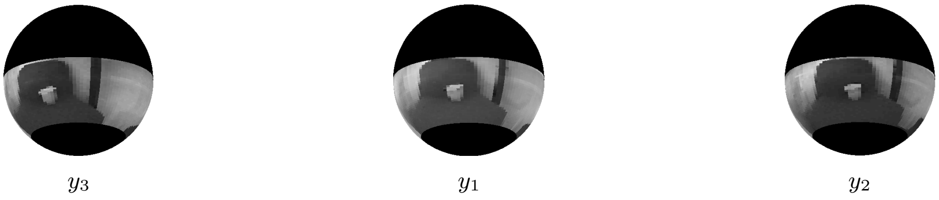


Figure 6.4: *Lab images.*

atom, the correspondences are sought in the neighborhood of the solid angle $\pi/4$. The camera pose $(\mathbf{R}, \mathbf{T})$ is recovered from 18 correspondences, as described in the Sec. 6.3. Due to the camera placement in a line, the rotation matrix between them should be equal to the identity matrix and the translation vector to $[1 \;\; 0 \;\; 0]^{\mathsf{T}}$, in the ideal case. With the proposed method, we obtain the following result for the rotation matrix:

$$\mathbf{R} = \begin{pmatrix} 1.0000 & 0.0000 & -0.0015 \\ -0.0002 & 0.9914 & -0.1307 \\ 0.0015 & 0.1307 & 0.9914 \end{pmatrix}$$

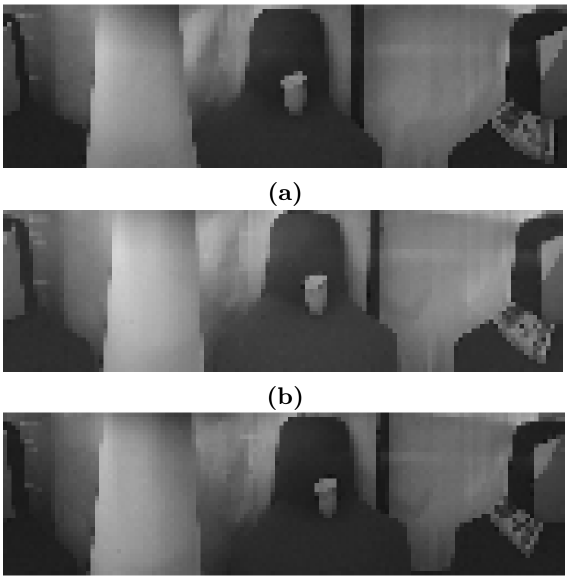
and translation vector: $\mathbf{T} = [0.9909 \;\; -0.1322 \;\; -0.0252]^{\mathsf{T}}$. Therefore, $\mathbf{R}$ and $\mathbf{T}$ have been recovered with a small error with respect to the ground truth values.

The unwrapped original natural images of the Lab scene are presented in Figure 6.5. Figure 6.6 shows the reconstructed image y_2 obtained by disparity mapping from images y_3 and y_1. This is performed by first identifying two sets of atom correspondences: atom pairs between images y_3 and y_2, and atom pairs between y_1 and y_2. Then, for each pixel $\mathbf{z}_2$ in y_2 we find its corresponding pixel either in y_3 or y_1. For this, we use Eq. (6.2) where $\mathbf{z}_1$ and γ_i can be from both sets of corresponding atom pairs, i.e., from both reference images y_3 and y_1. The reconstructed image $\hat{y}_2$ is then obtained by taking for each pixel the value of its corresponding pixel in y_3 or y_1. For the disparity mapping, we have used $s = 0.9$ in order to have a consistent and smooth disparity map. The differences between the images y_2 and reference images y_3 and y_1 are shown in Figure 6.7(a) and (b) respectively, with the range of [0, 1], where 1 (white) means no error. Figure 6.7(c) illustrates the difference between the original image y_2 and its reconstruction with the proposed disparity mapping. For a quantitative comparison of the three residual images, we evaluate their energies. The first residual $y_3 - y_2$ carries the highest energy of 89.2, followed by the $y_1 - y_2$ with the energy 77, while the residue after disparity mapping with the proposed scheme resulted in the lowest energy of 49.7, confirming the benefits of our method. Once again, we can see that the obtained disparity map succeeds in compensating the movements and transforms of objects in the scene, which can be advantageously used for the camera pose estimation, especially for low bit rate applications.

6.5.4 Improving camera pose estimation results with RANSAC

Finally, we have tested the camera pose estimation using atom matching on a larger set of omnidirectional images with different translation vectors. We have used our database "Mede", described in Section 4.6. Besides multiple views obtained by camera rotation, the database contains also some images taken from camera positions that include both translation and rotation. However, since the ground truth values of the exact rotation are not available, we use these images only for the purpose of evaluating the robustness of translation estimation to different camera rotations. We have used the images only from the first set in the database, since the second set is similar. As mentioned in Section 4.6, the images were mapped to spherical images of resolution $2B = 256$.

We have used the same generating function for the dictionary design as in the previous experiments. However, due to the higher image resolution and smaller objects, the scales are uniformly distributed on a logarithmic scale, from 2 to $2B/16$ for Gaussian atoms and from 4 to $2B/2$ for edge-like atoms. We have excluded the biggest atoms from the dictionary as they are rare to



(a)

(b)

(c)

Figure 6.5: *Unwrapped Lab images (*128×128*): a)* y_3*; b)* y_1*; c)* y_2*.*

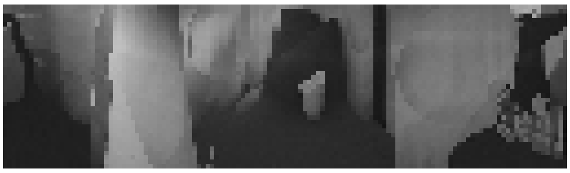
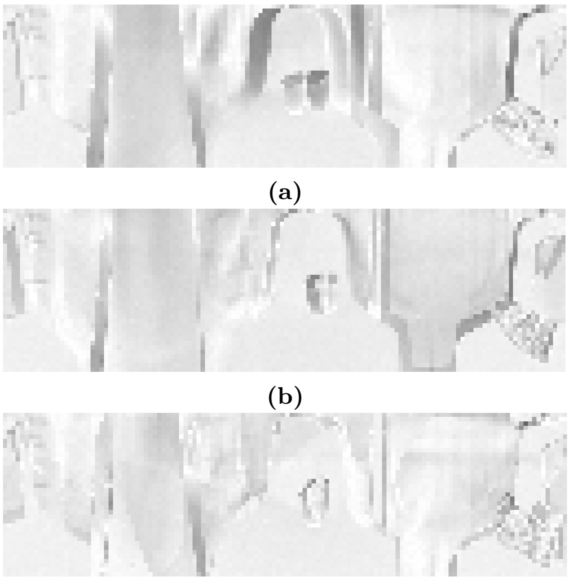


Figure 6.6: $\hat{y}_2$*: reconstructed* y_2 *using the disparity map from* y_3 *and* y_1*.*



(a)

(b)

(c)

Figure 6.7: *Differences between: a)* y_3 *and* y_2*, energy: 89.2; b)* y_1 *and* y_2*, energy: 77.0; c)* $\hat{y}_2$ *and* y_2*, energy: 49.7.*

appear, and they increase the computing time of MP. The rotation parameter ψ has been sampled from 0 to 2π with 256 different orientations. Images have been decomposed in 10 Gaussian atoms and 40 edge-like atoms. The atom matching has been performed using the shape similarity constraint ($s = 0.6$) and coefficient similarity. The correspondences have been sought in a neighborhood of $\pi/4$, as in previous experiments on the Lab images.

To see the influence of the local atom transforms on the pose estimation, we have performed the pose estimation in two cases: 1) when more point correspondences are used for each atom (as explained in Section 6.2), and 2) when only atom centers are used. In the first case, we use all points in the image where the selected atom has more than 90% of its highest value, i.e., when $w_{\gamma_i} = 0$ for $g_{\gamma_i}/\max(g_{\gamma_i}) < 0.9$. Those points basically lie close to the geometrical component represented by the atom and have probable epipolar matching. In the second correspondence set, we have used only atom centers as point correspondences, to investigate the correctness of the feature localization obtained by extracting atoms. Camera pose has been estimated for each of these two sets independently using the eight-point algorithm.

We have measured the Euclidean error of the estimated translation vector with respect to the ground truth. To illustrate how the value of the Euclidean error relates to the estimated and the target translation vectors, we give in Table 6.2 some random examples of target vectors, estimated vectors and errors between them. We can see that for the errors higher than 0.2 the direction of the camera movement cannot be determined.

Table 6.2: *The Euclidean error values for randomly chosen examples of the estimated and the target translation vectors. The table illustrates that for error values of* 0.2 *and above, the estimated translation vector cannot reliably point to the direction of the camera movement.*

Target $\mathbf{T}^\mathsf{T}$	Estimated $\mathbf{T}^\mathsf{T}$	Error
[1 0 0]	[0.999 0.025 0.036]	0.002
[1 0 0]	[0.990 0.081 0.115]	0.020
[1 0 0]	[0.980 0.114 0.162]	0.040
[1 0 0]	[0.900 0.251 0.355]	0.200
[1 0 0]	[0.800 0.346 0.489]	0.400
[0.707 0 0.707]	[0.706 0.053 0.706]	0.003
[0.707 0 0.707]	[0.702 0.145 0.697]	0.021
[0.707 0 0.707]	[0.682 0.263 0.682]	0.070
[0.707 0 0.707]	[0.617 0.432 0.657]	0.198

First two rows in Table 6.3 give the Euclidean error of the estimated translation vector for five different values of the target vector $\mathbf{T}$. We have used the camera pairs without rotation, i.e., when $\mathbf{R} = \mathbf{I}$. Error $e(\mathbf{T})$ refers to the error when the set with more points per atom has been used, while $e_c(\mathbf{T})$ is the error when only atoms centers have been used for the pose estimation. We can see that taking only atom centers gives better estimates of $\mathbf{T}$ in most of the cases. The fact that taking more points per atom to perform pose estimation can deteriorate the estimation performance might indicate that the shape of the atoms is not optimized for the task of correspondence matching. Learning the atoms that give good epipolar matching and image approximation performance is studied in the Chapter 7 of this thesis. The worst estimate of $\mathbf{T}$ has been found in the case of the target matrix $\mathbf{T} = [1 \;\; 1 \;\; 0]^\mathsf{T}$, since the camera movement along two axis represents a more difficult case for pose estimation.

Afterwards, RANSAC has been applied to refine the results. The total number of trials has been computed as in Eq. (6.4), with $p = 0.45$. The error tolerance has been set as the mean absolute error obtained when using eight-point method for all the points. The size of the acceptable consensus set t has been chosen to be 80% of the total number of correspondences. The errors of translation estimation after refinement with RANSAC are shown on the last two rows in Table 6.3, where $e_r(\mathbf{T})$ is the error for the correspondence set with more points per atoms, and $e_{cr}(\mathbf{T})$ is the error when only atom centers are used for estimation.

From Table 6.3 we see that applying RANSAC can sometimes lead to worse results than without RANSAC, especially in the case when more point correspondences are used per atom. This is probably due to the fact that these correspondences are erroneous (since the atom shape is not optimized) and hence make RANSAC converge to the wrong estimate. However, applying RANSAC on correspondence sets with only atom centers gives the best and the most reliable estimates of the direction of camera movement in most of the cases. It even gives a very good estimate of the target translation along both x and y axis.

Table 6.3: *Translation estimation errors for the "Mede" database, for pose estimation using atom matching and RANSAC.*

Target matrices	$\mathbf{T}=[100]^\top$ $\mathbf{R}=\mathbf{I}$	$\mathbf{T}=[-100]^\top$ $\mathbf{R}=\mathbf{I}$	$\mathbf{T}=[010]^\top$ $\mathbf{R}=\mathbf{I}$	$\mathbf{T}=[010]^\top$ $\mathbf{R}=\mathbf{I}$	$\mathbf{T}=[110]^\top$ $\mathbf{R}=\mathbf{I}$
$e(\mathbf{T})$	0.0827	0.0052	0.2756	0.2756	1.6808
$e_c(\mathbf{T})$	0.4479	0.0041	0.0936	0.0936	0.6334
$e_r(\mathbf{T})$	0.0461	1.6823	1.5078	1.9792	1.8631
$e_{cr}(\mathbf{T})$	0.1001	0.0151	0.0477	0.8273	0.0327

Finally, we study the influence of rotation between cameras on translation estimation. Given the images from two cameras with relative rotation $\mathbf{R}$ and translation $\mathbf{T}$, we show in Table 6.4 the estimation error only for the translation vector. Since we do not have the ground truth for $\mathbf{R}$, we do not report the estimates and error of rotation estimation. The goal of this experiment is to see if our translation estimation algorithm is robust to camera rotations. The errors are given in Table 6.4, showing similar performance as in the previous experiments with $\mathbf{R}=\mathbf{I}$. Namely, performing RANSAC on atom centers gives the best and the most reliable estimates, which are not affected by the camera rotation.

Table 6.4: *Translation estimation errors for the "Mede" database when images are rotated for an arbitrary angle between* 10 *and* 30 *degrees, for pose estimation using atom matching and RANSAC.*

Target matrices	$\mathbf{T}=[100]^\top$ $\mathbf{R}=\mathbf{R}_1$	$\mathbf{T}=[010]^\top$ $\mathbf{R}=\mathbf{R}_2$	$\mathbf{T}=[110]^\top$ $\mathbf{R}=\mathbf{R}_3$
$e(\mathbf{T})$	1.8188	0.1971	1.9783
$e_c(\mathbf{T})$	1.6054	0.0771	1.5955
$e_r(\mathbf{T})$	1.6073	0.2140	0.1143
$e_{cr}(\mathbf{T})$	0.0841	0.0572	0.0790

6.6 Conclusion

The contribution of this chapter is the application of the multi-view correlation model proposed in Chapter 4 to a problem of estimating the disparity map and camera pose from very coarse multi-view image descriptions. We have shown that the use of structured dictionaries for sparse decompositions enables pairing corresponding features among views and leads to good estimation of the disparity map and the camera pose. Additionally, we have included the RANSAC algorithm in camera pose estimation to improve its robustness to false correspondence matching. However, we have seen that the benefits of local transforms are not fully exploited in the camera pose estimation because the atoms are not optimized for epipolar matching. In the next chapter, we propose a dictionary learning algorithm that results in the dictionaries optimized for both approximation and epipolar matching. We will see in Chapter 7 that such learning can improve the pose estimation proposed in this chapter.

Chapter 7

Learning of Stereo Visual Dictionaries

7.1 Introduction

In previous chapters of this thesis we have proposed a multi-view image model based on sparse image approximations with overcomplete dictionaries of atoms. Each image is approximated by a linear combination of meaningful geometric features that represent the visual information of the scene. The multi-view model in Chapter 4 relates atoms across views with a local geometric transform that satisfies the multi-view geometry constraints. It leads to a compact multi-view image representation from which the 3D information is easily extractable, and which is additionally highly compressible. In Chapter 4 we have shown how the new model is beneficial for the compression of multi-view images in a distributed setting, while Chapter 6 demonstrated its application to camera pose estimation. Although in both applications the choice of discrete dictionary parameters was empirical, the method has shown good performance. Certainly, one would expect to obtain improved performance by optimizing the dictionary parameters for the particular case of multi-view image representation.

This chapter targets the problem of learning a dictionary adapted to the multi-view image representation model, where transforms between atoms satisfy the multi-view geometry constraints. Dictionary learning for overcomplete signal representations has become an extremely active area of research in the last few years. Researchers have realized that adapting the dictionary to a specific task or imposing a certain structure to the dictionary can yield significant performance improvement in target applications. Even though there has been recently a great amount of research done in the domain of dictionary learning for single images, there has been no work targeting the problem of learning stereo overcomplete dictionaries. A related problem of learning the receptive fields of binocular cells and the disparity tuning curves has been however investigated in the neuroscience research domain. Hoyer and Hyvärinen have applied independent component analysis (ICA) to learn the orthogonal basis of stereo images [Hoy00]. In their model, each stereo pair is a linear combination of stereo basis functions, which are composed of a left and a right component. Their algorithm resulted in Gabor-like basis functions tuned to different disparities. Okajima has proposed an Infomax learning approach, where the binocular receptive fields are learned by maximizing the mutual information between the stereo image model and the disparity [Oka04]. They have obtained results similar to Hoyer and Hyvärinen. The stereo dictionary learning method that we propose in this chapter differs from these state of the art methods for learning stereo basis functions in few aspects. First, we assume a sparse stereo image model and learn overcomplete dictionaries. Moreover, our method learns atoms from stereo image pairs while simultaneously performing the disparity estimation of the learned image features. The disparity estimation is included in the probabilistic model of stereo images, thus removing the need for disparity estimation as the preprocessing step. Finally, the stereo learning methods of Hoyer and Hyvärinen [Hoy00] and Okajima [Oka04] have been designed for the purpose of studying the receptive fields of binocular cells and the disparity tuning characteristics. However, the main target of the work presented in this chapter is to design stereo dictionaries that have the optimal properties for both image approximation and disparity or 3D scene structure estimation. To the best of our knowledge, such a problem has never been studied in the past.

We focus on the problem of two views and develop a maximum likelihood (ML) method for learning dictionaries that lead to improved image approximation under the sparsity prior. At the same time, the learned atoms give better multi-view geometry estimation from sparse image approximations. Our method builds upon the ML method for learning overcomplete dictionaries from natural monocular images, introduced by Olshausen and Field [Ols97, Ols96]. We propose an approach to stereo dictionary learning that includes the epipolar geometry in the probabilistic modeling, and hence matches the corresponding atoms within the learning process itself. The experimental results show that learning the scales of dictionary atoms for representing stereo omnidirectional images has significant benefits for the camera pose recovery and for the epipolar atom matching in distributed multi-view image representation. The novel stereo dictionary learning method has many other potential applications, such as the problem of modeling binocular cells in the primary visual cortex.

7.2 Preliminaries

7.2.1 Stereo image model

Developing the maximum likelihood dictionary learning method for stereo images requires first a definition of the stereo image model. We use our multi-view image model defined in Eq. (4.2), and consider two images: left image y_L and right image y_R, which have sparse representations in dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, respectively. In general, we consider those dictionaries as different, but they can also be equivalent in special cases. The images are approximated by sparse decompositions of m atoms up to an approximation error, i.e.:

$$\begin{aligned} y_L &= \mathbf{\Phi a} = \sum_{k=1}^{m} a_{l_k}\phi_{l_k} + e_L, \\ y_R &= \mathbf{\Psi b} = \sum_{k=1}^{m} b_{r_k}\psi_{r_k} + e_R, \end{aligned} \tag{7.1}$$

where $\mathcal{L} = \{l_k\}, \mathcal{R} = \{r_k\}, k = 1, ..., m$ label the sets of atoms that participate in the sparse decompositions of y_L and y_R, respectively. In other words, $\{l_k\}, \{r_k\}, k = 1, ..., m$ denote the atoms for which $a_{l_k} \neq 0$ and $b_{r_k} \neq 0$. Vectors $\mathbf{a}$ and $\mathbf{b}$ denote the vectors of sparse coefficients for the left and the right image, respectively. This model is a special case of the model in Eq. (4.2) when both stereo images are m-sparse, i.e., composed of m atoms. The motivation behind this model is that left and right images record the visual information from the same 3D environment and typically contain the image projections of the same 3D scene features. Hence, the number of sparse components will be approximately the same. We have also introduced a different notation in order to distinguish between the parameters of the left view and the parameters of the right view. As shown in Chapter 4, if the dictionary consists of localized and oriented atoms that represent well the edges, we can say that stereo images contain similar atoms that are locally transformed. Therefore, we further assume that signals y_L and y_R are correlated in the following way:

$$y_R = \sum_{k=1}^{m} b_{r_k}\psi_{r_k} + e_R = \sum_{k=1}^{m} b_{r_k} F_{l_k r_k}(\phi_{l_k}) + e_R, \tag{7.2}$$

where $F_{l_k r_k}(\cdot)$ denotes the transform of an atom ϕ_{l_k} in y_L to an atom ψ_{r_k} in y_R, and it differs for each $k = 1, ..., m$. This correlation model is a special case of the model given in Eq. (4.3) when the second summation in Eq. (4.3), representing the occlusions, is equal to zero. Since they do not participate in stereo matching, occlusions should not be considered for the learning of stereo dictionaries. Therefore, we assume in Eq. (7.2) that the occlusions in the scene are not dominant and that they can be included in the approximation errors e_R and e_L. As discussed in Chapter 4, object and atom transforms arising from the change of viewpoint can be usually captured using a structured parametric dictionary. We define our dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ as structured dictionaries built on the same generating function g, but using different sets of parameters: Γ_L for $\mathbf{\Phi}$, and Γ_R for $\mathbf{\Psi}$. To simplify the notation, we introduce the following equivalencies:

$$\phi_l \equiv g_{\gamma_l^{(L)}}, \qquad \gamma_l^{(L)} \in \Gamma_L, \quad \text{for } l = 1, ..., N, \tag{7.3}$$

$$\psi_r \equiv g_{\gamma_r^{(R)}}, \qquad \gamma_r^{(R)} \in \Gamma_R, \quad \text{for } r = 1, ..., N. \tag{7.4}$$

Using this notation, the transform F_{lr} of an atom ϕ_l in the left image to an atom ϕ_r in the right image is given as:

$$\begin{aligned} F_{lr}(\phi_l) &= F_{lr}(g_{\gamma_l^{(L)}}(\mathbf{v})) = U(\gamma')g_{\gamma_l^{(L)}}(\mathbf{v}) \\ &= \frac{1}{K} g_{\gamma_l^{(L)}}(Q_{lr}(\mathbf{v})) = g_{\gamma_r^{(R)}}(\mathbf{u}) = \phi_r, \end{aligned} \tag{7.5}$$

where $g_{\gamma_r^{(R)}} = \gamma' \circ g_{\gamma_l^{(L)}}$, as described in Section 4.3, and K is a normalization factor. Vector $\mathbf{v}$ denotes the coordinates of y_L and $\mathbf{u}$ denotes the coordinates obtained by transforming $\mathbf{v}$ with a linear transform Q_{lr}. As explained in Section 2.1.3, a linear transform Q is defined by atom parameters γ. Since the transform of one atom into another one in the structured dictionary translates into the transform of its parameters (see Section 4), the linear transform Q_{lr} is a function of $\gamma_l^{(L)}$ and $\gamma_r^{(R)}$.

The transforms $F_{l_k r_k}$, $k = 1, ..., m$, which relate the atoms in the left and the right view, satisfy the multi-view geometry constraint given by the epipolar constraint. In Section 4.5 we have defined the symmetric epipolar atom distance as a measure of epipolar matching between two atoms linked by a local transform. However, due to the computational complexity of evaluating this distance, we have chosen to use a simpler epipolar measure, presented in the next section.

7.2.2 Epipolar distance measure

The epipolar constraint gives a geometric relation between 3D points and their image projections, as explained in Chapter 2. Consider now that a point on the left image, given by the coordinates $\mathbf{v}$, and a point on the right image, $\mathbf{u}$, represent image projections of the same 3D point p from two camera positions with relative pose $(\mathbf{R}, \mathbf{T})$. $\mathbf{R} \in SO(3)$ is the relative orientation between cameras and $\mathbf{T} \in \mathbb{R}^3$ is their relative position. The epipolar geometry constraint is then:

$$\mathbf{u}^{\mathsf{T}}\hat{\mathbf{T}}\mathbf{R}\mathbf{v} = 0. \tag{7.6}$$

The matrix $\hat{\mathbf{T}}$ is obtained by representing the cross product of $\mathbf{T}$ with $\mathbf{R}\mathbf{x}_1$ as a matrix multiplication, as defined in Section 2.2.1. In the case when point $\mathbf{v}$ lies on the atom ϕ_l, as shown in Figure 7.1, our goal is to seek for a transform F_{lr} (and equivalently for Q_{lr}) such that $\mathbf{u} = Q_{lr}(\mathbf{v})$. In other words, point $\mathbf{u}$ should lie on the atom $\psi_r = F_{lr}(\phi_l)$. The epipolar geometry constraint is then:

$$[Q_{lr}(\mathbf{v})]^{\mathsf{T}}\hat{\mathbf{T}}\mathbf{R}\mathbf{v} = 0. \tag{7.7}$$

The epipolar constraint is rarely satisfied exactly, due to discrete spatial sampling of images, and can be only evaluated with a certain error ε_l. The estimated epipolar constraint d_{el} is thus given as:

$$d_{el} = [Q_{lr}(\mathbf{v})]^{\mathsf{T}}\hat{\mathbf{T}}\mathbf{R}\mathbf{v} + \varepsilon_l = d_L + \varepsilon_l. \tag{7.8}$$

Moreover, when there is uncertainty in epipolar geometry estimation, the epipolar measure is not symmetric. A different value is obtained when reversing the transform and considering the

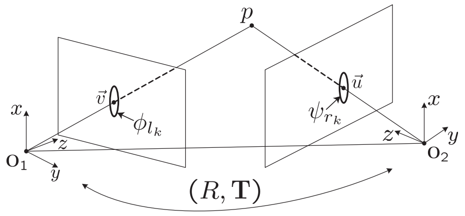


Figure 7.1: *Epipolar geometry between stereo atoms.*

epipolar constraint when point $\mathbf{u}$ on the second image is transformed into a point $\mathbf{v} = Q_{lr}^{-1}(\mathbf{u})$ on the first image. In that case, we have the epipolar geometry estimate d_{er}:

$$d_{er} = [Q_{lr}^{-1}(\mathbf{u})]^{\mathsf{T}} \hat{\mathbf{T}} \mathbf{R} \mathbf{u} + \varepsilon_r = d_R + \varepsilon_r. \tag{7.9}$$

where:

$$\begin{aligned} d_L &= [Q_{lr}(\mathbf{v})]^{\mathsf{T}} \hat{\mathbf{T}} \mathbf{R} \mathbf{v}, && (7.10) \\ d_R &= [Q_{lr}^{-1}(\mathbf{u})]^{\mathsf{T}} \hat{\mathbf{T}} \mathbf{R} \mathbf{u}. && (7.11) \end{aligned}$$

7.3 Maximum likelihood dictionary learning from natural images

In order to develop a stereo dictionary learning method, we will build upon the unsupervised learning of the overcomplete dictionary introduced by Olshausen and Field [Ols97, Ols96]. We first briefly overview their method, and then extend it to the dictionary learning for stereo images in the next section.

The work of Olshausen and Field was the earliest work addressing the problem of learning overcomplete dictionaries for image representation. They have developed a maximum likelihood (ML) dictionary learning method from natural images under the sparse coding assumption. The goal of their work was to give evidence that coding in the primary visual area V1 in the human cortex probably follows the sparse image representation model, using a dictionary of localized, oriented and bandpass atoms corresponding to the receptive fields of simple cells in V1. Their method is based on the maximum likelihood estimation of dictionary elements, given a sparse linear image model:

$$y = \mathbf{\Phi} \mathbf{a} = \sum_{k=1}^{M} a_k \phi_k, \tag{7.12}$$

where y denotes an image and $\mathbf{a}$ is a vector of coefficients. Columns of the matrix $\mathbf{\Phi}$ are atoms ϕ_k from a redundant dictionary $\mathcal{D} = \{\phi_k\}, k = 1, ..., M$. Note here that this generative image model is a linear combination of all atoms in the dictionary, and image representation will be sparse only if a small number of coefficients a_k are significant. The dictionary learning is performed by minimizing the Kullback-Leibler (KL) divergence between the probability distribution of natural images arising from the image model $y = \mathbf{\Phi} \mathbf{a}$ (given by the conditional probability $P(y|\mathbf{\Phi})$), and the actual distribution of natural images $P^*(y)$. This KL divergence is given as:

$$KL = \int_y P^*(y) log \frac{P^*(y)}{P(y|\mathbf{\Phi})} \mathrm{d}y. \tag{7.13}$$

Since $P^*(y)$ is constant, minimizing the KL divergence is equivalent to maximizing the log-likelihood $\log P(y|\mathbf{\Phi})$ that a set of natural images y arises from an overcomplete set of functions

$\mathbf{\Phi}$. Therefore, the goal of learning is to find the overcomplete dictionary $\mathbf{\Phi}^*$ such that:

$$\mathbf{\Phi}^* = \arg\max_{\mathbf{\Phi}} \langle \max_{\mathbf{a}} \log P(y|\mathbf{\Phi}) \rangle, \qquad \text{where } \; y = \mathbf{\Phi}\mathbf{a}. \tag{7.14}$$

Under the given sparse image model in Eq. (7.12), the probability $P(y|\mathbf{\Phi})$ can be evaluated by integrating $P(y, \mathbf{a}|\mathbf{\Phi})$ over all possible realizations of the coefficient vector $\mathbf{a}$, as:

$$P(y|\mathbf{\Phi}) = \int\limits_{\mathbf{a}} P(y|\mathbf{a}, \mathbf{\Phi}) P(\mathbf{a}) \mathrm{d}\mathbf{a}. \tag{7.15}$$

The probability that image y arises from a given realization of coefficients $\mathbf{a}$ and the dictionary $\mathbf{\Phi}$, denoted as $P(y|\mathbf{a}, \mathbf{\Phi})$, is simply modeled by the distribution of the noise that expresses the uncertainty of the imaging process. This noise η is considered to be additive and the image model becomes: $y = \mathbf{\Phi}\mathbf{a} + \eta$. Olshausen and Field have modeled $P(y|\mathbf{a}, \mathbf{\Phi})$ as white Gaussian noise with variance σ_I^2, obtaining:

$$P(y|\mathbf{a}, \mathbf{\Phi}) = \frac{1}{z_I} \exp\left(-\frac{\|y - \mathbf{\Phi}\mathbf{a}\|_2^2}{2\sigma_I^2} \right), \tag{7.16}$$

where $\| \cdot \|_2$ denotes the l_2 norm, and z_I is a normalization factor.

The hypothesis on sparse structure of the coefficient vector $\mathbf{a}$ is introduced by modeling the prior distribution on coefficients $P(\mathbf{a})$ with a distribution that is highly peaked at zero and heavy tailed. This encourages sparsity since only a small number of coefficients have values significantly different from zero. The distribution $P(\mathbf{a})$ is chosen to be factorial in a_i, i.e., $P(\mathbf{a}) = \prod_i P(a_i)$, thus assuming that the coefficients are independent of each other [Ols97]. The distribution of the coefficient magnitudes is chosen as:

$$P(a_i) = \frac{1}{z_\delta} \exp\left(-\delta S(a_i) \right), \tag{7.17}$$

where $S(a_i)$ defines the shape of the heavy tailed and zero peaked distribution, and δ controls its steepness. The factor z_δ is the normalization factor. Choosing the prior distribution of $\mathbf{a}$ to be tightly peaked at zero further permits us to approximate the integral in Eq. (7.15) only by its value at the maximum $P(y|\mathbf{a}, \mathbf{\Phi})P(\mathbf{a})$, thus modifying the optimization problem given in Eq. (7.14) into:

$$\mathbf{\Phi}^* = \arg\max_{\mathbf{\Phi}} \langle \max_{\mathbf{a}} \log P(y|\mathbf{a}, \mathbf{\Phi}) P(\mathbf{a}) \rangle. \tag{7.18}$$

This optimization problem can be also seen as an energy minimization problem:

$$\mathbf{\Phi}^* = \arg\min_{\mathbf{\Phi}} \langle \min_{\mathbf{a}} E(y, \mathbf{a}|\mathbf{\Phi}) \rangle, \tag{7.19}$$

where the energy function is:

$$E(y, \mathbf{a}|\mathbf{\Phi}) = -log P(y|\mathbf{a}, \mathbf{\Phi}) P(\mathbf{a}) = \|y - \mathbf{\Phi}\mathbf{a}\|^2 + \lambda \sum_i S(a_i). \tag{7.20}$$

and $\lambda = 2\sigma_N^2 \delta$.

The casted optimization problem can be solved by iterating between two steps. In the first step, $\mathbf{\Phi}$ is kept constant and the energy function is minimized with respect to the coefficient vector $\mathbf{a}$ by finding the equilibrium solution of the differential equation over $\mathbf{a}$ using a neural network. The second step keeps the obtained coefficients $\mathbf{a}$ constant, while performing the gradient descent on $\mathbf{\Phi}$ to minimize the energy $E(y, \mathbf{a}|\mathbf{\Phi})$. Therefore, the algorithm iterates between the sparse coding and the dictionary learning steps until convergence. Another way to observe this maximum likelihood learning is through the Expectation-Maximization (EM) algorithm, where images y are observed variables, $\mathbf{a}$ represent hidden or latent variables, and $\mathbf{\Phi}$ are parameters [Cul07]. Instead of maximizing directly $\log P(y|\mathbf{\Phi}) = \log \int P(y, \mathbf{a}|\mathbf{\Phi}) \mathrm{d}\mathbf{a}$, which is not computationally feasible, EM algorithm maximizes its lower bound:

$$\mathcal{L}(q, \mathbf{\Phi}) = \int\limits_{\mathbf{a}} q(\mathbf{a}|y) \log \frac{P(y, \mathbf{a}|\mathbf{\Phi})}{q(\mathbf{a}|y)} \mathrm{d}\mathbf{a}. \tag{7.21}$$

EM iterates between two steps: Expectation (E) and Maximization (M). In the E step of iteration t, $\mathcal{L}(q, \mathbf{\Phi}^{(t)})$ is maximized over the distribution $q(\mathbf{a}|y)$ when parameters $\mathbf{\Phi}$ are fixed. When the distribution $P(\mathbf{a}|y, \mathbf{\Phi}^{(t)})$ is taken as an estimate of $q^{(t+1)}(\mathbf{a}|y)$, $\mathcal{L}(q, \mathbf{\Phi})$ (and equivalently $\log P(y|\mathbf{\Phi})$) is maximized (see lemma 1 and lemma 2 in [Nea98]). Maximizing $\mathcal{L}(q, \mathbf{\Phi})$ over q involves, however, evaluating the expectation of $\log P(\mathbf{a}, y|\mathbf{\Phi}^{(t)})$ over q, which means integrating over $\mathbf{a}$. Nevertheless, if we use the same assumption that $P(\mathbf{a}, y|\mathbf{\Phi}^{(t)})$ can be approximated by its sample at the maximum, $\max_{\mathbf{a}} P(\mathbf{a}, y|\mathbf{\Phi}^{(t)})$, then we have [Cul07]:

$$\mathbf{a}^{(t+1)} = \arg\max_{\mathbf{a}} P(\mathbf{a}|y, \mathbf{\Phi}^{(t)}) = \arg\max_{\mathbf{a}} P(y|\mathbf{a}, \mathbf{\Phi}^{(t)}) P(\mathbf{a}) = \arg\min_{\mathbf{a}} E(y, \mathbf{a}|\mathbf{\Phi}^{(t)}). \tag{7.22}$$

In the M step, $\mathcal{L}(q, \mathbf{\Phi})$ is maximized with respect to $\mathbf{\Phi}$, using the $\mathbf{a}^{(t)}$ obtained in the E step. Again, taking $P(\mathbf{a}^{(t+1)}, y|\mathbf{\Phi}) \approx \max_{\mathbf{a}} P(\mathbf{a}^{(t+1)}, y|\mathbf{\Phi})$, we get:

$$\mathbf{\Phi}^{(t+1)} = \arg\max_{\mathbf{\Phi}} \mathcal{L}(q^{(t+1)}, \mathbf{\Phi}) \approx \arg\max_{\mathbf{\Phi}} \log P(y|\mathbf{a}^{(t+1)}, \mathbf{\Phi}) = \arg\min_{\mathbf{\Phi}} E(y, \mathbf{a}^{(t+1)}|\mathbf{\Phi}). \tag{7.23}$$

Therefore, the EM algorithm with the sampling of $P(\mathbf{a}|y, \mathbf{\Phi})$ at maximum point is equivalent to the iterative minimization of the energy function in the ML-based dictionary learning. The E step is essentially the inference (or sparse coding) step, while the M step corresponds to the learning step.

7.4 Maximum-likelihood learning of stereo dictionaries

7.4.1 Problem formulation

Following the similar approach as Olshausen and Field, we formulate the probabilistic framework for the maximum likelihood learning of overcomplete dictionaries $\mathbf{\Phi}, \mathbf{\Psi}$ that are used to represent stereo images y_L and y_R, respectively. We want to define the likelihood that stereo images captured by two cameras with a relative pose $(\mathbf{R}, \mathbf{T})$ arise from a set of atom pairs related by geometric transforms, under the sparsity prior. In other words, we want to learn the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ simultaneously. In order to infer what types of atoms are usually present in stereo (or multi-view) images, we need to maximize the probability that the observed stereo images y_L and y_R arise from dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ under a sparsity prior, and that the epipolar constraint between all corresponding points on y_L and y_R is equal to zero, i.e., $D = 0$. The epipolar geometry constraint is introduced in the probabilistic model in order to maximize the probability that the selected stereo pairs of atoms satisfy the disparity relation. Formally, we seek to solve the following optimization problem:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\max_{\mathbf{\Phi}, \mathbf{\Psi}} \langle \max_{\mathbf{a}, \mathbf{b}} \log P(y_L, y_R, D = 0|\mathbf{\Phi}, \mathbf{\Psi}) \rangle. \tag{7.24}$$

Marginalizing over $\mathbf{a}$ and $\mathbf{b}$ we have that:

$$P(y_L, y_R, D = 0|\mathbf{\Phi}, \mathbf{\Psi}) = \int\limits_{\mathbf{a}, \mathbf{b}} P(y_L, y_R, D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) \mathrm{d}\mathbf{a}\mathrm{d}\mathbf{b}. \tag{7.25}$$

7.4.2 Distribution of coefficient vectors a and b

We first need to define the joint distribution of coefficients $\mathbf{a}$ and $\mathbf{b}$, given dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, denoted as $P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})$. Let us assume that:

$$\forall k \ \ \mathit{and} \ \ \forall \mathbf{v} \ \ \mathit{s.t.} \ \ \phi_{l_k}(\mathbf{v}) \neq 0, \ \ \Rightarrow \ \ y_L(\mathbf{v}) = y_R(Q_{l_k r_k}(\mathbf{v})) = y_R(\mathbf{u}), \tag{7.26}$$

where $\mathbf{u} = Q_{l_k r_k}(\mathbf{v})$. This means that if the transform $Q_{l_k r_k}$ maps a pixel at position $\mathbf{v}$ on image y_L into a pixel at position $\mathbf{u}$ on image y_R, then those pixels have the same intensity. In other words, we assume that pixels keep their intensity values under the local transforms induced by the viewpoint change. This assumption holds in multi-view images when the scene is assumed to be Lambertian, and when atom transforms correctly represent the local object transforms. Under the assumption given in Eq. (7.26), we can use the Lemma A.1.1 to show that for an arbitrary stereo atom pair ϕ_l, ψ_r and their coefficients a_l, a_r we have the following equality:

$$\langle y_R, \psi_{r_k} \rangle = \frac{1}{\sqrt{J_{l_k r_k}}} \langle y_L, \phi_{l_k} \rangle, \tag{7.27}$$

where $J_{l_k r_k} = |\frac{\partial \mathbf{u}}{\partial \mathbf{v}}| = |\frac{\partial Q_{l_k r_k}(\mathbf{v})}{\partial \mathbf{v}}|$ is the Jacobian determinant, or just simply called the Jacobian, of the linear transform $Q_{l_k r_k}$. Using the sparse image model and Lemma A.1.1 we can obtain the following probabilities:

$$P(b_r|a_l, \phi_l, \psi_r) = P(a_l|b_r, \phi_l, \psi_r) = \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 \right), \tag{7.28}$$

where z_b is the normalization factor and σ_b is the standard deviation of the zero-mean Gaussian noise that models the difference between b_r and $a_l/\sqrt{J_{lr}}$. The detailed derivation of Eq. (7.28) is given in Appendix A.2.

The joint distribution of coefficients a_l, b_r given atoms ϕ_l, ψ_r can be decomposed in two ways:

$$\begin{aligned} P(a_l, b_r|\phi_l, \psi_r) &= P(b_r|a_l, \phi_l, \psi_r)P(a_l), && (7.29) \\ P(a_l, b_r|\phi_l, \psi_r) &= P(a_l|b_r, \phi_l, \psi_r)P(b_r), && (7.30) \end{aligned}$$

where we assume that priors on coefficients in each image $P(a_l)$ and $P(b_r)$ are independent of the atoms. Although in reality the distribution of the coefficients would depend on an arbitrarily chosen dictionary, imposing the independence of the coefficients with respect to the dictionary during learning would actually lead to inferring a dictionary that gives the same prior distribution of coefficients for all types of images. In other words, when the prior of coefficients is tightly picked at zero, the learning would lead to a universal dictionary in which all natural images have sparse decompositions.

For modeling the priors on coefficients, we will not take the same approach as Olshausen and Field, who have modeled this distribution as continuous and peaked at zero. Instead, we assume that the coefficients a_l and b_r are drawn from a Bernoulli distribution over the activity of coefficients $\mathcal{I}(a_l)$ and $\mathcal{I}(b_r)$, where $\mathcal{I}$ denotes the indicator function. These distributions are:

$$P(a_l) = \begin{cases} p & \text{if } \mathcal{I}(a_l) = 1; \\ q & \text{if } \mathcal{I}(a_l) = 0. \end{cases}$$

$$P(b_r) = \begin{cases} p & \text{if } \mathcal{I}(b_r) = 1; \\ q & \text{if } \mathcal{I}(b_r) = 0. \end{cases}$$

Choosing $p \ll q$ will introduce the sparsity assumption on the coefficients, saying that it is much more probable that the coefficient takes the value zero, than a value greater than zero. Let M we denote the cardinality of overcomplete dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$. When images y_L and y_R are m sparse, we obtain:

$$P(\mathbf{a}) = \prod_{l=1}^{M} P(a_l) = p^m (1-p)^{(M-m)}, \tag{7.31}$$

$$P(\mathbf{b}) = \prod_{r=1}^{M} P(b_r) = p^m (1-p)^{(M-m)}. \tag{7.32}$$

Let us now give a parametric form for p as $1/(1+e^{1/\lambda})$, where reducing the value of λ will increase the level of "sparseness" of coefficients. The coefficients a_l and b_r for $l, r = 1, .., M$ are i.i.d., and we have then:

$$P(\mathbf{a}) = (\frac{e^{1/\lambda}}{1+e^{1/\lambda}})^M \exp(-m/\lambda) = \frac{1}{z_\lambda} \exp\left(-\frac{\|\mathbf{a}\|_0}{\lambda}\right), \tag{7.33}$$

$$P(\mathbf{b}) = (\frac{e^{1/\lambda}}{1+e^{1/\lambda}})^M \exp(-m/\lambda) = \frac{1}{z_\lambda} \exp\left(-\frac{\|\mathbf{b}\|_0}{\lambda}\right), \tag{7.34}$$

where $\|\cdot\|_0$ denotes the l_0 norm. By multiplying Eq. (7.29) and Eq. (7.30) and taking the square root we get:

$$P(a_l, b_r|\phi_l, \psi_r) = \sqrt{P(b_r|a_l, \phi_l, \psi_r)P(a_l|b_r, \phi_l, \psi_r)P(a_l)P(b_r)}. \tag{7.35}$$

We further assume that pairs of coefficients (a_l, b_r) are independent, which is usually the case when image decompositions are sparse enough. Then, the distribution $P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})$ is factorial, i.e.:

$$P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) = \prod_{l=1}^{M}\prod_{r=1}^{M} P(a_l, b_r|\phi_l, \psi_r) = \prod_{l=1}^{M}\prod_{r=1}^{M} \sqrt{P(b_r|a_l, \phi_l, \psi_r)P(a_l|b_r, \phi_l, \psi_r)P(a_l)P(b_r)}. \tag{7.36}$$

Finally, we obtain from Eq. (7.28) and Eq. (7.33):

$$P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) = \frac{1}{z_b z_\lambda} \exp\left(-\frac{1}{2\sigma_b^2}\sum_{l=1}^{M}\sum_{r=1}^{M}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2\right) \exp\left(-\frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0)\right). \tag{7.37}$$

Having evaluated the probability $P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})$, we can now go back to the likelihood in Eq. (7.25). We can approximate the probability $P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})$ by its value at the maximum, since it is a product of a zero-mean Gaussian distribution and a discrete distribution tightly peaked at zero. Eq. (7.25) then becomes:

$$\begin{aligned} P(y_L, y_R, D=0|\mathbf{\Phi}, \mathbf{\Psi}) &\approx P(y_L, y_R, D=0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) && (7.38)\\ &= P(y_L, y_R|D=0, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})P(D=0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) && (7.39)\\ &= P(y_L, y_R|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})P(D=0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}), && (7.40) \end{aligned}$$

where Eq. (7.39) follows from the chain rule, and Eq. (7.40) holds since $D = 0$ does not bring more information to y_L, y_R than $\mathbf{\Phi}, \mathbf{\Psi}$. To evaluate our likelihood function, we next need to find the probability that the epipolar constraint D is equal to zero given the stereo image model in Eq. (7.2), i.e., we need to find $P(D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$.

7.4.3 Probability of epipolar matching of atoms

The epipolar matching of two points on the left image and the right image, which are the projections of the same point in the 3D space, is derived in Section 7.2.2 for the considered stereo image model. The estimation of the epipolar constraints d_{el} and d_{er} is erroneous, and we assume that the estimation errors ε_l and ε_r are i.i.d. white zero-mean Gaussian noises of variances σ_{dl}^2 and σ_{dr}^2, respectively:

$$\begin{aligned} P(\varepsilon_l) &= \frac{1}{z_{dl}} \exp\left(-\frac{\varepsilon_l^2}{2\sigma_{dl}^2}\right), && (7.41)\\ P(\varepsilon_r) &= \frac{1}{z_{dr}} \exp\left(-\frac{\varepsilon_r^2}{2\sigma_{dr}^2}\right), && (7.42) \end{aligned}$$

where z_{dl} and z_{dr} are the normalization factors. Equivalently, we can define the conditional probability of the random variables d_{el} and d_{er}, given a pair of points $\mathbf{v}$, $\mathbf{u}$ and atoms ϕ_l, ψ_r as:

$$P(d_{el}|\mathbf{v}, \mathbf{u}, \phi_l, \psi_r) = \frac{1}{z_{dl}} \exp\left(-\frac{(d_{el} - d_L)^2}{2\sigma_{dl}^2}\right), \tag{7.43}$$

$$P(d_{er}|\mathbf{v}, \mathbf{u}, \phi_l, \psi_r) = \frac{1}{z_{dr}} \exp\left(-\frac{(d_{er} - d_R)^2}{2\sigma_{dr}^2}\right). \tag{7.44}$$

Although atoms ϕ_l and ψ_r do not appear explicitly in functions d_L and d_R, they are implicitly there since the transform Q_{lr} depends on the parameters $\gamma_l^{(L)}$ an $\gamma_r^{(R)}$, which respectively define atoms $\phi_l = g_{\gamma_l^{(L)}}$ and $\psi_r = g_{\gamma_r^{(R)}}$. Since we want to find the probability that two atoms satisfy the epipolar constraint, we are interested in the probability of a particular realization of random variables d_{el} and d_{er} when they are simultaneously equal to zero. Therefore, we define the conditional probability of $d_L = 0, d_R = 0$, given $\mathbf{v}, \mathbf{u}, \phi_l, \psi_r$ as:

$$\begin{aligned} P(d_{el} = 0, d_{er} = 0|\mathbf{v}, \mathbf{u}, \phi_l, \psi_r) &= P(d_{el} = 0|\mathbf{v}, \mathbf{u}, \phi_l, \psi_r)P(d_{er} = 0|\mathbf{v}, \mathbf{u}, \phi_l, \psi_r) \\ &= \frac{1}{z_{dl} z_{dr}} \exp\left(-\frac{d_L^2}{2\sigma_{dl}^2}\right) \exp\left(-\frac{d_R^2}{2\sigma_{dr}^2}\right). \end{aligned} \tag{7.45}$$

Further on, we define the conditional probability that $d_{el} = 0$ and $d_{er} = 0$ for all pairs of pixels that are given by a transform between two corresponding atoms, i.e., for all $\mathbf{v}_i, \mathbf{u}_i$, $i = 1, ..., q$. Let D_{lr} denote the event when $d_{el}^{(i)} = 0$ and $d_{er}^{(i)} = 0$ for all $i = 1, ..., q$, where $d_{el}^{(i)} = d_L^{(i)} + \varepsilon_l = [Q_{lr}(\mathbf{v}_i)]^{\mathsf{T}} \hat{\mathbf{T}} \mathbf{R} \mathbf{v}_i + \varepsilon_l$ and $d_{er}^{(i)} = d_R^{(i)} + \varepsilon_l = [Q_{lr}^{-1}(\mathbf{u}_i)]^{\mathsf{T}} \hat{\mathbf{T}} \mathbf{R} \mathbf{u}_i + \varepsilon_r$. If we assume that estimation of the epipolar distance for each pixel pair i is done independently of the others, we have that:

$$\begin{aligned} P(D_{lr} = 0|\phi_l, \psi_r) &= \prod_{i=1}^{q} P(d_{el}^{(i)} = 0, d_{er}^{(i)} = 0|\mathbf{v}_i, \mathbf{u}_i, \phi_l, \psi_r) & (7.46) \\ &= \prod_{i=1}^{q} \frac{1}{z_{dl}^{(i)} z_{dr}^{(i)}} \exp\left(-\frac{(d_L^{(i)})^2}{2(\sigma_{dl}^{(i)})^2}\right) \exp\left(-\frac{(d_R^{(i)})^2}{2(\sigma_{dr}^{(i)})^2}\right) & (7.47) \\ &= \frac{1}{z_D} \exp\left(-\frac{1}{2\sigma_D^2} \sum_{i=1}^{q} \left(w_l^{(i)} (d_L^{(i)})^2 + w_r^{(i)} (d_R^{(i)})^2\right)\right) & (7.48) \\ &= \frac{1}{z_D} \exp\left(-\frac{1}{2\sigma_D^2} D_E(\gamma_l^{(L)}, \gamma_r^{(R)})\right), & (7.49) \end{aligned}$$

where $z_D = \prod_{i=1}^{q} z_{dl}^{(i)} z_{dr}^{(i)}$, $w_l^{(i)} = \sigma_D^2/(\sigma_{dl}^{(i)})^2$ and $w_r^{(i)} = \sigma_D^2/(\sigma_{dr}^{(i)})^2$. Introducing the weights $w_l^{(i)}, w_r^{(i)}$ permits us to put more importance on the epipolar constraint for points that are closer to the geometric discontinuity represented by the atom, where estimation of the epipolar constraint is more reliable. For the ease of notation, we have replaced the summation over i in Eq. (7.48) with a function D_E that depends on the parameters $\gamma_l^{(L)}, \gamma_r^{(R)}$ of the stereo atom pair. Finally, the probability of epipolar matching for the stereo image pair is the product of probabilities of epipolar matching for pairs of active atoms. The active atoms participate in sparse decompositions of the left and the right image with their respective coefficients a_l and b_r, which are different from zero. Then, we can model the probability $P(D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$ as:

$$P(D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) = \frac{1}{z_D} \exp\left(-\frac{1}{2\sigma_D^2} \sum_{l=1}^{M} \sum_{r=1}^{M} \mathcal{I}(a_l)\mathcal{I}(b_r) D_E(\gamma_l^{(L)}, \gamma_r^{(R)})\right). \tag{7.50}$$

7.4.4 The energy function

At this point, we have defined all components of the objective maximum likelihood function in Eq. (7.39), except $P(y_L, y_R|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$. This probability can be modeled by a Gaussian white noise:

$$P(y_L, y_R|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) = P(e_L + e_R) = \frac{1}{z_I} \exp\left(-\frac{1}{2\sigma_I^2}(\|y_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|y_R - \mathbf{\Psi}\mathbf{b}\|_2^2)\right), \tag{7.51}$$

where we have used the fact that the sum of two zero-mean Gaussian random variables is also a zero-mean Gaussian random variable. We can now rewrite the maximum likelihood transform learning problem in Eq. (7.52) as:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\max_{\mathbf{\Phi},\mathbf{\Psi}} \left[\max_{\mathbf{a},\mathbf{b}} (\log P(y_L, y_R|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(D = 0|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}))\right]. \tag{7.52}$$

This is equivalent to solving the following minimization problem:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\min_{\mathbf{\Phi},\mathbf{\Psi}} \langle \min_{\mathbf{a},\mathbf{b}} E(y_L, y_R, D = 0, \mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})\rangle, \tag{7.53}$$

where E denotes the energy function given as:

$$E(y_L, y_R, D = 0, \mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) = \frac{1}{2\sigma_I^2}(\|y_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|y_R - \mathbf{\Psi}\mathbf{b}\|_2^2) \tag{7.54}$$

$$+ \frac{1}{2\sigma_D^2} \sum_{l=1}^{M} \sum_{r=1}^{M} \mathcal{I}(a_l)\mathcal{I}(b_r) D_E(\gamma_l^{(L)}, \gamma_r^{(R)}) \tag{7.55}$$

$$+ \frac{1}{2\sigma_b^2} \sum_{l=1}^{M} \sum_{r=1}^{M} (b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \tag{7.56}$$

The energy function thus consists of four main summation terms:

1. the data fidelity term, expressed by the energy of the approximation error after sparse approximation of images y_L and y_R,
2. the epipolar constraint term, measuring the epipolar matching of atoms in sparse decompositions of a stereo image pair,
3. the coefficient similarity term, measuring the correlation of coefficients of stereo atom pairs under a local transform,
4. the sparsity term, expressing the degree of sparsity of a stereo image pair.

We can see that the first three terms depend on the choice of dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, while the last term depends only on the coefficient vectors $\mathbf{a}$ and $\mathbf{b}$. Therefore, we group the first three terms into a function $f(y_L, y_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$ and express the energy function as:

$$E(y_L, y_R, D = 0, \mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi}) = f(y_L, y_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \tag{7.57}$$

Our target optimization problem in Eq. (7.53) thus becomes the following energy minimization:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\min_{\mathbf{\Phi},\mathbf{\Psi}} \left[\min_{\mathbf{a},\mathbf{b}} \left(f(y_L, y_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0)\right)\right]. \tag{7.58}$$

7.5 Energy minimization by Expectation-Maximization

As explained in Section 7.3, we can solve the optimization problem in Eq. (7.58) in two iterative steps: 1) minimization of the energy function with respect to coefficients, while keeping the atoms fixed (inference); and 2) minimization of the energy with respect to atoms, when the coefficients are taken from the previous step (learning).

7.5.1 Minimization with respect to coefficients

In the inference step, we need to find the coefficients **a** and **b** as:

$$(\mathbf{a}, \mathbf{b})^* = \arg\min_{\mathbf{a},\mathbf{b}} f(y_L, y_R, \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \tag{7.59}$$

We can see that this problem is similar to the constrained sparse approximation problem in Eq. (2.4) when casted as an unconstrained problem, where $\frac{1}{2\lambda}$ is a trade-off parameter between minimizing the energy in f and the sparsity of coefficient vectors **a** and **b**. Since finding the global minimum of such a problem is NP-hard [Nat95], we will use a greedy approach to find a locally optimal solution. Although they are not guaranteed to find the sparsest solution for the problem, greedy algorithms have performed quite well in practice, giving high residue energy decay rate [Mal93, Fig06, Fro00]. The particular advantage of using a greedy approach here is that it leads to a signal approximation with a small set of coefficient different from zero. This is necessary in our multi-view framework due to the component of the energy function related to the epipolar constraint, which includes the indicator function of coefficients. Sparse approximation methods that minimize the l_1 norm can also be applied using a different prior on the coefficients. However, the computational complexity would be much higher because these methods result in many coefficients that are close to zero, but not exactly zero. Other possibility that can be envisaged in the future is to implement the sparse decomposition step using hard thresholding in the locally competitive algorithm, which thresholds insignificant coefficients to zero [Roz08].

We propose a greedy algorithm that chooses at each iteration k the pair of atoms ϕ_{l_k}, ψ_{r_k} that give the minimal value of the function:

$$\begin{aligned}(\phi_{l_k}, \psi_{r_k}) &= \arg\min_{\phi_l,\psi_r}[\frac{1}{2\sigma_I^2}(\|h_l^{(k-1)} - \langle h_l^{(k-1)}, \phi_l\rangle\phi_l\|_2^2 + \|h_r^{(k-1)} - \langle h_r^{(k-1)}, \psi_r\rangle\psi_r\|_2^2) \\ &+ \frac{1}{2\sigma_D^2} D_E(\gamma_l^{(L)}, \gamma_r^{(R)}) + \frac{1}{2\sigma_b^2}(\langle h_r^{(k-1)}, \psi_r\rangle - \frac{\langle h_l^{(k-1)}, \phi_l\rangle}{J_{lr}})^2],\end{aligned} \tag{7.60}$$

where $h_l^{(k-1)}$ and $h_r^{(k-1)}$ are the residues of the left and right images respectively, after $k-1$ iterations. At the beginning, the residues are: $h_L^{(0)} = y_L$ and $h_R^{(0)} = y_R$ and they are updated at each step k as:

$$\begin{aligned}h_L^{(k)} &= h_L^{(k-1)} - \langle h_L^{(k-1)}, \phi_{l_k}\rangle\phi_{l_k}, \\ h_R^{(k)} &= h_R^{(k-1)} - \langle h_R^{(k-1)}, \psi_{r_k}\rangle\psi_{r_k}.\end{aligned} \tag{7.61}$$

The coefficients a_{l_k} and b_{r_k} are simply evaluated as:

$$\begin{aligned}a_{l_k} &= \langle h_L^{(k-1)}, \phi_{l_k}\rangle, \\ b_{r_k} &= \langle h_R^{(k-1)}, \psi_{r_k}\rangle.\end{aligned} \tag{7.62}$$

We will refer to this algorithm as Multi-view Matching Pursuit (MVMP)[1]. It is easy to see that the proposed greedy algorithm is essentially the Weak Matching Pursuit (WMP) [Tem00]. WMP chooses at each iteration an atom such that its coefficient differs from the maximal coefficient up to a scale factor $\kappa \in (0, 1]$. Therefore, the chosen coefficients satisfy:

$$\begin{aligned}a_{l_k} &\geqslant \kappa_l \max_{\gamma_l^{(L)}}\langle h_l^{(k-1)}, \phi_{l_k}\rangle, \\ b_{r_k} &\geqslant \kappa_r \max_{\gamma_r^{(R)}}\langle h_r^{(k-1)}, \psi_{r_k}\rangle,\end{aligned} \tag{7.63}$$

---

[1] Although we take here only two images, we can generalize the algorithm to more than two images by pairwise image correspondence.

where κ_l and κ_r depend on the last two summation terms in Eq. (7.60). The bound of the approximation rate of MVMP is given in Appendix A.3.

Since we use parametric dictionaries where $\phi_l = g_{\gamma_l^{(L)}}$ and $\psi_r = g_{\gamma_r^{(R)}}$, Eq. (7.60) can be rewritten as:

$$\begin{aligned}(\gamma_{l_k}^{(L)},\gamma_{r_k}^{(R)}) = \arg\min_{\gamma_l^{(L)},\gamma_r^{(R)}} & [\frac{1}{2\sigma_I^2}(\|h_l^{(k-1)} - \langle h_l^{(k-1)}, g_{\gamma_l^{(L)}}\rangle g_{\gamma_l^{(L)}}\|_2^2 + \|h_r^{(k-1)} - \langle h_r^{(k-1)}, g_{\gamma_r^{(R)}}\rangle g_{\gamma_r^{(R)}}\|_2^2) \\ & + \frac{1}{2\sigma_D^2} D_E(\gamma_l^{(L)},\gamma_r^{(R)}) + \frac{1}{2\sigma_b^2}(\langle h_r^{(k-1)}, g_{\gamma_r^{(R)}}\rangle - \frac{\langle h_l^{(k-1)}, g_{\gamma_l^{(L)}}\rangle}{J_{lr}})^2]. \qquad (7.64)\end{aligned}$$

7.5.2 Minimization with respect to atom scale parameters

In the previous step, the parameters $\gamma_{l_k}^{(L)},\gamma_{r_k}^{(R)}, k = 1,...,m$ and the atoms $\phi_{l_k}, \psi_{r_k}, k = 1,...,m$ that participate in the decomposition of images y_L and y_R have been found by MVMP. Once their coefficients have been evaluated, they are kept fixed while the atoms are updated by minimizing the energy function. Knowing the selected atoms, the energy function at iteration t of EM becomes:

$$\begin{aligned}E^{(t)} &= \frac{1}{2\sigma_I^2}(\|y_L - \sum_{k=1}^{m} a_{l_k} g_{\gamma_{l_k}^{(L)}}\|_2^2 + \|y_R - \sum_{k=1}^{m} b_{r_k} g_{\gamma_{r_k}^{(R)}}\|_2^2) + \frac{1}{2\sigma_D^2}\sum_{k=1}^{m} D_E(\gamma_{l_k}^{(L)},\gamma_{r_k}^{(R)}) \\ &+ \frac{1}{2\sigma_b^2}\sum_{k=1}^{m}(b_{r_k} - \frac{a_{l_k}}{\sqrt{J_{l_k r_k}}})^2 + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \qquad (7.65)\end{aligned}$$

The energy function $E^{(t)}$ is an analytic function of parameters $\{\gamma_{l_k}^{(L)}\}$ and $\{\gamma_{r_k}^{(R)}\}$ and we can calculate its derivatives with respect to each parameter. Therefore, one can use the multivariate gradient descent or the multivariate conjugate gradient method to find the local minimum of $E^{(t)}$ with respect to $\gamma_{l_k}^{(L)}$ and $\gamma_{r_k}^{(R)}$, given the coefficients $\mathbf{a}$ and $\mathbf{b}$.

The sparse coding and the learning steps are iteratively repeated until the convergence is achieved. The inference should be performed from a large set of data, i.e., from different multi-view pairs and with different camera poses. This is achieved by performing the sparse coding step on a randomly selected set of pairs yielding sparse coefficients for each pair, and then by averaging the energy function over all pairs to perform the learning step.

7.6 Experimental results

The correlation model for multi-view omnidirectional images described in Section 7.2 assumes that atoms in sparse decompositions of different views are correlated with a geometric transform. In practical applications, the benefits of this correlation model depend on the discretization of the parameters that are used for dictionary construction: translations, rotations and scaling. Among those, the scaling parameters are the most important since they directly define the shape of atoms (they become elongated as the scales become anisotropic). On the other hand, translations and rotations depend highly on the distance and orientation of cameras, so learning these parameters is meaningful only when camera positions are fixed. Therefore, we choose in this work to learn the scaling parameters of atoms in overcomplete parametric dictionaries optimized for stereo image representation. We want to perform learning of atom scales that are present in sparse approximations of stereo views, which additionally satisfy the epipolar geometry constraint.

7.6.1 Experimental setup

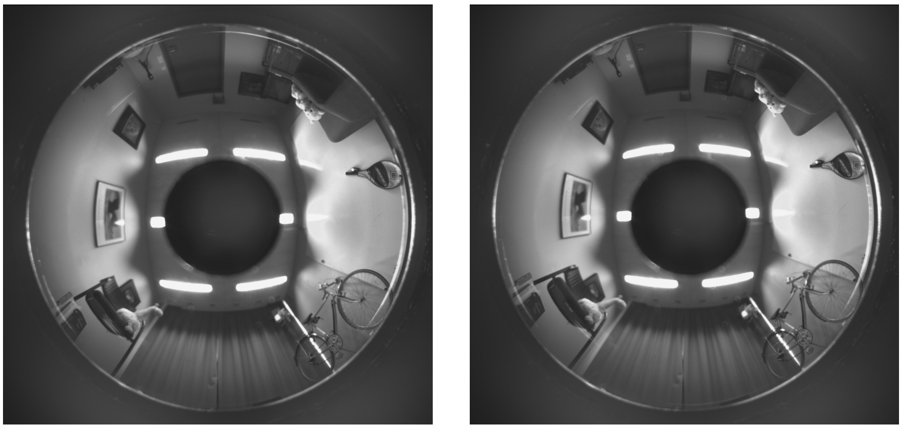
The stereo image model given in Eq. (7.2) does not put any assumption on the type of cameras used for stereo image acquisition. It can be applied to planar or omnidirectional multi-view images by defining the dictionary for the considered type of images, and by introducing the epipolar geometry constraints that are defined for that particular image projection geometry. Since this thesis proposes the use of omnidirectional cameras for 3D scene representation with multi-view images, we perform learning of stereo dictionaries for omnidirectional images. As explained in Section 2.3.1, omnidirectional images obtained by catadioptric cameras can be appropriately mapped to spherical images. For representing spherical images, we use the formulation of a dictionary on the 2-D unit sphere proposed in Chapter 3. The epipolar geometry for the spherical camera model is formulated in the same manner as for the perspective camera model, as explained in detail in Section 2.4.1. The transform $\mathbf{u} = Q_{lr}(\mathbf{v})$ that relates a point $\mathbf{v}$ on atom $\phi_l = g_{\gamma_l^{(L)}}$ to its corresponding transformed point $\mathbf{u}$ on the transformed atom $\psi_r = g_{\gamma_r^{(R)}}$ has been defined in Section 4.5. It is derived from the local transforms between geometric atoms via a linear transform of the coordinate system. When the atoms ϕ_l and ψ_r are defined by the parameters $\gamma_l^{(L)} = (\tau_l^{(L)}, \nu_l^{(L)}, \psi_l^{(L)}, \alpha_l^{(L)}, \beta_l^{(L)})$ and $\gamma_r^{(R)} = (\tau_r^{(R)}, \nu_r^{(R)}, \psi_r^{(R)}, \alpha_r^{(R)}, \beta_r^{(R)})$, respectively, the point $\mathbf{u}$ can be written as:

$$\mathbf{u} = \mathbf{R}^{-1}_{\gamma_r^{(R)}} \cdot \zeta(\mathbf{R}_{\gamma_l^{(L)}} \cdot \mathbf{v}). \tag{7.66}$$

Matrices $\mathbf{R}_{\gamma_l^{(L)}}$ and $\mathbf{R}_{\gamma_r^{(R)}}$ are rotation matrices given respectively by Euler angles $(\tau_l^{(L)}, \nu_l^{(L)}, \psi_l^{(L)})$ and $(\tau_r^{(R)}, \nu_r^{(R)}, \psi_r^{(R)})$. The grid transform $\zeta(\cdot)$ due to anisotropic scaling is given in Eq. (4.14).

Since we are interested in learning only scales $\alpha^{(L)}, \beta^{(L)}$ from the set of atom parameters $\gamma^{(L)}$, and $\alpha^{(R)}, \beta^{(R)}$ from $\gamma^{(R)}$, the energy function is minimized only with respect to these four parameters. We have performed the minimization using the conjugate gradient method[2]. Motion parameters (τ, ν) are constant, and include the positions of all pixels in an image. Rotations ψ are taken from 0 to π in uniform steps, with resolution N_r. We have taken the same motion and rotation parameters for the left and the right dictionary.

We have tested the proposed stereo dictionary learning algorithm on our "Mede" omnidirectional multi-view database, described in Section 6.5.4. The database consists of 54 omnidirectional images of an indoor environment, grouped into two sets: set without plants (27 images), and set with plants (27 images). The role of plants in the second set is to introduce some image statistics of an outdoor environment, since capturing real outdoor images was not performed due to the specific hardware requirements. Two views from each of the sets are shown in Figure 7.2 for the set without plants and in Figure 7.3 for the set with plants. We have formed 216 pairs of images with different distances between cameras, and hence different translation vectors $\mathbf{T}$ on the $x - y$ plane (pairs were formed within each set). Since we know the camera positions and the rotation is identity, $\mathbf{T}$ and $\mathbf{R}$ are known for each image pair.

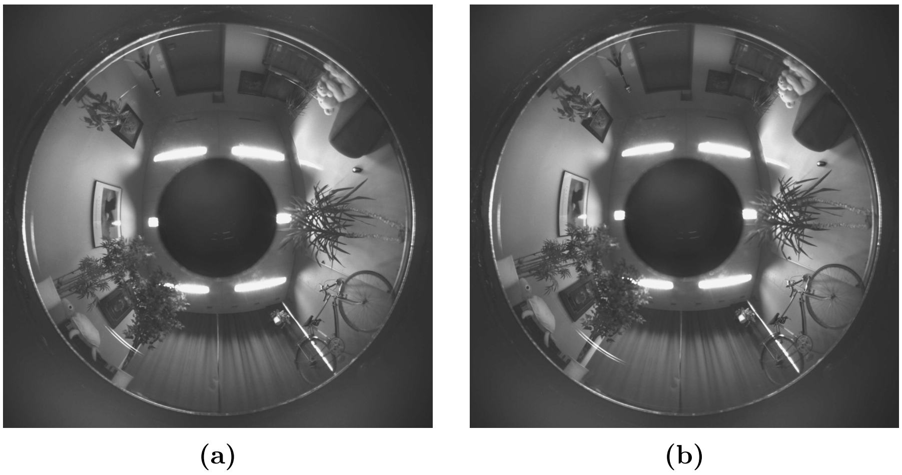


(a) **(b)**

Figure 7.2: *Two views from the "Mede" database, no plants.*

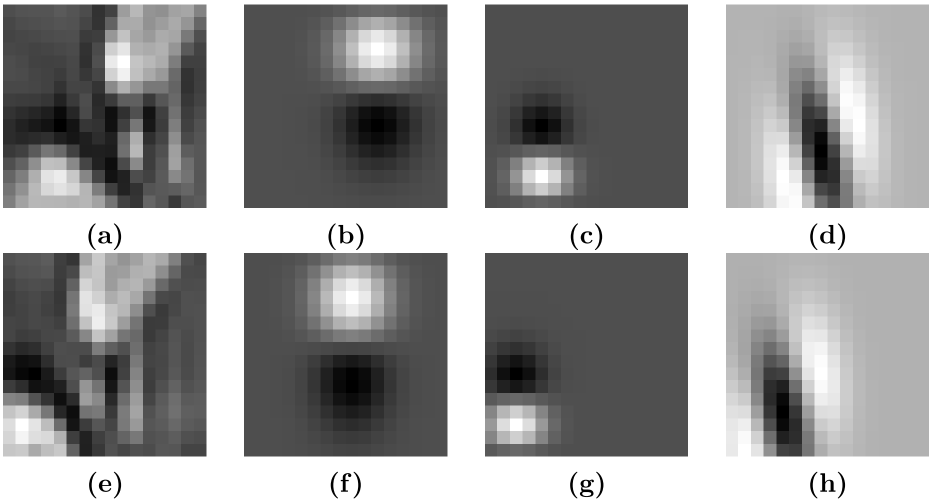
---

[2] http://www.kyb.tuebingen.mpg.de/bs/people/carl/code/minimize/



(a) **(b)**

Figure 7.3: *Two views from the "Mede" database, with plants.*



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

(e) **(f)** **(g)** **(h)**

Figure 7.4: *Example of a pair of stereo patches and their MVMP selected atoms: a) left patch, b)-d) the first three atoms in the MVMP decomposition of the left patch; e) right patch, f)-h) the first three atoms in the MVMP decomposition of the right patch.*

The first step in the learning algorithm (i.e., the expectation (E) step implemented by MVMP) needs to be performed on a big set of statistically different stereo images in order to result in a meaningful learning. To limit the complexity of the whole learning process, but still include the image diversity, we select small patches of $N_c \times N_c$ pixels from the spherical images obtained by mapping the omnidirectional images to the unit sphere (see Chapter 2). As cropping a square image patch from a spherical image is feasible only when its center lies on the equator, we rotate the sphere such that the center of the patch coincides with the equator and then we crop the patch. This rotation is taken into account when we estimate the epipolar geometry. Therefore, in the E step we form a set of S_p pairs of stereo patches. For each $p = 1, ..., S_p$ we randomly choose an image pair from the database. We then randomly choose a point on the sphere and extract two patches from two stereo images with the center at the chosen point. Hence, image components exhibit a disparity between the two stereo patches, as shown in Figures 7.4(a) and (e). MVMP is then performed on each pair of patches independently, and N_{at} atoms are selected. Examples of atoms are shown in Figures 7.4(b)-(d) and (f)-(h). The dictionary learning step (M step) is then performed by minimizing the sum of the energy function given by Eq. (7.65) for all patches.

In our experiments, we have taken $S_p = 50$ pairs of patches, randomly selected in each E step during EM iterations. The dictionary was constructed using 12×12 different atom positions, thus spanning all pixels in a 12×12 image patch. To avoid border effects, we have cropped slightly larger patches of 16×16 pixels. The number of rotations for dictionary construction has been chosen as four. Finally, the pairs of scales have been independently randomly initialized for the left and the right dictionary, using five pairs of anisotropic scales in a range $(5, 15)$ (from big to small atoms). The number of atoms selected by the MVMP is three atoms per patch. Since the patch size is small, three atoms are usually enough to represent the main geometrical components in a patch. We have used the edge-like atoms on the sphere, based on the generating function which is a Gaussian in one direction, and its second derivative in the orthogonal direction:

$$g_{HF}(\theta, \varphi) = -\frac{1}{K_A}\left(16\alpha^2 \tan^2 \frac{\theta}{2} \cos^2 \varphi - 2\right) \exp\left(-4 \tan^2 \frac{\theta}{2}\left(\alpha^2 \cos^2 \varphi + \beta^2 \sin^2 \varphi\right)\right). \quad (7.67)$$

These are the same atoms as those used in the compression of omnidirectional images in Section 3.2. Samples of edge-like functions are shown in Figure 3.2. For the weighting function in Eq. (7.48) we have used a Gaussian envelope of the form:

$$w(\theta, \varphi) = \frac{1}{K_G} \exp\left(-4 \tan^2 \frac{\theta}{2}\left(12\alpha^2 \cos^2 \varphi + \beta^2 \sin^2 \varphi\right)\right), \quad (7.68)$$

which gives positive weights to the points on the main (central) lobe of the atom $g_{HF}(\theta, \varphi)$, while outside of the main lobe the weights are close to zero. Weights are higher towards the axis of the discontinuity represented by the atom. Choosing such a weight function for the epipolar geometry estimation enables us to use points that are likely to satisfy the epipolar constraint, and to exclude the points represented by the ripples of the second derivative of the Gaussian.

We have first performed MVMP on a set of randomly selected patches to estimate the variances $\sigma_D^2 = 2.7 \cdot 10^{-3}$ and $\sigma_b^2 = 0.047$. Within the preprocessing step, all patches have been normalized to the same variance 0.1 to equalize the importance of each patch. Moreover, the patches have been whitened to flatten the image spectrum and make all frequencies equally important [Ols97]. The whitening has been performed by spherical filtering [Tos05].

7.6.2 Scale learning results

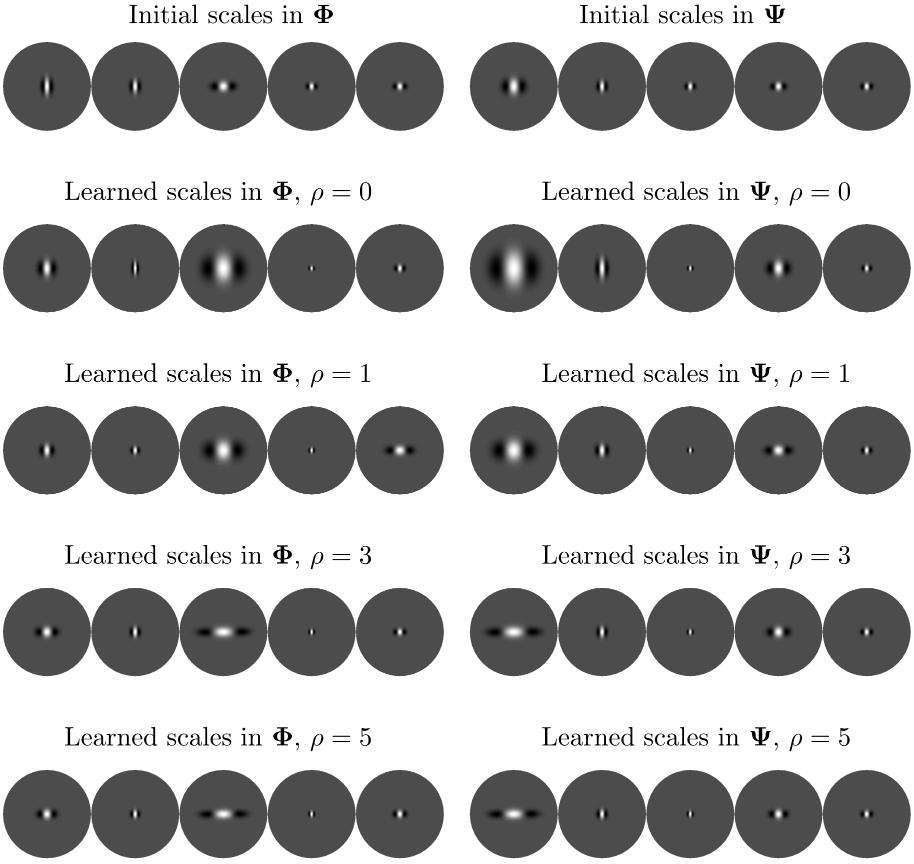
The initial values of scales $\alpha^{(L)}$, $\beta^{(L)}$, $\alpha^{(R)}$ and $\beta^{(R)}$ for the learning algorithm have been chosen randomly, and they are given in the first two columns in Table 7.1. The atoms of the initial scales are shown in the first row in Figure 7.5. The whole dictionary is built from these atoms by shifting them at all pixel locations and rotating in four orientations. To see the influence of the part of the objective function, which relies on the multi-view constraint, we have introduced a factor ρ in the energy function:

$$\begin{aligned} \tilde{E}(y_L, y_R, D = 0, \mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}) &= \frac{1}{2\sigma_I^2}(\|y_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|y_R - \mathbf{\Psi}\mathbf{b}\|_2^2) & (7.69) \\ &+ \rho \frac{1}{2\sigma_D^2} \sum_{l=1}^{M} \sum_{r=1}^{M} \mathcal{I}(a_l)\mathcal{I}(b_r) D_E(\gamma_l^{(L)}, \gamma_r^{(R)}) & (7.70) \\ &+ \rho \frac{1}{2\sigma_b^2} \sum_{l=1}^{M} \sum_{r=1}^{M} (b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). & (7.71) \end{aligned}$$

We can see that for $\rho = 1$, the energy function in Eq. (7.69) is equal to the one in Eq. (7.57). On the other hand, for $\rho = 0$, there are no multi-view constraints in the energy function, and dictionary learning is based only on minimization of the residual energy of sparse representations of stereo images. Columns 3-10 in Table 7.1 give the learned values for the scales $\alpha^{(L)}$, $\beta^{(L)}$, $\alpha^{(R)}$ and $\beta^{(R)}$, for $\rho = 0, 1, 3, 5$. These atoms are shown in Figure 7.5. The learned scales have been

Table 7.1: *Initial and learned scale parameters for the left and the right image, for different values of the parameter* ρ.

Initial dictionary		Learned dictionary							
		$\rho = 0$		$\rho = 1$		$\rho = 3$		$\rho = 5$	
$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$
13.15	5.98	8.61	6.34	10.82	8.68	6.86	10.17	7.84	10.92
14.06	7.78	22.19	7.30	16.92	13.72	14.84	9.95	16.53	12.36
6.27	10.47	3.40	3.56	3.81	5.05	2.82	10.26	3.17	11.39
14.13	14.58	25.88	22.95	26.00	19.73	25.19	18.21	24.36	17.04
11.32	14.65	14.52	14.78	5.57	11.25	12.73	15.63	11.61	13.92
$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$
6.58	6.42	2.94	2.69	3.58	4.73	2.72	9.36	3.01	10.77
14.71	9.22	12.18	5.04	11.72	8.43	14.79	9.66	15.19	11.00
14.57	14.16	25.93	20.30	25.57	18.94	24.75	17.86	23.94	16.55
9.85	12.92	6.60	6.80	5.70	10.56	6.72	10.13	8.22	11.72
13.00	14.59	15.87	16.05	15.08	14.52	13.08	14.97	13.25	14.85



Initial scales in $\mathbf{\Phi}$

Initial scales in $\mathbf{\Psi}$

Learned scales in $\mathbf{\Phi}$, $\rho = 0$

Learned scales in $\mathbf{\Psi}$, $\rho = 0$

Learned scales in $\mathbf{\Phi}$, $\rho = 1$

Learned scales in $\mathbf{\Psi}$, $\rho = 1$

Learned scales in $\mathbf{\Phi}$, $\rho = 3$

Learned scales in $\mathbf{\Psi}$, $\rho = 3$

Learned scales in $\mathbf{\Phi}$, $\rho = 5$

Learned scales in $\mathbf{\Psi}$, $\rho = 5$

Figure 7.5: *Initial and learned scales of the atoms for the left and the right dictionaries. All atoms are on the North pole.*

obtained after 50 iterations of the EM algorithm, after which the change in parameters becomes small, i.e., the solution is stabilized. We have used the same initial scales for all values of ρ.

The learned dictionaries include atoms of different scales, which are hence able to approximate signals at various scales. When $\rho = 0$, the learned atoms are more elongated along the Gaussian direction, while narrower on the direction of the second derivative of the Gaussian. These results are in consistency with the previous work on dictionary learning for image representation in the single view case. However, when we increase ρ, and hence include the geometry constraints in the stereo learning, we obtain different results for atoms scales. The atoms become more elongated along the direction of the Gaussian second derivative and narrower in the direction of the Gaussian, which is the opposite effect than in the case of $\rho = 0$. In addition, for $\rho > 0$ the learned scales generally tend to give smaller atoms than for $\rho = 0$. These two effects of including multi-view geometry in the dictionary learning process are most probably due to the local nature of the epipolar constraint. Namely, the depth of the scene changes rapidly around object boundaries leading to different disparity and epipolar matching in these areas. Since the object boundaries are represented by 2D discontinuities on the image of a 3D scene, the epipolar geometry is satisfied along the discontinuity and in a limited area. This makes the learned atoms become anisotropic and small. Thus, we conclude that for stereo dictionary learning, the geometric constraints need to be considered, otherwise significantly different results are obtained. We can also see that the results for $\rho = 3$ and $\rho = 5$ are close, leading to the conclusion that further increasing ρ is not necessary.

7.6.3 Verification of the learned dictionaries in the distributed setting

We further want to verify if the proposed dictionary learning method for stereo image representation improves the correspondence matching when the atoms are selected independently from the left and the right image. Namely, we do not search for corresponding atoms during sparse approximation using MVMP, but we rather apply MP independently on each image and then match the corresponding atoms. This corresponds to the distributed image approximation method shown in Figure 7.6. If the atoms that form the learned dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ represent the statistically optimal atoms for both image approximation and epipolar geometry, one should expect that using the learned dictionary results in more corresponding atom pairs than a randomly initialized dictionary.

Two image patches y_L and y_R with the same center have been randomly selected from a randomly chosen pair of omnidirectional images in "Mede" database. The size of the patches has been chosen as 40×40 pixels, which is slightly larger than for learning in order to select more atoms using MP. After independent MP decomposition of the left and the right patch using 10 atoms per patch, epipolar constraint has been evaluated for all possible pairs of left and right atoms $d_A(\phi_l, \psi_r)$, $l = 1, ..., 10$, $r = 1, ..., 10$. The epipolar measure d_A is equal to half of the value D_E in Eq. (7.49), i.e., it represents the epipolar atom distance per view.

Figure 7.7(a) plots on the y-axis the number of atom pairs (ϕ_l, ψ_r) that have the epipolar distance $d_A(\phi_l, \psi_r)$ smaller or equal than a threshold value plotted on the x-axis. We will call

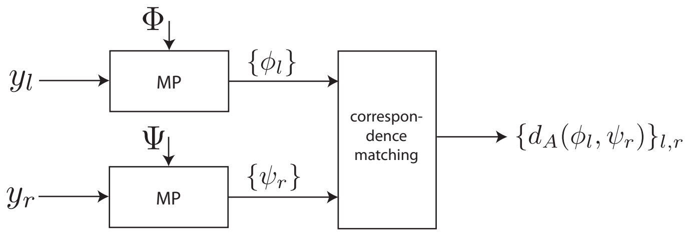
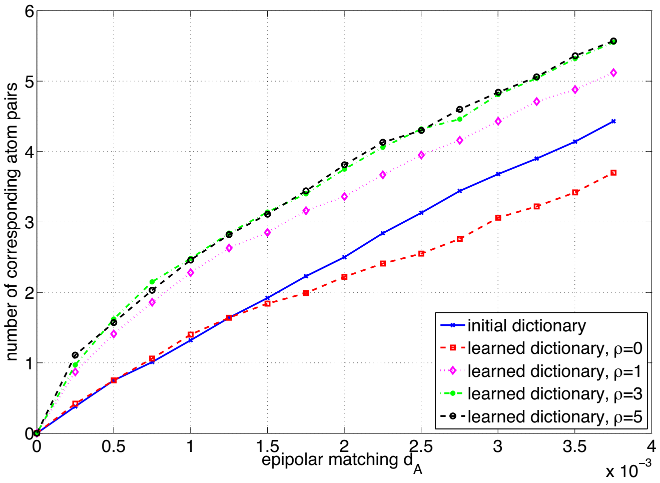
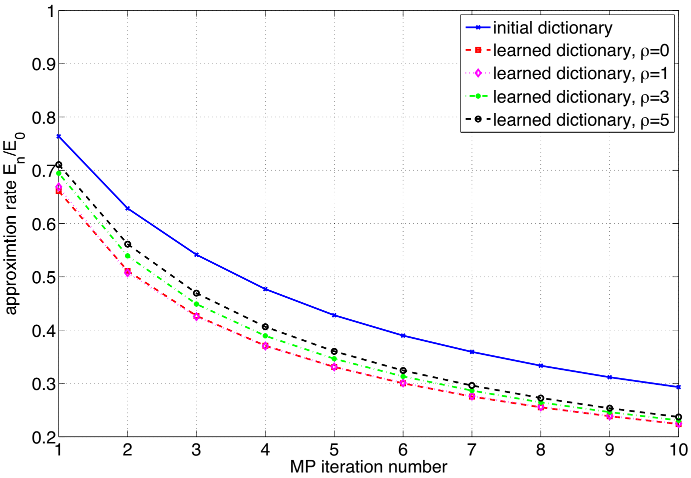


Figure 7.6: *Distributed setting for the correspondence matching from sparse representations of stereo images.*



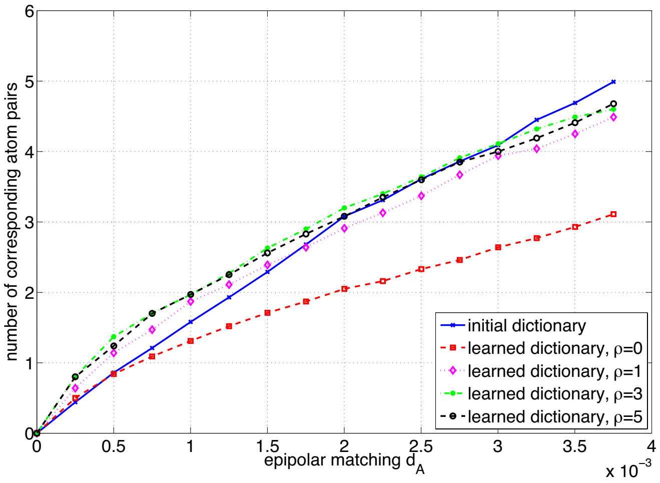
(a)



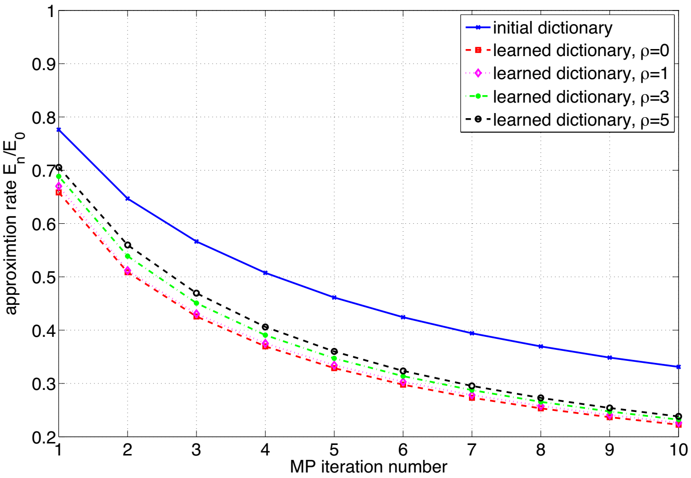
(b)

Figure 7.7: *Performance of the learned dictionaries in a distributed setting: a) Cumulative correspondence number (CCN) curves for initial dictionary and learned dictionaries, for* $\rho = 0, 1, 3, 5$*; b) MP energy decay for* y_L *and* y_R*.*

this curve the cumulative correspondence number (CCN) curve. The left part of CCN curves with smaller d_A is more important than the right part, because the correspondences are more reliable when their epipolar distance is smaller. All CCN curves have been averaged over 100 randomly chosen image pairs. We can see that CCN curves for $\rho = 1$, $\rho = 3$ and $\rho = 5$ are all above the CCN curve for the randomly initialized dictionary, thus confirming that our stereo dictionary learning results in dictionaries that can give higher number of correspondences. On the other hand, for $\rho = 0$, the CCN curve is either close to the CCN curve of a random dictionary, or below it. This makes us conclude that designing dictionaries to best approximate the images, without considering the geometry, can lead to suboptimal stereo dictionaries. Figure 7.7(b) shows the average approximation rate (energy decay) during the iterations of MP for images y_L and y_R. We plot the ratio between the sum of the residues of the left and right images after k iterations and the sum of their initial energies, versus the iteration number k. We can see that for all values of ρ the approximation rate using learned dictionaries is better than using random dictionaries. Moreover,



(a)



(b)

Figure 7.8: *Performance of the learned dictionaries in a distributed setting, with averaging over random initial dictionaries: a) Average Cumulative correspondence number (CCN) curve for 100 random initial dictionaries and CCN curves for learned dictionaries, for* $\rho = 0, 1, 3, 5$*; b) MP energy decay for* y_L *and* y_R*.*

increasing ρ slows down the approximation rate for a very small amount, hence optimizing the dictionaries for stereo matching does not induce a big penalty on the approximation rate.

To verify that the superior performance of the learned dictionaries over the initial ones is not due to the unlucky selection of the random initial dictionaries, we compare the performance of the learned dictionaries to an average performance of different randomly selected initial dictionaries. We select randomly 100 initial dictionaries and 100 stereo image pairs, and plot the average CCN curve for the initial dictionary. This curve is shown with the solid line in Figure 7.8(a). The learned dictionaries are fixed as obtained previously, and the CCN curves are averaged over 100 image pairs. We see that the learned dictionaries still give more correspondences than the random ones, for the smaller value of d_A. Therefore, they lead to more atom pairs with a better epipolar matching. The comparison of the energy decay in this case is shown in Figure 7.8(b), and it is similar to the energy decay in Figure 7.7(b).

7.6.4 Benefits of learning for distributed coding and camera pose estimation

Finally, we discuss how the proposed learning algorithm can be useful for distributed coding presented in Chapter 5 and camera pose estimation presented in Chapter 6. The advantages of using learned dictionaries for distributed coding can be easily understood by looking at Figures 7.7 and 7.8. Since learned dictionaries result in a higher number of correspondences in a distributed setting, the Wyner-Ziv decoder in Section 5.3.3 would be able to find more corresponding atoms and improve the final image reconstruction. At the same time, learned dictionaries for $\rho > 0$ have a small decrease in the approximation rate with respect to the learned dictionary for $\rho = 0$ which is optimized for the best approximation performance. This decrease in approximation rate would be negligible in the overall distributed coding performance.

Since the advantages of using learned dictionaries for camera pose estimation cannot be straightforwardly deduced from the above figures, we have performed two experiments. In the first experiment the camera pose has been estimated using the initial dictionaries, while in the second one we use the learned dictionaries for $\rho = 3$. Two pairs of images from the "Mede" database have been selected, with translation $\mathbf{T} = [1\ 0\ 0]^\mathsf{T}$ and $\mathbf{T} = [0\ 1\ 0]^\mathsf{T}$. For each image pair, we have randomly chosen 20 patches of 40×40 pixels. The same image patches have been decomposed by MP using initial and learned dictionaries, giving 3 atoms for each patch and thus 60 atoms per image. Matching atoms from left and right images and estimation of the camera pose has been performed using the algorithm proposed in Chapter 6, with and without Ransac. We have taken more points per atom, as explained in Section 6.5.4, in order to see the effect of learning the atom shape. The estimated translation matrices are shown in Table 7.2 and Table 7.3, for target matrices $\mathbf{T} = [1\ 0\ 0]^\mathsf{T}$ and $\mathbf{T} = [0\ 1\ 0]^\mathsf{T}$, respectively. We can see that using the learned dictionaries for pose estimation gives significantly better performance than using randomly initialized dictionaries. The learned dictionaries lead to a precise estimation of the translation, while the initial dictionaries cannot even determine the direction of camera motion. Moreover, applying Ransac does not improve the performance in the case of the learned dictionary. This leads to the conclusion that all points on the learned atoms give reliable matches without gross outliers.

Table 7.2: *Comparison of camera pose estimation with random initial and respectively learned dictionaries, for target translation* $\mathbf{T}^\mathsf{T} = [1\ 0\ 0]$.

Target matrices	$\mathbf{T}^\mathsf{T} = [1\ 0\ 0]$ $\mathbf{R} = \mathbf{I}$	
	learned dictionary	initial dictionary
estimated $\mathbf{T}^\mathsf{T}$, without Ransac	[0.9778 -0.1519 0.1441]	[0.1910 -0.7284 0.6580]
estimated $\mathbf{T}^\mathsf{T}$, with Ransac	[0.8144 -0.5436 0.2032]	[0.7540 0.6067 0.2518]

Table 7.3: *Comparison of camera pose estimation with random initial and respectively learned dictionaries, for target translation* $\mathbf{T}^\mathsf{T} = [0\ 1\ 0]$.

Target matrices	$\mathbf{T}^\mathsf{T} = [0\ 1\ 0]$ $\mathbf{R} = \mathbf{I}$	
	learned dictionary	initial dictionary
estimated $\mathbf{T}^\mathsf{T}$, without Ransac	[-0.0385 0.9951 0.0907]	[0.9067 0.3484 0.2376]
estimated $\mathbf{T}^\mathsf{T}$, with Ransac	[-0.1317 0.9424 0.3075]	[0.9003 0.3934 0.1866]

7.7 Related work on dictionary learning

As mentioned in Section 7.3, Olshausen and Field have introduced the earliest work on learning overcomplete dictionaries for image representation [Ols97, Ols96]. Their maximum likelihood (ML) method has been also extended to dictionary learning from natural videos [Ols03, Ols07, Cad08].

The probabilistic inference approach to overcomplete dictionary learning has been later adopted by other researchers. Engan et al. have introduced a method of optimal directions (MOD), which includes the sparse coding and dictionary update steps that iteratively optimize the objective ML function [Eng99a, Eng99b]. Their method differs from the work of Olshausen and Field in two aspects. First, while in the work of Olshausen and Field the sparse coding step involves finding the equilibrium solution of the differential equation over $\mathbf{a}$, MOD uses either OMP [Eng99a] or FOCUSS [Eng99b] algorithms to find a sparse vector $\mathbf{a}$. Second, the dictionary $\mathbf{\Phi}$ is updated as the solution of the differential equation $\partial E/\partial \mathbf{\Phi} = 0$, where E is the energy function that is, in this case, equal to the residue $\|y - \mathbf{\Phi}\mathbf{a}\|_F^2$ ($\|\cdot\|_F$ denotes the Frobenius norm). These two modifications make the MOD approach faster compared to the ML method of Olshausen and Field. Maximum a posteriori (MAP) dictionary learning method, proposed by Kreutz-Delgado et al. [Kre03] belongs also to the family of two-step iterative algorithms based on probabilistic inference. Instead of maximizing the likelihood $P(y|\mathbf{\Phi})$, MAP method maximizes the posterior probability $P(\mathbf{\Phi}, \mathbf{a}|y)$. This essentially reduces to the same two-step (sparse coding-dictionary update) algorithm, where dictionary update includes an additional constraint on the dictionary that can be for example the unit Frobenius norm of $\mathbf{\Phi}$ or the unit l_2 norm of all atoms in the dictionary. Sparse coding step is performed with FOCUSS [Gor97].

A slightly different line of dictionary learning techniques is based on vector quantization (VQ) achieved by K-means clustering. The VQ approach for dictionary learning has been first proposed by Schmid-Saugeon and Zakhor, within the Matching Pursuit video coding application [Sch04, Sch01]. Their algorithm optimizes a dictionary given a set of image patches by first grouping patterns such that their distance to a given atom is minimal. It then updates the atom such that the overall distance in the group of patterns is minimal. The implicit assumption here is that each patch can be represented by a single atom with a coefficient equal to one. This reduces the learning procedure to K-means clustering. Since each patch is represented by only one atom, the sparse coding step is trivial here. A generalization of the K-means for dictionary learning, called K-SVD algorithm, has been proposed by Aharon et al. [Aha06]. After the sparse coding step (where any pursuit algorithm can be employed), the dictionary update is performed by sequentially updating each column of $\mathbf{\Phi}$ using the singular value decomposition (SVD) to minimize the approximation error. The update step is hence a generalized K-means since each patch can be represented by more atoms with different weights.

Finally, there exist other approaches for learning special types of dictionaries, like unions of orthonormal basis [Gri03], shift-invariant dictionaries [Jos06b], or block-based dictionaries and constrained overlapping dictionaries [Eng07]. A comparison of all state of the art dictionary learning methods is difficult to make. The efficiency of existing algorithms differs with the dictionary size and the training data. However, ML and MAP methods are characterized by a flexibility of extending the probabilistic modeling to higher-dimensional data, like videos [Ols07, Cad08] or stereo images. It is also possible to include different modalities that have correlated nature, such as audio and visual signals in order to learn audio-visual dictionaries [Mon08a, Mon08b]. Because of this property of the ML approach, we have selected it for learning parametric dictionaries in the stereo case. Since the definition of the parametric dictionary already includes atoms normalization, the MAP approach with prior on the dictionary norm would not be beneficial in our case.

7.8 Conclusion

Our contribution in this chapter is a novel method for learning the overcomplete dictionaries that have optimal performance in representing stereo images. The multi-view image model presented in Chapter 4 has served as basis for developing a maximum likelihood (ML) method for stereo dictionary learning. The experimental results have shown that one has to consider the geometric constraints in order to obtain optimal stereo atoms. The learned dictionaries give both better stereo matching and approximation properties than randomly selected dictionaries. Finally, we have shown that dictionary learning for optimal scene representation has important benefits for the two applications addressed in this thesis, the distributed coding (Chapter 5) and the camera pose estimation (Chapter 6). Therefore, learning of stereo dictionaries constitutes an important step in image-based 3D scene representation with sparse approximations.

Chapter 8

Conclusion

This thesis has addressed the problem of efficiently representing a 3D scene captured by multiple distributed cameras. We have targeted one of the most important challenges in such a framework: modeling the multi-view images and the underlying multi-view correlation between them, in order to simultaneously achieve efficient compression and scene geometry estimation. In particular, we have considered the use of omnidirectional cameras for scene representation, which are particularly advantageous due to their wide field of view.

First, we proposed a compression scheme for omnidirectional images based on sparse approximations of spherical images using a redundant dictionary of oriented and anisotropic atoms on the sphere. Sparse approximations are interesting solutions to the image representation problem as they usually lead to high compression gains, and are also believed to be a strategy of image coding in the early stages of visual information processing in the mammalian brain. Besides omnidirectional image compression, we have applied sparse approximations on the sphere to 3D object coding.

The next contribution of this thesis is the development of a new multi-view correlation model that has numerous advantages compared to existing models. Namely, our model is capable of representing diverse transforms between image features in multiple views, and it deals with the spatial locality of these transforms. Multi-view images are represented as sparse approximations over transform-invariant redundant dictionaries of geometric atoms; this leads to an elegant multi-view correlation model that relates geometrically corresponding atoms by local transforms. The proposed model has been successfully used for the design of a distributed coding scheme for multi-view omnidirectional images, which alleviates the necessity of inter-camera communication at the encoder side. At the same time, the coding scheme effectively removes the inter-view correlation to achieve high compression gains at low rates, where the coding of image geometry is dominant. Moreover, we have shown that the proposed model can be used to efficiently estimate the camera pose and a coarse depth map of the scene from low bit rate descriptions of multi-view omnidirectional images.

Finally, we have proposed a novel method for stereo dictionary learning based on our multi-view image model. We have developed a maximum likelihood method for stereo dictionary learning, which includes the epipolar geometry constraint in a probabilistic model. Experimental results show that we have to consider the geometry constraints to obtain optimal stereo atoms, which give both better stereo matching and approximation properties than randomly selected dictionaries, even in distributed settings. Moreover, learning the dictionaries for optimal scene representation improves camera pose estimation and can be beneficial for distributed coding due to the increased number of atom correspondences.

There are many exciting directions that this research can be taken in. First, in distributed coding for camera networks the proposed correlation modeling offers a solution for low bit rate coding, i.e., for encoding the geometry or structure of multi-view images. However, the correlation model is purely geometric and does not consider textures. Therefore, one can envision a hybrid coding method where the proposed model would serve as a geometry estimation and compensation step, followed by texture coding step with texture-based dictionaries. Alternative sparse approximation algorithms, such as Basis Pursuit Denoising or Sparse Bayesian Learning, can be applied instead of MP in the distributed coder. The complexity and memory limitations due to the huge dictionary size can be overcome by dictionary splitting. The coset design can also be a topic of future research. One can envision a joint dictionary partitioning strategy for all dictionary parameters, where atoms with small epipolar distance that form curves in the dictionary parameters space, are distributed in different cosets.

The proposed correlation model can be also applied within the field of distributed compressed sensing. The proposed MVMP algorithm based on the new multi-view image model can be applied to the reconstruction of a 3D scene from random measurements taken from multiple views. Besides camera pose estimation, our atom pairing model can be used in other problems in computer vision, like multi-view object localization or motion tracking. Multi-view tracking based on sparse representations and compressed sensing has been recently proposed by Reddy et al. [Red08]. However, their multi-view image model does not include the geometry constraints. Applying our model in this setting would thus contribute important tracking information. Finally, with an appropriate experimental setup, the novel stereo dictionary learning method can be used as a mathematical framework to study the receptive fields of binocular cells in human vision and their relation to vergence eye movements.

APPENDIX

Appendix A

Appendix

A.1 Recovery of sparse correlated signals by thresholding

We establish here the conditions for the recovery (identification) of sparse components in correlated signals, using distributed thresholding. In particular, we consider the signals that follow the correlation model established in Chapter 4, where sparse components of different signals are linked with local transforms. Thresholding is a fast algorithm where sparse components are simply chosen as those that have the highest inner product with the signal. Due to its low complexity, thresholding represents an interesting alternative to MP in applications proposed in Chapters 5 and 6. However, this algorithm does not guarantee in general to find the correct signal elements. The sufficient condition for the correct signal recovery by thresholding has been given in [Gri08b]. The contribution of this part of the thesis is the theoretical analysis of the conditions that correlated signals have to fulfill such that their sparse support can be identified by thresholding. We consider here the recovery of correlated signals and look at the general problem, where the local transforms can be arbitrary. Besides multi-view images, a variety of correlated signal sets can be modeled this way, such as videos or seismic signals. For example, these signals can be obtained by a set of sensors that look at the same event, but record different observations, as shown in Figure A.1.

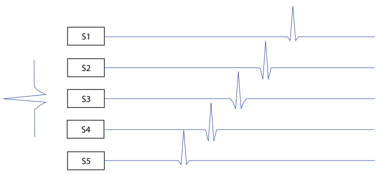


Figure A.1: *A set of sensors that observe the same event and record correlated observations.*

Recovery of correlated signals by thresholding has been considered in [Sch06], where the correlated signals share a common sparse support and the same coefficients, but are observed under different noisy conditions. The authors have shown that the variability of noise in different observations can boost some sparse components and therefore help the thresholding algorithm to find the correct signal support. Sparse representation of correlated signals has been further studied within the concept of distributed compressed sensing [Dua05]. Two different correlation models have been introduced: common sparse component with sparse innovations, and common sparse supports. The recovery algorithms proposed for these two models are based on greedy algorithms.

The thresholding recovery analysis for correlated signals that we present here differs from the previous work [Sch06, Dua05] in one major assumption: we do not require the signals to share the same support (i.e., to have exactly the same atoms in the representation). Instead, we allow each atom in one signal to have its corresponding atom in another signal, which is obtained by a local transform such as shift (translation), scaling, rotation or any combination of those. We validate the sufficient recovery condition on randomly generated 1D signals and illustrate its usage in the particular case of seismic signals.

A.1.1 Sparse signal correlation model

We consider two signals y_1 and y_2 that have sparse representations in dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, respectively. In general, we consider those dictionaries as different, but they can also be equivalent in special cases. We assume that the signals are not exactly sparse, but they can be approximated by sparse decompositions of m atoms up to an approximation error, i.e.:

$$\begin{aligned} y_1 &= \mathbf{\Phi}_I a_1 + e_1 = \sum_{k=1}^{m} a_{1,k}\phi_{i_k} + e_1, \\ y_2 &= \mathbf{\Psi}_J a_2 + e_2 = \sum_{k=1}^{m} a_{2,k}\psi_{j_k} + e_2, \end{aligned} \tag{A.1}$$

where $I = \{i_k\}$ and $J = \{j_k\}, k = 1, ...m$ label the sets of atoms that participate in the sparse decompositions of y_1 and y_2, respectively. Matrices $\mathbf{\Phi}_I$ and $\mathbf{\Psi}_J$ denote respectively sub-matrices of $\mathbf{\Phi}$ and $\mathbf{\Psi}$ with respect to I and J. We are particularly interested in signals that are correlated by local transforms of atoms in different signals. The correlation model is described in the rest of this section through two assumptions.

Assumption 1. We assume that signals y_1 and y_2 are correlated in the following way:

$$y_2 = \sum_{k=1}^{m} a_{2,k}\psi_{j_k} + e_2 = \sum_{k=1}^{m} a_{2,k}F_k(\phi_{i_k}) + e_2, \tag{A.2}$$

where $F_k(\cdot)$ denotes the transform of an atom ϕ_{i_k} in y_1 to an atom ψ_{j_k}, and it differs for each $k = 1, ..., m$.

In particular, we consider a special class of transforms F, which result from a linear transform of the coordinate system in the space where the signal and the dictionaries are defined. Let $\mathbf{v}$ denote the unit vector of coordinates and $\mathbf{u}$ denote the vector of coordinates obtained by transforming $\mathbf{v}$ with an arbitrary linear transform Q, i.e., $\mathbf{u} = Q(\mathbf{v})$. Let further an atom ϕ be defined as a continuous function in a Hilbert space $\mathcal{H}$ normalized to have the l_2 norm equal to one, i.e., $\phi = g(\mathbf{v})/\|g\|$. Equivalently, we define $\psi = h(\mathbf{v})/\|h\|$. We consider the atom transforms F that satisfy:

$$h(\mathbf{v}) = F(g(\mathbf{v})) = g(Q(\mathbf{v})) = g(\mathbf{u}). \tag{A.3}$$

The second equality in Eq. (A.3) directly defines the class of transforms F considered in this work, which result from a linear transform T. These types of transforms have been shown to be of great practical use, for example in dictionary design for sparse image approximation [Fig06]. The considered transforms can be illustrated by the following example.

Example 1. *Consider a function $g(x) \in \mathbb{R}$ and an overcomplete dictionary obtained by local transforms of $g(x)$, such as shifts and scaling. These transforms can be realized by a single transform of the x coordinate, which includes translation b and anisotropic scaling s; i.e., $x' = (x - b)/s$. This class of transforms obviously satisfies Eq. (A.3), where $\mathbf{v} = x$ and $\mathbf{u} = x'$.*

We assume further that the signals satisfy the local transforms applied to each atom, within the local support of that atom:

Assumption 2. For all k and $\mathbf{v}$ such that $\phi_{i_k}(\mathbf{v}) \neq 0$, it holds:

$$y_2(\mathbf{v}) = y_1(Q_k(\mathbf{v})) = y_1(\mathbf{u}). \tag{A.4}$$

The type of signal correlation under different local transforms given by Assumptions 1 and 2 can be found in practical cases where the same signal is observed by sensors at different positions. Locality of the transforms is highly important in practice as different parts of the signal can be captured under different transforms, like for example at different distances to the sensor. Since the signal correlation model includes a noise component, slight deviations from the assumed model (e.g., occlusions or interference) can be considered as noise components and hence the signal correlation model is not very restrictive.

We are now interested in establishing the conditions under which thresholding, performed independently on each signal, recovers the correct sparse representations of signals y_1 and y_2. This can be stated as follows:

Problem 1. *Assume that we are given two correlated signals* y_1 *and* y_2 *in Eq. (A.1), and the assumptions 1 and 2 hold. Suppose that thresholding recovers the correct sparsity pattern* I *of the signal* y_1*. We want to derive the sufficient condition for the correct sparse recovery of the sparsity pattern* J *of the signal* y_2 *using thresholding.*

A.1.2 Single signal thresholding

We review here the conditions under which simple thresholding algorithm recovers correctly the sparse representation of the signals [Gri08b]. Before describing their result, let us define some functions that will be used throughout the analysis. *Setwise cumulative coherence function*, defined in [Gri08a], is given as:

$$\mu_1(\mathbf{\Phi}, I) := \sup_{k \notin I} \sum_{i \in I} |\langle \phi_k, \phi_i \rangle|. \tag{A.5}$$

Note however that this function is different from the cumulative coherence given in [Tro04]. Next, the *Dictionary Inter Symbol Interference* is given as:

$$ISI(\mathbf{\Phi}, I) := \mu_1(\mathbf{\Phi}, I) + \sup_{l \in I} \mu_1(\mathbf{\Phi}_I, I \setminus \{l\}). \tag{A.6}$$

This function measures the interference of atoms in the sparse decomposition that can lead to incorrect recovery. We will denote the second term in the expression for $ISI(\mathbf{\Phi}, I)$ as $\chi(\mathbf{\Phi}_I, I)$, i.e.,

$$\chi(\mathbf{\Phi}_I, I) := \sup_{l \in I} \mu_1(\mathbf{\Phi}_I, I \setminus \{l\}). \tag{A.7}$$

The recovery condition is given by the following theorem:

Theorem A.1.1 (Gribonval, Nielsen, Vandergheynst [Gri06])**.** *Let* $y = \mathbf{\Phi} x + e$ *be a noisy sparse representation of the data. Moreover, assume* $x_{l_k}, k = 1, ...|I|$ *are the* $|I|$ *nonzero components of* x *in decreasing order of magnitude, i.e.,* $|x_{l_1}| \geqslant |x_{l_2}| \geqslant ... \geqslant |x_{l_{|I|}}|$*. If the following condition is satisfied:*

$$\frac{|x|_{l_m}}{||x||_\infty} > \frac{||\mathbf{\Phi}_I^* e||_\infty + ||\mathbf{\Phi}_{\bar{I}}^* e||_\infty}{||x||_\infty} + ISI(\mathbf{\Phi}, I) \tag{A.8}$$

then each inner product of the observed data y *with the atoms* $\{\phi_{l_i}\}_{1 \leqslant i \leqslant m}$ *exceeds all the inner products with the atoms* $\{\phi_i\}_{i \in I \setminus \{l_1, \ldots, l_m\}}$ *indexed by the complementary set* $\bar{I}$*. In particular the* $m = |I|$ *largest inner products correspond exactly to the support* I *of* x*.*

For the general proof, please see the generalization of the theorem to the multi-channel case in [Gri08b].

A.1.3 Thresholding of correlated signals

We first assume that the sparsity pattern I of the signal y_1 can be recovered by thresholding, i.e., that signal y_1 satisfies:

$$\frac{|a_{1,m}|}{||a_1||_\infty} > \frac{||\mathbf{\Phi}_I^* e_1||_\infty + ||\mathbf{\Phi}_{\bar{I}}^* e_1||_\infty}{||a_1||_\infty} + ISI(\mathbf{\Phi}, I). \tag{A.9}$$

Before establishing the recovery conditions for the signal y_2, we prove the following lemma:

Lemma A.1.1. *Let two correlated signals* y_1 *and* y_2 *in Eq. (A.1) be correlated by the model in Eq. (A.2), with transforms defined by Eq. (A.3). Let further assume that the condition in Eq. (A.4) holds. Then, for all* $k = 1, ..., m$, *it holds:*

$$\langle y_2, \psi_{j_k} \rangle = C_k \langle y_1, \phi_{i_k} \rangle, \tag{A.10}$$

where:

$$C_k = \frac{1}{\sqrt{|\frac{\partial Q_k(\mathbf{v})}{\partial \mathbf{v}}|}} = \frac{1}{\sqrt{|\frac{\partial \mathbf{u}}{\partial \mathbf{v}}|}} = \frac{1}{\sqrt{J_k}}, \tag{A.11}$$

and $J_k = |\frac{\partial \mathbf{u}}{\partial \mathbf{v}}|$ *is the Jacobian of the linear transform* Q_k.

Proof. From the definition of the inner product, we have:

$$\langle y_2, \psi_{j_k} \rangle = \int\limits_{\widetilde{S}_k} y_2(\mathbf{v}) \psi_{j_k}(\mathbf{v}) d\mathbf{v}, \tag{A.12}$$

where $\widetilde{S}_k$ represents the subspace where $\psi_{j_k}(\mathbf{v}) \neq 0$. Substituting $\psi_{j_k} = h(\mathbf{v})/\|h\|$, we get:

$$\langle y_2, \psi_{j_k} \rangle = \int\limits_{\widetilde{S}_k} y_2(\mathbf{v}) \frac{h(\mathbf{v})}{\|h\|} d\mathbf{v}. \tag{A.13}$$

The l_2 norm $\|h\|$ can be evaluated as follows:

$$\|h\| = \sqrt{\langle h, h \rangle} = \sqrt{\int\limits_{\widetilde{S}_k} h^2(\mathbf{v}) d\mathbf{v}} = \sqrt{\int\limits_{S_k} g^2(\mathbf{u}) |\frac{\partial \mathbf{v}}{\partial \mathbf{u}}| d\mathbf{u}}. \tag{A.14}$$

The last equality is obtained by applying a change of variables theorem, using Eq. (A.3) and Eq. (A.4) for $\mathbf{u} = Q_k(\mathbf{v})$. Since Q_k is defined as a linear transform of coordinates, the mapping $S_k \to \widetilde{S}_k$ is smooth, and the change of variables theorem holds. Furthermore, we have that $|\partial \mathbf{v}/\partial \mathbf{u}| = 1/J_k$ does not depend on $\mathbf{v}$, and it can go in front of the integral:

$$\|h\| = \frac{1}{\sqrt{J_k}} \|g\|. \tag{A.15}$$

We can now go back to Eq. (A.13) and similarly apply a change of variables, using Eq. (A.3), Eq. (A.4) and Eq. (A.15) and obtain:

$$\langle y_2, \psi_{j_k} \rangle = \int\limits_{S_k} y_1(\mathbf{u}) \sqrt{J_k} \frac{g(\mathbf{u})}{\|g\|} \frac{1}{J_k} d\mathbf{u}. \tag{A.16}$$

Finally, $1/\sqrt{J_k}$ can go in front of the integral as a constant C_k and we get:

$$\langle y_2, \psi_{j_k} \rangle = 1/\sqrt{J_k} \int\limits_{S_k} y_1(\mathbf{u}) \phi_{j_k}(\mathbf{u}) d\mathbf{u} = C_k \langle y_1, \phi_{i_k} \rangle. \tag{A.17}$$

□

If we go back now to our Example 1, we have the Jacobian that is equal to: $|\frac{\partial(x')}{\partial(x)}| = 1/s_k$ and hence $C_k = \sqrt{s_k}$, where s_k is the scale transform for each pair of correlated atoms (ϕ_{i_k}, ψ_{j_k}).

We can now give the recovery condition for the signal y_2:

Theorem A.1.2. *Suppose we are given two correlated signals* $y_1 = \mathbf{\Phi}_I a_1 + e_1$ *and* $y_2 = \mathbf{\Psi}_J a_2 + e_2$, *which satisfy assumptions 1 and 2. Furthermore, suppose that thresholding recovers the correct sparsity pattern* I *of the signal* y_1, *i.e., the condition given by Eq. (A.9) is satisfied. If for all* $k = 1, ..., m$ *the following sufficient condition is satisfied:*

$$C_k |a_{1,m}| > C_k \|a_1\|_\infty \chi(\mathbf{\Phi}, I) + C_k \|\mathbf{\Phi}_{\bar{I}}^* e_1\|_\infty + \|\mathbf{\Psi}_{\bar{J}}^* e_2\|_\infty + \|a_2\|_\infty \mu_1(\mathbf{\Psi}, J) \tag{A.18}$$

where $C_k = 1/\sqrt{|\frac{\partial Q_k(\mathbf{v})}{\partial \mathbf{v}}|}$, *then thresholding recovers all sparse components of the signal* y_2, *i.e., each inner product of the signal* y_2 *with the atoms* $\{\psi_{j_k}\}_{1 \leqslant k \leqslant m}$ *exceeds all the inner products with the atoms* $\{\psi_i\}_{i \in J \setminus \{j_1, ..., j_m\}}$ *indexed by the complementary set* $\bar{J}$.

Proof. We start the proof by bounding the inner products of the signal y_1 with $\{\phi_{i_k}\}, k = 1, ..., m$, similarly to the proof of the Theorem 9 in [Gri08b]:

$$\{\langle y_1, \phi_{i_k} \rangle\}_{i_k \in I} = \mathbf{\Phi}_I^* y_1 = \mathbf{\Phi}_I^* \mathbf{\Phi}_I a_1 + \mathbf{\Phi}_I^* e_1 = a_1 + (\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d) a_1 + \mathbf{\Phi}_I^* e_1,$$

where the i_k-th term $1 \leqslant k \leqslant m$ can be bounded as follows:

$$\begin{aligned} |\langle y_1, \phi_{i_k} \rangle| &\geqslant |a_{1,k}| - \|(\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d) a_1\|_\infty - \|\mathbf{\Phi}_I^* e_1\|_\infty \\ &\geqslant |a_{1,m}| - \|(\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d) a_1\|_\infty - \|\mathbf{\Phi}_I^* e_1\|_\infty. \end{aligned} \tag{A.19}$$

From Lemma 1, we have $\langle y_2, \psi_{j_k} \rangle = C_k \langle y_1, \phi_{i_k} \rangle$. When combined with Eq. (A.19) it gives the following inequality:

$$|\langle y_2, \psi_{j_k} \rangle| \geqslant C_k (|a_{1,m}| - \|(\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d) a_1\|_\infty - \|\mathbf{\Phi}_I^* e_1\|_\infty). \tag{A.20}$$

In order for $|\langle y_2, \psi_{j_k} \rangle|$ to be recovered by thresholding, the following condition has to be satisfied for all $1 \leqslant k \leqslant m$:

$$|\langle y_2, \psi_{j_k} \rangle| > \sup_{l \notin J} |\langle y_2, \psi_l \rangle|. \tag{A.21}$$

We have that:

$$\sup_{l \notin J} |\langle y_2, \psi_l \rangle| = \|\mathbf{\Psi}_{\bar{J}}^* y_2\|_\infty \leqslant \|\mathbf{\Psi}_{\bar{J}}^* e_2\|_\infty + \|\mathbf{\Psi}_{\bar{J}}^* \mathbf{\Psi}_J a_2\|_\infty, \tag{A.22}$$

so we have to show that the condition in Eq. (A.18) implies the following inequality:

$$C_k (|a_{1,m}| - \|(\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d) a_1\|_\infty - \|\mathbf{\Phi}_I^* e_1\|_\infty) > \|\mathbf{\Psi}_{\bar{J}}^* e_2\|_\infty + \|\mathbf{\Psi}_{\bar{J}}^* \mathbf{\Psi}_J a_2\|_\infty, \tag{A.23}$$

or equivalently:

$$C_k |a_{1,m}| > C_k \|(\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d) a_1\|_\infty + C_k \|\mathbf{\Phi}_I^* e_1\|_\infty + \|\mathbf{\Psi}_{\bar{J}}^* e_2\|_\infty + \|\mathbf{\Psi}_{\bar{J}}^* \mathbf{\Psi}_J a_2\|_\infty. \tag{A.24}$$

Tropp has shown that the following inequalities hold [Tro04]:

$$\frac{\|(\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d) x\|_\infty}{\|x\|_\infty} \leqslant |||(\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d)|||_{\infty,\infty} = |||(\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d)|||_{1,1} = \sup_{l \in I} \mu_1(\mathbf{\Phi}_I, I \setminus \{l\}), \tag{A.25}$$

$$\frac{\|\mathbf{\Phi}_{\bar{I}}^* \mathbf{\Phi}_I x\|_\infty}{\|x\|_\infty} \leqslant |||\mathbf{\Phi}_{\bar{I}}^* \mathbf{\Phi}_I|||_{\infty,\infty} = |||\mathbf{\Phi}_I^* \mathbf{\Phi}_{\bar{I}}|||_{1,1} = \mu_1(\mathbf{\Phi}, I). \tag{A.26}$$

Therefore, we have:

$$\|(\mathbf{\Phi}_I^* \mathbf{\Phi}_I - I_d) a_1\|_\infty \leqslant \|a_1\|_\infty \sup_{l \in I} \mu_1(\mathbf{\Phi}_I, I \setminus \{l\}) = \|a_1\|_\infty \chi(\mathbf{\Phi}, I), \tag{A.27}$$

$$\|\mathbf{\Psi}_{\bar{J}}^* \mathbf{\Psi}_J a_2\|_\infty \leqslant \|a_2\|_\infty \mu_1(\mathbf{\Psi}, J). \tag{A.28}$$

We can thus conclude that the condition given in Eq. (A.18) which has the same right hand side term as Eq. (A.24), but lower bounded by Eq. (A.25) and Eq. (A.26) implies also the condition in Eq. (A.24). □

The derived condition in Eq. (A.18) represents the worst case analysis solution and in general case, it is not tight. However, the novel condition does not include the value of $\chi(\mathbf{\Psi}, J)$ as the condition in Eq. (A.8) for signal y_2 would include when the correlation model is not considered. Therefore, the new condition is less constraining than Eq. (A.8) for the considered correlation model.

A.1.4 Verification of the results

A.1.4.1 Randomly generated 1D signals — The sufficient condition given by the novel theorem A.1.2 has been verified on pairs of one-dimensional synthetic signals. We have generated a dictionary of size M=1000, for signals of length N=700. We have constructed a parametric dictionary, where a generating function undergoes random shift and scaling operations to generate different atoms in the dictionary. We have used the second derivative of the Gaussian as the generating function, i.e., $g(x) = (4x^2 - 2)\exp(-(x^2))$. The dictionary has been constructed by applying the coordinate transform $x' = (x - b)/s$. The shifts b have been selected randomly from 1 to N, while the scales s were chosen randomly from a uniform distribution on the logarithmic scale from -1 to 3. All atoms have been normalized to have the unit norm. The Jacobian of this transform is $1/s$, and hence we have the constant $C = \sqrt{s}$. The same dictionary has been used for both signals, hence $\mathbf{\Phi} = \mathbf{\Psi}$.

We have performed experiments in noiseless and noisy scenarios. In both cases the signal y_1 has been chosen such that the condition in Eq. (A.9) is fulfilled, for different values of the sparsity m. The sparsity pattern I and the coefficients a_1 have been chosen randomly. To construct the correlated signal y_2 we have randomly chosen different transforms $Q_k, k = 1, ..., m$ defined by b_k and s_k. For each atom in the sparse support of the signal, b is chosen in the range $(-2, 2)$ and the scale s is chosen within $(1, s_{max})$. The transformed atoms from I then yield the sparse support for the signal y_2, denoted as J, and also give the values of C_k for each pair of atoms $(\phi_{i,k}, \psi_{j,k}), k = 1, ..., m$. The two signals have been constructed so that they verify the Lemma A.1.1.

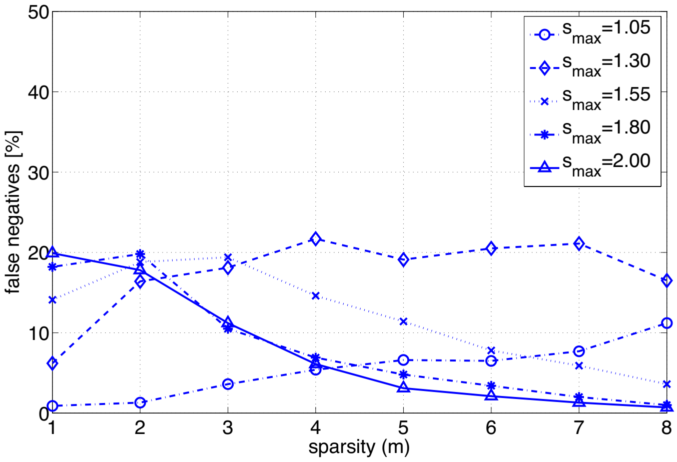


Figure A.2: *Number of false negatives versus the sparsity m, for different values of the maximal scaling parameters.*

We have verified the sufficient condition in Eq. (A.18) by running experiments over 10 different realizations of the dictionary. For each dictionary we have performed 100 trials on y_1 and y_2 constructed as explained above. No false positive has been recorded, and all components where Eq. (A.18) holds have been recovered. This confirms that the condition is indeed sufficient. The condition of Eq. (A.18) is however not necessary, as it is based on a worst case analysis. In order to evaluate the quality of the bound, we count the number of false negatives (when the condition is not fulfilled but thresholding still recovers the correct J). In Figure A.2 we plot the percentage of false negatives (FN) depending on the number of sparse components m, for different maximal values of the transform scale s_{max} between all pairs of signals y_1 and y_2. The maximum number of false negatives has been recorded for small sparsity values m and for larger values of s_{max}. For large m and s_{max}, the number of FN reduces. Although this might look counter-intuitive, it can be easily explained. In order for the signal y_1 to be recovered by thresholding, it needs to have a low value of $\chi(\mathbf{\Phi}, I)$, which then reduces the first term on the right hand side of Eq. (A.18), thus

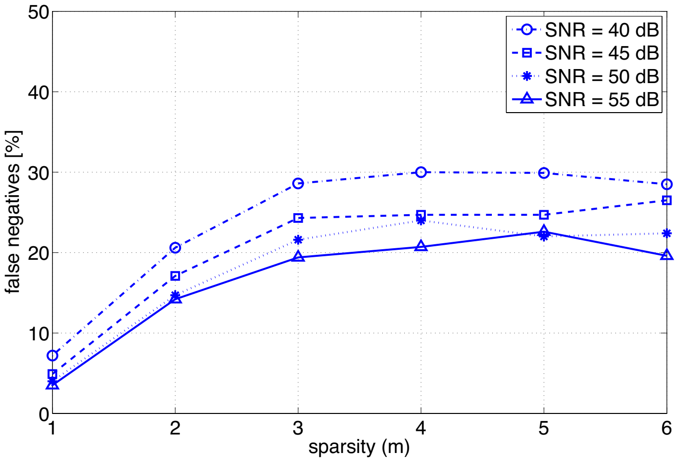


Figure A.3: *Number of false negatives versus the the sparsity m for different SNR, the maximum scale is set constant $s = 1.5$.*

the condition holds in more cases. Finally, we can see that the number of FN is small for most cases.

Finally, Figure A.3 shows the number of FN as a function of m, in the case where signals are distorted by additive white Gaussian noise. We can see that FN is now higher for smaller SNR value but then tends to the values obtained in the noiseless case when SNR increases. The influence of noise is smaller for small m or equivalently, for higher sparsity.

A.1.4.2 Seismic 1D signals — Seismic signals captured at neighboring locations exhibit the type of correlation assumed by Assumption 1. Two seismic signals shown in Figure A.4 are obviously correlated and the second signal is shifted towards the front with respect to the first signal. This shift is important to detect in seismic signals as it represents the propagation of the seismic wave. In the following, we approximate these signals with Gabor atoms:

$$g(x) = \frac{1}{K} \exp\left(-\pi\left(\frac{x-b}{s}\right)^2\right) \sin\left(2\pi \frac{w}{N}(x-b)\right), \tag{A.29}$$

where K is a normalization constant. Atoms are chosen from a dictionary, which is constructed by the discretization of parameters (s, b, w) that respectively represent scale, translation and frequency. The scales are discretized in a dyadic manner, i.e., $s = 2^j, j = 1, ..., log_2(N)$, where N is the signal length. Translation (shift) parameter b is chosen uniformly from 1 to N with step 2, such that the dictionary is overcomplete and its ISI is not too high. Finally, to construct the dictionary that is invariant to shifts and scales as given in Eq. (A.3), frequency has to be linked to the scale as: $w = w_0/s$, where w_0 is the basic modulating frequency and it is constant. We have chosen it to be $5N$, which is the approximate frequency of the given seismic signals. Seismic signals y_1 and y_2 are approximated by one Gabor atom per signal ($m = 1$), and the approximated signals are p_1 and p_2, respectively (see Figure A.4). The Gabor atoms recovered by independent thresholding on two signals have the following parameters: $s_1 = 128, s_2 = 128, b_1 = 571, b_2 = 553$, and the recovery conditions in Eq. (A.8) for the signal y_1 and Eq. (A.18) for the signal y_2 are shown to be satisfied. Therefore, the observed seismic signals are sparse in the chosen dictionary, correlated by the proposed model, and the derived recovery condition holds. Interestingly, the condition in Eq. (A.8) evaluated for the signal y_2 does not hold (false negative), thus giving evidence that our new condition in Eq. (A.18) is less conservative than the condition in Eq. (A.8). Moreover, the recovered atoms directly give us the shift between signals. The recovered shift is equal to the shift evaluated by the cross-correlation of original signals, thus it is correctly recovered.

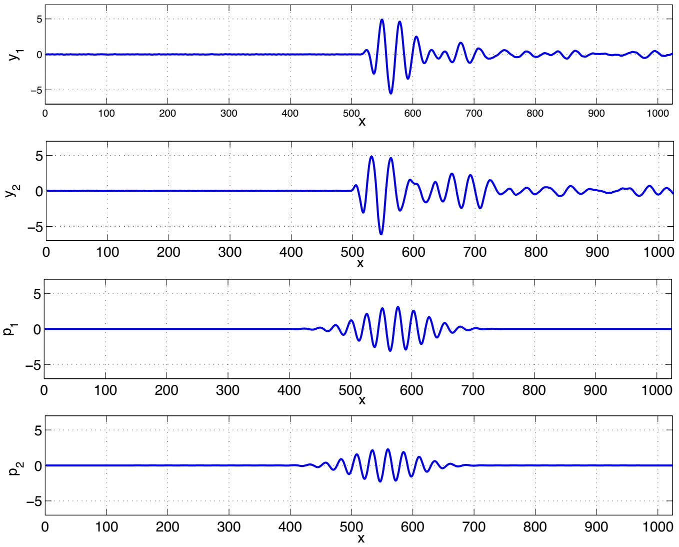


Figure A.4: *Seismic signals y_1 and y_2 captured at two neighboring locations and their respective approximations p_1 and p_2 with one Gabor atom per signal. The signals y_1 and y_2 are correlated by a shift on the x axis, which is correctly captured by shifted Gabor atoms.*

The distributed thresholding represent a low-complexity solution for sparse approximation of multi-view images in camera networks. The derived recovery condition permits us to identify the situations when complex sparse approximation algorithms, such as MP, can be replaced with the fast thresholding algorithm. This would result in significant energy savings in a camera network.

A.2 Derivation of the conditional probabilities $P(b_r|a_l, \phi_l, \psi_r)$ and $P(a_l|b_r, \phi_l, \psi_r)$ for the ML based stereo learning method

We first replace the expansions for y_L and y_R from Eq. (7.1) in Eq. (7.27), and get for all $k = 1, ..., m$:

$$\langle\sum_{i=1}^{m} b_{r_i}\psi_{r_i}, \psi_{r_k}\rangle + \langle e_R, \psi_{r_k}\rangle = \frac{1}{\sqrt{J_{l_k r_k}}}\langle\sum_{i=1}^{m} a_{l_i}\phi_{l_i}, \phi_{l_k}\rangle + \frac{1}{\sqrt{J_{l_k r_k}}}\langle e_L, \phi_{l_k}\rangle, \tag{A.30}$$

which can be rewritten as:

$$b_{r_k} + \sum_{\substack{i=1\\ i\neq k}}^{m}\langle b_{r_i}\psi_{r_i}, \psi_{r_k}\rangle + \langle e_R, \psi_{r_k}\rangle = \frac{1}{\sqrt{J_{l_k r_k}}} a_{l_k} + \frac{1}{\sqrt{J_{l_k r_k}}}\sum_{\substack{i=1\\ i\neq k}}^{m}\langle a_{l_i}\phi_{l_i}, \phi_{l_k}\rangle + \frac{1}{\sqrt{J_{l_k r_k}}}\langle e_L, \phi_{l_k}\rangle, \tag{A.31}$$

or simply:

$$b_{r_k} = \frac{a_{l_k}}{\sqrt{J_{l_k r_k}}} + \eta', \tag{A.32}$$

where:

$$\eta' = \frac{1}{\sqrt{J_{l_k r_k}}}\sum_{\substack{i=1\\ i\neq k}}^{m}\langle a_{l_i}\phi_{l_i}, \phi_{l_k}\rangle + \frac{1}{\sqrt{J_{l_k r_k}}}\langle e_L, \phi_{l_k}\rangle - \sum_{\substack{i=1\\ i\neq k}}^{m}\langle b_{r_i}\psi_{r_i}, \psi_{r_k}\rangle - langlee_R, \psi_{r_k}\rangle. \tag{A.33}$$

We will further assume that η' is a small value, since it is a sum of the projection of some noise to a chosen atom, and a linear combination of inner products of a chosen atom with other atoms in the image decomposition. When the image decomposition is sparse and the dictionary is overcomplete, the assumption is usually verified. However, we cannot use directly the expression in Eq. (A.34) to derive the distribution $P(\mathbf{a}, \mathbf{b}|\mathbf{\Phi}, \mathbf{\Psi})$ because the sparse support of the stereo images is not known, and hence also the indexes l_k, r_k and $k = 1, ..., m$. Therefore, we say that an arbitrary stereo atom pair ϕ_l, ψ_r and their coefficients a_l, a_r satisfy Eq. (A.34) up to a certain error η_1, which includes also η'. Therefore, we have:

$$b_r = \frac{1}{\sqrt{J_{lr}}} a_l + \eta_1, \tag{A.34}$$

where J_{lr} is the Jacobian of the linear transform of the coordinate system induced by the transform between atoms ϕ_l and ψ_r. When a_l and b_r are the coefficients of a stereo pair, then they satisfy Eq. (A.34) with a small value of the noise η_1. Otherwise, a_l and b_r are not significant in sparse decompositions of stereo images (according to the model in Eq. (7.1)) and hence the noise η_1 is also small. Therefore, we can model the noise η_1 with a white Gaussian noise of variance σ_b^2 and get:

$$P(\eta_1) = P(b_r|a_l, \phi_l, \psi_r) = \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2\right). \tag{A.35}$$

Although ϕ_l, ψ_r are not explicitly contained in the probability expression, they are implicitly there since J_{lr} is evaluated as a Jacobian of a transform between ϕ_l and ψ_r. Multiplying Eq. (A.34) with $\sqrt{J_{lr}}$, we can get a symmetric relation:

$$\begin{aligned} P(\eta_2) = P(a_l|b_r, \phi_l, \psi_r) &= \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_a^2}(a_l - \sqrt{J_{lr}} b_r)^2\right) \\ &= \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2 J_{lr}}(a_l - \sqrt{J_{lr}} b_r)^2\right) \\ &= \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2}(b_r - \frac{a_l}{\sqrt{J_{lr}}})^2\right), \end{aligned} \tag{A.36}$$

where we used the fact that variance of the noise $\eta_2 = \sqrt{J_{lr}}\eta_1$ can be evaluated as $\sigma_a = \sigma_b \sqrt{J_{lr}}$. Note that the same expression for the conditional probability $P(a_l|b_r, \phi_l, \psi_r)$ would be obtained if we consider the inverse transform F_{rl} from atom ψ_r to atom ϕ_l because the Jacobian of the linear transform satisfies: $J(Q_{lr}^{-1}) = 1/J_{lr}$.

A.3 Approximation rate of the MVMP algorithm

The MVMP algorithm in Chapter 7 is a modification of the baseline Matching Pursuit algorithm (or simply called Matching Pursuit [Mal93]). We examine here its approximation properties. At each iteration k of the Matching Pursuit (MP) the best atom is chosen from the dictionary $\mathcal{D} = \{\phi_n\}, n = 1, ..., N$ such that:

$$\phi_b^{(k)} = \arg\min_{\phi \in \mathcal{D}}(\|h_b^{(k-1)} - \langle h_b^{(k-1)}, \phi\rangle\phi\|^2) = \arg\min_{\phi \in \mathcal{D}}(\|h_b^{(k-1)}\|^2 - \langle h_b^{(k-1)}, \phi\rangle^2), \tag{A.37}$$

where $h_b^{(k-1)}$ is the signal residue from the previous iteration $k-1$. We have introduced a subscript b in the residue and the selected atoms to denote that they are specific to the MP. In the k^{th} iteration, the energy of the residue obtained with MP is equal to:

$$\|h_b^{(k)}\|^2 = \min_{\phi}(\|h_b^{(k-1)}\|^2 - \langle h_b^{(k-1)}, \phi\rangle^2). \tag{A.38}$$

Mallat and Zhang have shown that the residue in the MP algorithm decays exponentially, i.e., that there exists $\lambda > 0$ such that for all $k \geqslant 0$ it holds [Mal93]:

$$\|h_b^{(k)}\| \leqslant 2^{-\lambda}\|h_b^{(k-1)}\| \leqslant 2^{-\lambda k}\|h^{(0)}\|, \tag{A.39}$$

where $h^{(0)}$ is the initial residue equal to the signal that is approximated. Therefore, the approximation rate of MP is exponential.

However, MVMP differs from the MP because it belongs to a class of algorithms that we will call Constrained Matching Pursuit.

Definition A.3.1. *The constrained Matching Pursuit is a modified Matching Pursuit where at each iteration, the best atom is chosen according to the following criterion:*

$$\phi^{(k)} = \arg\min_{\phi}(\|h^{(k-1)} - \langle h^{(k-1)}, \phi\rangle\phi\|^2 + \rho A(\phi)) = \arg\min_{\phi}(\|h^{(k-1)}\|^2 - \langle h^{(k-1)}, \phi\rangle^2 + \rho A(\phi)) \tag{A.40}$$

Function $A(\phi)$ is a nonnegative constraining function, i.e., $A(\phi) \geqslant 0$, and ρ is the constraining factor. The residue is then updated in the same manner as in MP:

$$h^{(k)} = h^{(k-1)} - \langle h^{(k-1)}, \phi^{(k)}\rangle\phi^{(k)}. \tag{A.41}$$

For $\rho = 0$, the constrained MP is equivalent to the MP.

It is straightforward to see that the energy of the residue for the constrained MP is monotonically decreasing with k, since:

$$\|h^{(k)}\|^2 = \|h^{(k-1)}\|^2 - \langle h^{(k-1)}, \phi^{(k)}\rangle^2, \tag{A.42}$$

and the second term is always greater or equal to zero. However, Eq. (A.42) does not tell us anything about the residue energy decay or the approximation rate.

In the following theorem we derive the approximation bound for the constrained MP, and therefore for the MVMP as well.

Theorem A.3.1. *There exists $\lambda > 0$ such that for all $k \geqslant 0$, the energy of the residue in the constrained MP defined in Definition A.3.1 is bounded in the following way:*

$$\|h^{(k)}\|^2 \leqslant 2^{-2\lambda k}\|h^{(0)}\|^2 + \rho\bar{A}\frac{1 - 2^{-2\lambda(k+1)}}{1 - 2^{-2\lambda}}, \tag{A.43}$$

where $\bar{A} = \max_\phi A(\phi) - \min_\phi A(\phi)$ denotes the range of the constraining function $A(\phi)$.

Proof. We start the proof by relating the energies of the residues for the constrained MP and the MP, at iteration k. For the constrained MP we have that:

$$\|h^{(k-1)}\|^2 - \langle h^{(k-1)}, \phi^{(k)}\rangle^2 + \rho A(\phi^{(k)}) = \min_{\phi}(\|h^{(k-1)}\|^2 - \langle h^{(k-1)}, \phi\rangle^2 + \rho A(\phi)). \tag{A.44}$$

Moving $A(\phi^{(k)})$ to the right hand side we get:

$$\|h^{(k-1)}\|^2 - \langle h^{(k-1)}, \phi^{(k)}\rangle^2 = \min_{\phi}\left[\|h^{(k-1)}\|^2 - \langle h^{(k-1)}, \phi\rangle^2 + \rho A(\phi)) - \rho A(\phi^{(k)}\right]. \tag{A.45}$$

Using the triangle inequality we obtain:

$$\|h^{(k-1)}\|^2 - \langle h^{(k-1)}, \phi^{(k)}\rangle^2 \leqslant \min_{\phi}(\|h^{(k-1)}\|^2 - \langle h^{(k-1)}, \phi\rangle^2) + \rho\max_{\phi} A(\phi) - \rho\min_{\phi} A(\phi). \tag{A.46}$$

We can see that the first term on the right hand side is equal to the energy of the residue when computed by MP, at iteration k. Therefore, we have:

$$\|h^{(k)}\|^2 \leqslant \|h_b^{(k)}\|^2 + \rho\bar{A}. \tag{A.47}$$

The bound in Eq. (A.43) can be now obtained by induction. For $k = 1$, we have:

$$\|h^{(1)}\|^2 \leqslant \|h_b^{(1)}\|^2 + \rho\bar{A} \leqslant 2^{-2\lambda}\|h^{(0)}\|^2 + \rho\bar{A}, \tag{A.48}$$

where the second inequality follows from Eq. (A.39). Similarly, for $k = 2$, we have:

$$\|h^{(2)}\|^2 \leqslant \|h_b^{(2)}\|^2 + \rho\bar{A} \leqslant 2^{-2\lambda}\|h^{(1)}\|^2 + \rho\bar{A}. \tag{A.49}$$

Substituting $h^{(1)}$ from Eq. (A.48) into Eq. (A.49), we have:

$$\|h^{(2)}\|^2 \leqslant 2^{-4\lambda}\|h^{(0)}\|^2 + \rho\bar{A}2^{-2\lambda} + \rho\bar{A}. \tag{A.50}$$

Therefore, at iteration k, we finally obtain:

$$\|h^{(k)}\|^2 \leqslant 2^{-2\lambda k}\|h^{(0)}\|^2 + \rho\bar{A}\sum_{i=0}^{k} 2^{-2\lambda i} = 2^{-2\lambda k}\|h^{(0)}\|^2 + \rho\bar{A}\frac{1 - 2^{-2\lambda(k+1)}}{1 - 2^{-2\lambda}}. \tag{A.51}$$

□

The approximation rate of the MVMP is thus lower compared to the approximation rate of the MP, for $\rho > 0$ and $\bar{A} > 0$. The penalty on the approximation rate due to the constraining of MP depends on the constraining factor ρ. Knowing the range of the constraining function A, one can use the bound in Eq. (A.43) to choose ρ in order to reach the desired energy decay.

Bibliography

[Ada01] Adams M. The JPEG-2000 Still Image Compression Standard. Technical report, ISO/IEC JTC1/SC29/WG1, 2001.

[Ade91] Adelson E. H. and Bergen J. R. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*. edited by Landy M. and Movshon J. A., MIT Press, 1991.

[Aha06] Aharon M., Elad M. and Bruckstein A. K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Transactions on Signal Processing*, 54(11):4311– 4322, 2006.

[All01] Alliez P. and Desbrun M. Progressive Compression for Lossless Transmission of Triangle Meshes. In *Proceedings of ACM SIGGRAPH*, 2001.

[Ana06] Anantrasirichai N., Canagarajah C. N., Redmill D. W. and Bull D. R. Dynamic Programming for Multi-View Disparity/Depth Estimation. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.

[Ant98] Antoine J. and Vandergheynst P. Wavelets on the n-sphere and related manifolds. *Journal of Mathematical Physics*, 39(8):3987–4008, 1998.

[Ant99] Antoine J. and Vandergheynst P. Wavelets on the 2-sphere : a group theoretical approach. *Applied and Computational Harmonic Analysis*, 7(3):262–291, 1999.

[Ant02a] Antoine J., Demanet L., Jacques L. and Vandergheynst P. Wavelets on the sphere : Implementation and approximations. *Applied and Computational Harmonic Analysis*, 13(3):177–200, 2002.

[Ant02b] Antone M. and Teller S. Scalable Extrinsic Calibration of Omni-Directional Image Networks. *International Journal of Computer Vision*, 49(2-3):143–174, 2002.

[Asp02] Aspert N., Santa-Cruz D. and Ebrahimi T. Mesh: Measuring errors between surfaces using the hausdorff distance. In *Proceedings of the IEEE International Conference in Multimedia and Expo*, 2002.

[Bak99] Baker S. and Nayar S. K. A Theory of Single-Viewpoint Catadioptric Image Formation. *International Journal of Computer Vision*, 35(2):1–22, 1999.

[Bak03] Bakstein H., Pajdla T. and Večerka D. Rendering Almost Perspective Views from a Sparse Set of Omnidirectional Images. In *Proceedings of the British Machine Vision Conference*, 2003.

[Bar05] Barreto J. and Araujo H. Geometric properties of central catadioptric line images and their application in calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1327 – 1333, 2005.

[Bau06] Bauermann I., Mielke M. and Steinbach E. H. 264 Based Coding of Omnidirectional Video. In *Proceedings of the International Conference on Computer Vision and Graphics*, 2006.

[Bog04] Bogdanova I., Vandergheynst P., Antoine J., Jacques L. and Morvidone M. Discrete Wavelet Frames on the Sphere. In *Proceedings of the European Signal Processing Conference*, 2004.

[Bog05] Bogdanova I., Vandergheynst P., Antoine J., Jacques L. and Morvidone M. Stereographic Wavelet Frames on the Sphere. *Applied and Computational Harmonic Analysis*, 19(2):223–252, 2005.

[Bou04] Boult T. E., Gao X., Micheals R. and Eckmann M. Omni-directional visual surveillance. *Image and Vision Computing*, 22(7):515–534, 2004.

[Bur83] Burt P. J. and Adelson E. H. The Laplacian Pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[Cad08] Cadieu C. and Olshausen B. A. Learning Transformational Invariants from Time-Varying Natural Images. In *Proceedings of the Conference on Neural Information Processing Systems*, 2008.

[Can99] Candès E. J. and Donoho D. L. Curvelets – A surprisingly effective nonadaptive representation for objects with edges. In *Curves and Surfaces.* edited by Rabut C., Cohen A., and Schumaker L. L., Vanderbilt University Press, Nashville, TN, 1999.

[Che99] Chen S., Donoho D. and Saunders M. Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.

[Che07] Cheung N. and Ortega A. Flexible video decoding: A distributed source coding approach. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, 2007.

[Coh01] Cohen-Steiner D. and Da F. A Greedy Delaunay Based Surface Reconstruction Algorithm. ECG Technical Report ECG-TR-124202-01, INRIA, 2001.

[Cov91] Cover T. M. and Thomas J. A. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

[Cul07] Culpepper B. J. Learning 'what' and 'where' from movies. Master's thesis, UC Berkeley, 2007.

[Dav97] Davis G., Mallat S. and Avellaneda M. Greedy adaptive approximation. *Journal on Constructive Approximation*, 13(1):57–98, 1997.

[Dee95] Deering M. Geometric Compression. In *Proceedings of ACM SIGGRAPH*, 1995.

[Do 01] Do M. N., Dragotti P. L., Shukla R. and Vetterli M. On the compression of two dimensional piecewise smooth functions. In *Proceedings of the IEEE International Conference on Image Processing*, 2001.

[Do 05] Do M. N. and Vetterli M. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106, 2005.

[Dri94] Driscoll J. R. and Healy D. M. Computing Fourier Transform and Convolutions on the 2-Sphere. *Advances in Applied Mathematics*, 15(2):202–250, 1994.

[Dua05] Duarte M. F., Sarvotham S., Baron D., Wakin M. B. and Baraniuk R. G. Distributed Compressed Sensing of Jointly Sparse Signals. In *Proceedings of Asilomar Conference on Signals, Systems and Computers*, 2005.

[Eng99a] Engan K., Aase S. and Hakon Husoy J. Method of optimal directions for frame design. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1999.

[Eng99b] Engan K., Rao B. D. and Kreutz-Delgado K. Frame design using FOCUSS with method of optimal directions (MOD). In *Proceedings of the Norwegian Signal Processing Symposium*, 1999.

[Eng07] Engan K., Skretting K. and Husřy J. Family of iterative LS-based dictionary learning algorithms, ILS-DLA, for sparse signal representation. *Digital Signal Processing*, 17(1):32–49, 2007.

[Fig02] Figueras i Ventura R. M., Granai L. and Vandergheynst P. R-D analysis of adaptive edge representations. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, 2002.

[Fig06] Figueras i Ventura R. M., Vandergheynst P. and Frossard P. Low rate and flexible image coding with redundant representations. *IEEE Transactions on Image Processing*, 15(3):726–739, 2006.

[Fis81] Fischler M. and Bolles R. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[Fli06] Flierl M. and Vandergheynst P. Distributed Coding of Highly Correlated Image Sequences with Motion-Compensated Temporal Wavelets. *EURASIP Journal on Applied Signal Processing*, Article ID 46747:10 pages, 2006.

[Fro00] Frossard P. *Robust and Multiresolution Video Delivery: From H.26x to Matching Pursuit Based Technologies.* Phd thesis, EPFL, 2000.

[Fro04] Frossard P., Vandergheynst P., Figueras i Ventura R. M. and Kunt M. A posteriori quantization of progressive matching pursuit streams. *IEEE Transactions on Signal Processing*, 52(2):525–535, 2004.

[Fuc04] Fuchs J.-J. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50(6):1341–1344, 2004.

[Fuj04] Fujii T. and Tanimoto M. Free-Viewpoint TV (FTV) System. In *Proceedings of the Pacific Rim Conference on Multimedia*, 2004.

[Gar06] Garbas J.-U., Fecker U., Troger T. and Kaup A. 4D Scalable Multi-View Video Coding Using Disparity Compensated View Filtering and Motion Compensated Temporal Filtering. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, 2006.

[Geh05] Gehrig N. and Dragotti P. L. DIFFERENT - distributed and fully flexible image encoders for camera sensor networks. In *Proceedings of the IEEE International Conference on Image Processing*, 2005.

[Geh07] Gehrig N. and Dragotti P. L. Distributed Compression of Multi-View Images using a Geometrical Coding Approach. In *Proceedings of the IEEE International Conference on Image Processing*, 2007.

[Geh09] Gehrig N. and Dragotti P. L. Geometry-Driven Distributed Compression of the Plenoptic Function: Performance Bounds and Constructive Algorithms. *IEEE Transactions on Image Processing*, 18(3):457– 470, 2009.

[Ger92] Gersho A. and Gray R. M. *Vector Quantization and Signal Compression.* Springer, 1992.

[Gey01] Geyer C. and Daniilidis K. Catadioptric Projective Geometry. *International Journal of Computer Vision*, 45(3):223 – 243, 2001.

[Gey02] Geyer C. and Daniilidis K. Paracatadioptric camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):687–695, 2002.

[Gir05] Girod B., Aaron A., Rane S. and Rebollo-Monedero D. Distributed video coding. *Proceedings of the IEEE*, 93(1):71 – 83, 2005.

[Gor97] Gorodnitsky I. and Rao B. Sparse Signal Reconstruction from Limited Data Using FOCUSS: a Re-weighted Minimum Norm Algorithm. *IEEE Transactions on Signal Processing*, 45(3):600–616, 1997.

[Gri03] Gribonval R. and Nielsen M. Sparse representations in unions of bases. *IEEE Transactions on Information Theory*, 49(12):3320–3325, 2003.

[Gri06] Gribonval R., Nielsen M. and Vandergheynst P. Towards an adaptive computational strategy for sparse signal approximation. *preprint of the Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA)*, 2006.

[Gri08a] Gribonval R. and Nielsen M. Beyond sparsity: Recovering structured representations by l1 minimization and greedy algorithms. *Advances in computational mathematics*, 28(1):23–41, 2008.

[Gri08b] Gribonval R., Rauhut H., Schnass K. and Vandergheynst P. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *Journal of Fourier Analysis and Applications*, 14(5):655–687, 2008.

[Gui07] Guillemot C., Pereira F., Torres L., Ebrahimi T., Leonardi, R. and Ostermann, J. Distributed Monoview and Multiview Video Coding. *IEEE Signal Processing Magazine*, 24(5):67–76, 2007.

[Guo04] Guo X. and Huang Q. Multiview Video Coding Based on Global Motion Model. In *Proceedings of the Pacific Rim Conference on Multimedia*, 2004.

[Guo06] Guo X., Lu Y., Wu F., Gao W. and Li S. Distributed multi-view video coding. In *Proceedings of the SPIE - Visual Communication and Image Processing*, 2006.

[Har97] Hartley R. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):580 – 593, 1997.

[Har00] Hartley R.I. and Zisserman A. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[Hea03] Healy Jr. D., Rockmore D., Kostelec P. and Moore S. FFTs for the 2-Sphere - Improvements and Variations. *Journal of Fourier Analysis and Applications*, 9(4):341 – 385, 2003.

[Hec97] Hecht E. and Zajac A. *Optics*. Addison-Wesley: Reading, MA, 1997.

[Hop92] Hoppe H., DeRose T., Duchamp T., McDonald J. and Stuetzle W. Surface reconstruction from unorganized points. In *Proceedings of ACM SIGGRAPH*, 1992.

[Hop96] Hoppe H. Progressive meshes. In *Proceedings of ACM SIGGRAPH*, 1996.

[Hop03] Hoppe H. and Praun E. Shape compression using spherical geometry images. In *Proceedings of the Symposium on Multiresolution in Geometric Modeling*, 2003.

[Hoy00] Hoyer P. and Hyvärinen A. Independent component analysis applied to feature extraction from colour and stereo images. *Network: Computation in Neural Systems*, 11(3):191–210, 2000.

[Jon92a] Jones D. and Malik J. A Computational Framework for Determining Stereo Correspondence from a Set of Linear Spatial Filters. In *Proceedings of the European Conference on Computer Vision*, 1992.

[Jon92b] Jones D. and Malik J. Determining Three-Dimensional Shape from Orientation and Spatial Frequency Disparities. In *Proceedings of the European Conference on Computer Vision*, 1992.

[Jos06a] Jost P., Vandergheynst P. and Frossard P. Tree-Based Pursuit: Algorithm and Properties. *IEEE Transactions on Signal Processing*, 54(12):4685 – 4697, 2006.

[Jos06b] Jost P., Vandergheynst P., Lesage S. and Gribonval R. MoTIF: an efficient algorithm for learning translation invariant dictionaries. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2006.

[Kan00] Kang S. B. Catadioptric Self-Calibration. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2000.

[Kar00] Karni Z. and Gotsman C. Spectral compression of mesh geometry. In *Proceedings of ACM SIGGRAPH*, pages 279–286, 2000.

[Kho00] Khodakovsky A., Schröder P. and Sweldens W. Progressive Geometry Compression. In *Proceedings of ACM SIGGRAPH*, 2000.

[Kol02] Kolmogorov V. and Zabih R. Multi-camera scene reconstruction via graph cuts. In *Proceedings of the European Conference on Computer Vision*, 2002.

[Kon00] Konrad J. and Lan Z.-D. Dense disparity estimation from feature correspondences. In *Proceedings of SPIE Symposium on Electronic Imaging*, 2000.

[Kos04] Košeckà J. and Yang X. Global localization and relative pose estimation based on scale-invariant features. In *Proceedings of the International Conference on Pattern Recognition*, 2004.

[Kos08] Kostelec P. and Rockmore D. FFTs on the Rotation Group. *Journal of Fourier Analysis and Applications*, 14(2):145–179, 2008.

[Kre03] Kreutz-Delgado K., Murray J., Rao B., Engan K., Lee T.-W. and Sejnowski T. J. Dictionary Learning Algorithms for Sparse Representation. *Neural Computation*, 15(2):349–396, 2003.

[Kub07] Kubota A., Smolic A., Magnor M., Tanimoto M., Chen T. and Zhang C. Multiview Imaging and 3DTV. *IEEE Signal Processing Magazine*, 24(6):10–21, 2007.

[LDP] Methods for constructing LDPC codes: Available in URL http://www.cs.utoronto.ca/pub/radford/LDPC-2001-05-04/pchk.html.

[Liu06] Liu J. and Hubbold R. Automatic Camera Calibration and Scene Reconstruction with Scale-Invariant Features. In *International Symposium on Visual Computing*, 2006.

[Llo08a] Llonch R. S., Kokiopoulou E., Tošić I. and Frossard P. 3D Face Recognition using Sparse Spherical Representations. In *Proceedings of the International Conference on Pattern Recognition*, 2008.

[Llo08b] Llonch R. S., Kokiopoulou E., Tošić I. and Frossard P. 3D Face Recognition with Sparse Spherical Representations. *submitted to Pattern Recognition*, 2008.

[Low04] Lowe D. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[Ma 04] Ma Y., Soatto S., Košeckà J. and Sastry S. S. *An Invitation to 3-D Vision: From Images to Geometric Models.* Springer, 2004.

[Mak03] Makadia A. and Daniilidis K. Direct 3D-rotation estimation from spherical images via a generalized shift theorem. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2003.

[Mak07] Makadia A., Geyer C. and Daniilidis K. Correspondenceless Structure from Motion. *International Journal of Computer Vision*, 75(3):311–327, 2007.

[Mal93] Mallat S. G. and Zhang Z. Matching Pursuits With Time-Frequency Dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[Mal08] Mallat S. and Peyré G. *A Wavelet Tour of Signal Processing: The Sparse Way.* Academic Press, 2008.

[Mar06a] Marpe D., Wiegand T. and Sullivan G. The H. 264/MPEG4 advanced video coding standard and its applications. *IEEE Communications Magazine*, 44(8):134–143, 2006.

[Mar06b] Martinian E., Behrens A., Xin J., Vetro A. and Sun H. Extensions of H. 264/AVC for multiview video compression. In *Proceedings of the IEEE International Conference on Image Processing*, 2006.

[Mat04] Matusik W. and Pfister H. 3D TV: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Transactions on Graphics*, 23:814–824, 2004.

[Mer06] Merkle P., Müller K., Smolic A. and Wiegand T. Efficient Compression of Multi-view Video Exploiting Inter-view Dependencies Based on H.264/MPEG4-AVC. In *Proceedings of the IEEE International Conference in Multimedia and Expo*, 2006.

[Mer07] Merkle P., Smolic A., Müller K. and Wiegand T. Efficient Prediction Structures for Multiview Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11):1461 – 1473, 2007.

[Mic04] Mičušík B. and Pajdla T. Autocalibration & 3D reconstruction with non-central catadioptric cameras. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

[Mon08a] Monaci G., Sommer F. and Vandergheynst P. Learning sparse generative models of audiovisual signals. In *Proceedings of the European Signal Processing Conference*, 2008.

[Mon08b] Monaci G., Sommer F. and Vandergheynst P. Learning sparse generative models of audiovisual signals. *submitted to IEEE Transactions on Neural Networks*, 2008.

[Nat95] Natarajan B. K. Sparse Approximate Solutions to Linear Systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

[Nay97a] Nayar S. Catadioptric omnidirectional camera. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 482–488, 1997.

[Nay97b] Nayar S. K. Omnidirectional vision. In *Proceedings of the International Symposium of Robotics Research*, 1997.

[Nea98] Neal R. M. and Hinton G. E. *A view of the EM algorithm that justifies incremental, sparse, and other variants*, pages 355–368. in *Learning in Graphical Models*, edited by M. I. Jordan, Dordrecht: Kluwer Academic Publishers, 1998.

[Nef97] Neff R. and Zakhor A. Very Low Bit-Rate Video Coding based on Matching Pursuits. *IEEE Transactions on Circuits and Systems for Video Technology*, 7(1):158–171, 1997.

[Nei97] Neider J., Davis T. and Woo M. OpenGL programming guide, 1997.

[Nen98] Nene S. and Nayar S. K. Stereo with mirrors. In *Proceedings of the International Conference on Computer Vision*, pages 1087–1094, 1998.

[Oka04] Okajima K. Binocular disparity encoding cells generated through an Infomax based learning algorithm. *Neural Networks*, 17(7):953–962, 2004.

[Ols96] Olshausen B. A. and Field D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. Springer-Verlag New YORK INC.

[Ols97] Olshausen B. and Field D. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–25, 1997.

[Ols03] Olshausen B. A. Learning sparse, overcomplete representations of time-varying natural images. In *Proceedings of the IEEE International Conference on Image Processing*, 2003.

[Ols07] Olshausen B. A., Cadieu C., Culpepper B. J., and Warland D. K. Bilinear Models of Natural Images. In *Proceedings of the SPIE Conference on Human Vision and Electronic Imaging*, 2007.

[Oua06a] Ouaret M., Dufaux F. and Ebrahimi T. Fusion-based multiview distributed video coding. In *Proceedings of ACM International Workshop on Video Surveillance and Sensor Networks*, pages 139 – 144, 2006.

[Oua06b] Ouaret M., Dufaux F. and Ebrahimi T. Fusion-based multiview distributed video coding. In *Proceedings of the 4th ACM international workshop on Video surveillance and sensor networks*, 2006.

[Pra03] Pradhan S. S. and Ramchandran K. Distributed source coding using syndromes (DISCUS). *IEEE Transactions on Information Theory*, 49(3):626–643, 2003.

[Pur07] Puri R. and Ramchandran K. PRISM: A Video Coding Paradigm With Motion Estimation at the Decoder. *IEEE Transactions on Image Processing*, 16(10):2436–2448, 2007.

[Qia04] Qian G., Chellappa R. and Zheng Q. Robust Bayesian cameras motion estimation using random sampling. In *Proceedings of the IEEE International Conference on Image Processing*, 2004.

[Rah06] Rahmoune A., Vandergheynst P. and Frossard P. Flexible Motion-Adaptive Video Coding with Redundant Expansions. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(2):178–190, 2006.

[Red08] Reddy D., Sankaranarayanan A. C., Cevher V. and Chellappa R. Compressed Sensing for Multi-View Tracking and 3-D Voxel Reconstruction. In *Proceedings of the IEEE International Conference on Image Processing*, 2008.

[Roz08] Rozell C. J., Johnson D. H., Baraniuk R. G. and Olshausen B. A. Sparse coding via thresholding and local competition in neural circuits. *Neural Computation*, 20(10):2526–2563, 2008.

[Sai96] Said A. and Pearlman W. A New, Fast, and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(3):243 – 250, 1996.

[Sca06] Scaramuzza D., Martinelli A. and Siegwart R. A Flexible Technique for Accurate Omnidirectional Camera Calibration and Structure from Motion. In *Proceedings of the IEEE International Conference on Computer Vision Systems*, 2006.

[Sch95] Schröder P. and Sweldens W. Spherical Wavelets: Efficiently Representing Functions on the Sphere. In *Proceedings of ACM SIGGRAPH*, 1995.

[Sch01] Schmid-Saugeon P. and Zakhor A. Learning dictionaries for matching pursuits based video coders. In *Proceedings of the IEEE International Conference on Image Processing*, 2001.

[Sch04] Schmid-Saugeon P. and Zakhor A. Dictionary design for matching pursuit and application to motion-compensated video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(6):880– 886, 2004.

[Sch06] Schnass K., Vandergheynst P. and Frossard P. Distributed Sensing of Noisy Signals by Thresholding of Redundant Expansions. In *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, 2006.

[Se 01] Se S., Lowe D. and Little J. Vision-based mobile robot localization and mapping using scale-invariant features. In *Proceedings of the IEEE International Conference on Robotics and Automation*, 2001.

[Seh04] Sehgal A., Jagmohan A. and Ahuja N. Wyner-Ziv coding of video: An error-resilient compression framework. *IEEE Transactions on Multimedia*, 6(2):249–258, 2004.

[She06] Shen L. and Makedon F. Spherical mapping for processing of 3D closed surfaces. *Image and Vision Computing*, 24(7):743–761, 2006.

[Sil91] Silicon Graphics Inc. GL programming guide, 1991.

[Sle73] Slepian D. and Wolf J. K. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, 1973.

[Smo06] Smolic A., Mueller K., Merkle P., Fehn C., Kauff P., Eisert P. and Wiegand T. 3D Video and Free Viewpoint Video–Technologies, Applications and MPEG Standards. In *Proceedings of the IEEE International Conference in Multimedia and Expo*, 2006.

[Smo07] Smolic A., Mueller K., Stefanoski N., Ostermann J., Gotchev, A., Akar G. B., Triantafyllidis G. and Koz A. Coding Algorithms for 3DTV - A Survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(11):1606–1621, 2007.

[Son07] Song B., Tuncel E. and Roy-Chowdhury A. K. Towards A Multi-Terminal Video Compression Algorithm By Integrating Distributed Source Coding With Geometrical Constraints. *Journal Of Multimedia*, 2(3):9–16, 2007.

[Sta03] Starck J., Elad M. and Donoho D. Image decomposition: Separation of texture from piecewise smooth content. In *Proceedings of the SPIE Conference on Wavelets: Applications in Signal and Image Processing*, 2003.

[Stu05] Sturm, P. Multi-view geometry for general camera models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2005.

[Sun03] Sun J., Zheng N.-N. and Shum H.-Y. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7):787– 800, 2003.

[Svo98] Svoboda T., Pajdla T. and Hlaváč V. Epipolar Geometry for Panoramic Cameras. In *Proceedings of the European Conference on Computer Vision*, pages 218–231, 1998.

[Sze99] Szeliski R. A multi-view approach to motion and stereo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1999.

[Tau98a] Taubin G. and Rossignac J. Geometry compression through Topological surgery. *ACM Transactions on Graphics*, 17(2):84–115, 1998.

[Tau98b] Taubin G., Guéziec A. and Horn W. Progressive forest split compression. In *Proceedings of ACM SIGGRAPH*, 1998.

[Tau01] Taubman D. S. and Marcellin M. W. *JPEG2000: Image Compression Fundamentals, Standards, and Practice*. Kluwer Academic Publishers, 2001.

[Tem00] Temlyakov V. N. Weak greedy algorithms. *Advances in Computational Mathematics*, 12(2-3):213–227, 2000.

[Thi09] Thirumalai V. and Frossard P. Bit Rate Allocation for Disparity Estimation from Compressed Images. *Proceedings of the Picture Coding Symposium*, 2009.

[Tor93] Torr P. H. S. and Murray D. W. Outlier detection and motion segmentation. In *Proceedings of the SPIE Conference on Sensor Fusion*, 1993.

[Tor97] Torr P. H. S. and Murray D. W. The Development and Comparison of Robust Methods for Estimating the Fundamental Matrix. *International Journal of Computer Vision*, 24(3):271–300, 1997.

[Tor05] Torii A., Imiya A. and Ohnishi N. Two- and Three- View Geometry for Spherical Cameras. In *Proceedings of the 6th Workshop on Omnidirectional Vision, Camera Networks and Non-classical cameras*, 2005.

[Tos05] Tošić I., Bogdanova I., Frossard P. and Vandergheynst P. Multiresolution Motion Estimation for Omnidirectional Images. In *Proceedings of the European Signal Processing Conference*, 2005.

[Tos06] Tošić I., Frossard P. and Vandergheynst P. Progressive Coding of 3-D Objects Based on Overcomplete Decompositions. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(11):1338–1349, 2006.

[Tos09] Tošić I. and Frossard P. *Spherical imaging in omni-directional camera networks.* in *Multi-Camera Networks, Principles and Applications*, edited by Aghajan H. and Cavallaro A., Academic press, 2009.

[Tou98] Touma C. and Gotsman C. Triangle Mesh Compression. In *Proceedings of the Graphics Interface Conference*, 1998.

[Tro04] Tropp J. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[Tro06] Tropp J. A. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

[Van01] Vandergheynst P. and Frossard P. Efficient Image Representation By Anisotropic Refinement In Matching Pursuit. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2001.

[Vel06] Velisavljević V., Beferull-Lozano B., Vetterli M. and Dragotti P. L. Directionlets: anisotropic multidirectional representation with separable filtering. *IEEE Transactions on Image Processing*, 15(7):1916 – 1933, 2006.

[Vie06] Vielva P., Wiaux Y., Martinez-Gonzalez E. and Vandergheynst P. Steerable wavelet analysis of CMB structures alignment. *New Astronomy Reviews*, 50(11-12):880–888, 2006.

[Wag03] Wagner R., Nowak R. and Baraniuk R. Distributed image compression for sensor networks using correspondence analysis and super-resolution. In *Proceedings of the IEEE International Conference on Image Processing*, 2003.

[Wia08] Wiaux Y., McEwen J. D., Vandergheynst P. and Blanc O. Exact reconstruction with directional wavelets on the sphere. *Monthly Notices of the Royal Astronomical Society*, 388(2):770–788, 2008.

[Wip04] Wipf D. P. and Rao B. D. Sparse Bayesian learning for basis selection. *IEEE Transactions on Signal Processing*, 52(8):2153–2164, 2004.

[Wit87] Witten I. H., Neal R. M. and Cleary J. G. Artihmetic Coding for Data Compression. *Communications of the ACM*, 30(6):520–540, 1987.

[Wyn74] Wyner A. Recent results in the Shannon theory. *IEEE Transactions on Information Theory*, IT-m(1), 1974.

[Wyn76] Wyner A. D. and Ziv J. The rate-distortion function for source coding with side-information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, 1976.

[Yag95] Yagi Y., Nishizawa Y. and Yachida M. Map-based navigation for a mobile robot with omnidirectional image sensor COPIS. *IEEE Transactions on Robotics and Automation*, 11(5):634–648, 1995.

[Yag99] Yagi Y. Omnidirectional sensing and its applications. *IEICE Transactions On Information And Systems*, E82-D(3):568–579, 1999.

[Yan06] Yang W., Lu Y., Wu F., Cai J., Ngan K. and Li S. 4-D Wavelet-Based Multiview Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(11):1385–1396, 2006.

[Yan07a] Yang Y., Stanković V., Zhao W. and Xiong Z. Multiterminal video coding. In *Proceedings of IEEE Workshop on Information Theory and its Applications*, 2007.

[Yan07b] Yang Y., Stanković V., Zhao W. and Xiong Z. Multiterminal video coding. In *Proceedings of the IEEE International Conference on Image Processing*, 2007.

[Yeo07] Yeo C. and Ramchandran K. Robust distributed multiview video compression for wireless camera networks. In *Proceedings of the SPIE - Visual Communication and Image Processing*, 2007.

[Yin04] Ying X. and Hu Z. Catadioptric camera calibration using geometric invariants. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1260 – 1271, 2004.

[Zhu03] Zhu X., Aaron A. and Girod B. Distributed compression for large camera arrays. In *Proceedings of the IEEE Workshop on Statistical Signal Processing*, 2003.

Curriculum Vitae

CONTACT

Ivana Tošić
Address: EPFL STI IEL LTS4, Station 11, Lausanne, CH 1015
Tel: 41 21 6934712
email: ivana.tosic@epfl.ch
webpage: http://lts4www.epfl.ch/~tosic/

EDUCATION

- **Ph.D. program in Computer and Communication Sciences** 2004-Present
  Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland
  School of Computer and Communication Sciences
  Research domain: Signal processing and image communications
- **Graduate program in Computer and Communication Sciences** 2003-2004
  Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland
  School of Computer and Communication Sciences
  Orientation: multimedia technologies
- **Dipl.Ing. degree in Electrical Engineering** 1997-2003
  University of Niš, Serbia
  Faculty of Electronic Engineering/Department of telecommunications
  GPA 9.80/10, best graduate award

WORK EXPERIENCE

- **Research and Teaching assistant** September 2004 - Present
  Swiss Federal Institute of Technology in Lausanne (EPFL), Switzerland
  Institute of Electrical Engineering, Signal Processing Laboratory LTS4
  Research topics: Sparse representations, Multi-view imaging,
  Data compression, Wyner-Ziv coding, Omnidirectional imaging
- **Research visit** January 2008 - April 2008
  University of California, Berkeley, US
  Redwood Center for Theoretical Neuroscience
  Research topic: unsupervised dictionary learning
- **Student internship** October 2001 - December 2001
  Mataró School of Engineering (EUPMT), Spain
  Project name: Mammogram Image Analysis in Diagnosing Breast Cancer

RESEARCH PROJECTS

- Developed a method for learning overcomplete dictionaries for representing stereo images (during the research stay at UC Berkeley)
- Developed an occlusion-resilient distributed coder for omnidirectional camera networks based on a novel geometric multi-view correlation model
- Built a low rate and robust coarse scene geometry and camera pose estimator
- Designed and implemented a 3D objects coder based on the Spherical Matching Pursuit and the dictionary on the SO(3) group. Spherical shape representation exploited additionally for 3D face recognition.
- Developed an interpolation method for spherical signals
- Developed a multiresolution motion estimator for correlated omnidirectional images
- Developed a mammogram image analysis method for diagnosing breast cancer (at EUPMT, Mataró, Spain)

TEACHING EXPERIENCE

- Teaching Assistant for the doctoral course *Advanced digital image processing*, EPFL, fall 2004, fall 2005
- Teaching Assistant for the master course *Image Communications*, EPFL, spring 2006
- Teaching Assistant for the bachelor course *Digital Signal Processing (Traitement numérique des signaux)*, teaching in french, EPFL, spring 2009
- Supervisor of master theses for three M.Sc. students at EPFL
- Gave lectures on Distributed coding within the *Image Communications* course, EPFL, spring 2007, 2008

AWARDS

- **IBM T.J. Watson Research Center award** 2008
  "Emerging Leaders in Multimedia 2008", top 12 students world-wide

- **EPFL Fellowship for the doctoral studies** 2003
  for the academic year 2003-2004

- **The Certificate of Recognition of the University of Niš** 2003
  Faculty of Electronic Engineering Best Graduate Award

- **Royal Family Karadjordjevic Award** 2002
  for the 100 best students in Serbia

- **Norway Government Award** 2001
  for the 1000 best students in Serbia

PROFESSIONAL ACTIVITIES

- Finance Chair for the 16th Packet Video Workshop, Lausanne, 2007.
- Technical Program Committee member for: ICME 2007, MMM 2009, 3DTV-CON 2009
- Reviewer for journals: IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for video technology, IEEE Transactions on Multimedia, EURASIP Journal on Advances in Signal Processing, Computer Vision and Image Understanding, EURASIP Journal on Image and Video Processing
- Reviewer for conferences: ICIP, ICASSP, EUSIPCO, ICME, MMSP, ISCAS, Eurographics, MMM, 3DTV-CON, GlobeCom, PCS

INVITED TALKS

- Invited talks at international conferences
    1. Wyner-Ziv coding of multi-view omnidirectional images with overcomplete decompositions, joint work with P. Frossard, invited at the IEEE International Conference on Image Processing, Atlanta (US), 2007
    2. Coarse scene geometry estimation from sparse approximations of multi-view omnidirectional images, joint work with P. Frossard, invited at the European Signal Processing Conference, Poznan (Poland), 2007
    3. A Geometrical Framework for Distributed Coding of Multiview Images, joint work with P. Frossard, invited at the DISCOVER workshop, Lisbon (Portugal), 2007
    4. Balanced Distributed Coding of Omnidirectional Images, joint work with V.Thirumalai and P. Frossard, invited at the SPIE - Visual Communication and Image Processing Conference, San Jose (US), 2008
    5. Distributed coding in camera networks with learned dictionaries, joint work with P. Frossard, invited to MobiMedia, International Mobile Multimedia Communications Conference, London (UK), 2009
    6. Stereo dictionary learning for multiview scene representation, joint work with P. Frossard, invited to APSIPA Annual Summit and Conference, Sapporo (Japan), 2009
- Invited talks at institutes
    1. Geometry-based distributed coding of multi-view images with sparse approximations, joint work with P. Frossard, invited at the Net/Comm/DSP Seminar Series, UC Berkeley (US), 2008
    2. Coarse scene geometry estimation from multi-view omnidirectional images, joint work with P. Frossard, invited at the Heterogeneous Sensor Networks (HSN) Seminar Series, UC Berkeley (US), 2008
    3. Geometry-based distributed coding of multi-view images with sparse approximations, joint work with P. Frossard, invited at the Emerging Leaders in Multimedia workshop, IBM Watson Research Center (US), 2008

LANGUAGES

- **English:** fluent
- **French:** fluent
- **Spanish:** basic
- **Serbian:** mother tongue

Personal Publications

- **Book Chapter**

  I. Tošić and P. Frossard, Spherical Imaging in Omnidirectional Camera Networks, in Multi-Camera Networks: Concepts and Applications, edited by Hamid Aghajan and Andrea Cavallaro, 2009, in press.

- **Journal Papers**

  J.1 I. Tošić and P. Frossard, Learning stereo visual dictionaries, in preparation for IEEE Transactions on Image Processing.

  J.2 I. Tošić and P. Frossard, Geometry-Based Distributed Scene Representation With Omnidirectional Vision Sensors, IEEE Transactions on Image Processing, Vol. 17, Nr. 7, pp. 1033-1046, 2008.

  J.3 R. Sala Llonch, E. Kokiopoulou, I. Tošić and P. Frossard, 3D Face Recognition with Sparse Spherical Representations, submitted to Pattern Recognition, 2008.

  J.4 V. Thirumalai, I. Tošić and P. Frossard, Symmetric Distributed Coding of Stereo Omnidirectional Images, Signal Processing: Image Communication, Special issue on Distributed Video Coding, Vol. 23, Nr. 5, pp. 379-390, 2008.

  J.5 I. Tošić, P. Frossard and P. Vandergheynst, Progressive Coding of 3-D Objects Based on Overcomplete Decompositions, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 16, Nr. 11, pp. 1338-1349, 2006.

  J.6 Z. H. Perić, I. Tošić, Piecewise uniform switched vector quantization of the memoryless two-dimensional Laplass source, "Data Recording, Storage & Processing". 2004. - Vol.6. - No.1. -pp.20-33.

- **Refereed Conference Papers**

  C.1 I. Tošić and P. Frossard, Conditions for recovery of sparse signals correlated by local transforms, accepted to the IEEE International Symposium on Information Theory, 2009.

  C.2 I. Tošić and P. Frossard, Low bit-rate compression of omnidirectional images, accepted to the Picture Coding Symposium, 2009.

  C.3 T. Tošić, I. Tošić and P. Frossard, Nonparametric Least Squares Regression for Image Reconstruction on the Sphere, accepted to the Picture Coding Symposium, 2009.

  C.4 I. Tošić and P. Frossard, Geometry-based distributed coding of multi-view omnidirectional images, Proceedings of the IEEE International Conference on Image Processing, 2008.

  C.5 R. S. Llonch, E. Kokiopoulou, I. Tošić and P. Frossard, 3D Face Recognition using Sparse Spherical Representations, Proceedings of the International Conference on Pattern Recognition, 2008.

C.6 V. Thirumalai, I. Tošić and P. Frossard, Balanced Distributed Coding of Omnidirectional Images, Proceedings of the SPIE - Visual Communication and Image Processing, 2008 (invited paper)

C.7 I. Tošić, P. Frossard, Wyner-Ziv coding of multi-view omnidirectional images with overcomplete decompositions, Proceedings of the IEEE International Conference on Image Processing, 2007 (invited paper)

C.8 V. Thirumalai, I. Tošić, P. Frossard, Distributed coding of multiresolution omnidirectional images, Proceedings of the IEEE International Conference on Image Processing, 2007

C.9 I. Tošić and P. Frossard, Coarse scene geometry estimation from sparse approximations of multi-view omnidirectional images, Proceedings of the European Signal Processing Conference, 2007 (invited paper)

C.10 I. Tošić, P. Frossard, Omnidirectional views selection for scene representation, Proceedings of the IEEE International Conference on Image Processing, 2006

C.11 I. Tošić, P. Frossard, FST-based Reconstruction of 3D-models from Non-Uniformly Sampled Datasets on the Sphere, Proceedings of the Picture Coding Symposium, 2006

C.12 I. Tošić, I. Bogdanova, P. Frossard and P. Vandergheynst, Multiresolution Motion Estimation for Omnidirectional Images, Proceedings of the European Signal Processing Conference, 2005

C.13 I. Tošić, P. Frossard and P. Vandergheynst, Progressive low bit rate coding of simple 3D objects with Matching Pursuit, Proceedings of the IEEE Data Compression Conference, 2005

C.14 I. Tošić, D. Djordjević, Analysis of various linearization methods for two-dimensional memoryless Laplass source, TELFOR 2002, Belgrade

C.15 D. Drača and I. Tošić, Probability of error for IM-DD optical communication system in the presence of quantum noise, thermal noise in the receiver and disturbances in the fiber, ICEST 2002, Niš

- **Technical reports**

T.1 V. Thirumalai, I. Tošić and P. Frossard, Symmetric Distributed Coding of Stereo Omnidirectional Images, No TR-ITS-2008.13, February 2008

T.2 I. Tošić and P. Frossard, Geometry-based scene representation with distributed vision sensors., No TR-ITS-2007.11, August 2007

T.3 I. Tošić, P. Frossard and P. Vandergheynst, Progressive coding of 3D objects based on overcomplete decompositions, No TR-ITS-2005.026, October 2005