

IMAGE/VIDEO REPRESENTATION AND SCALABLE CODING USING REDUNDANT DICTIONARIES

THÈSE N° 3316 (2005)

PRÉSENTÉE À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Institut de traitement des signaux

SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Adel RAHMOUNE

ingénieur d'état en génie électrique, Université de Boumerdes, Algérie
de nationalité algérienne

acceptée sur proposition du jury:

Prof. P. Vandergheynst, directeur de thèse
Prof. P. Frossard, rapporteur
Prof. B. Pesquet-Popescu, rapporteur
Dr P. Schelkens, rapporteur

Lausanne, EPFL
2005

Abstract

Compact or efficient representation for either images or image sequences is key operation to performing image and video processing tasks, such as compression, analysis, etc. The efficiency of an approximation is evaluated by the sparsity measure of the approximation, i.e., the sparsest the representation is, the more efficient it is. For image processing tasks, it is often desired to decompose the image into a linear combination of few visual primitives or features selected from a large collection of waveforms, called the dictionary. These primitives are usually designed in such a way to achieve some good approximation performances by assuming a given class of functions to model images. A common class of functions for image modeling is the set of functions having discontinuities along smooth contours or boundaries, delineating smooth geometrical regions.

It was shown that the well-known separable isotropically scaled two-dimensional wavelets fail to capture the geometrical regularity inherent to images. Motivated by this issue, we designed two rich dictionaries \mathcal{D}_s and \mathcal{D}_{st} , composed of multi-scaled ridge-like functions satisfying the anisotropy and the directionality features, for image and video expansions, respectively. Regarding the spatio-temporal dictionary \mathcal{D}_{st} , we defined a warping operator W , which aligns these functions along the coherent motion trajectories in order to exploit the nature of the temporal evolution in the video signal.

To obtain a compact representation for the visual signal over the dictionaries \mathcal{D}_s and \mathcal{D}_{st} , only the primitives that best match the signal components must be selected. This problem is well studied in approximation theory and it is known as the problem of sparse approximation in unrestricted dictionaries, whose optimal solution is *NP-hard*. Greedy approaches, such as MP and OMP, have been proposed to provide sub-optimal solutions by iteratively choosing one atom at a time. However, the computational complexity of these algorithms is cumbersome and limits their applicability. To alleviate this issue, we introduced a greedy algorithm, called the M-Term Pursuit MTP, whose performances are very close to those related to MP. Then we employed both algorithms, MP and MTP, for image and image sequence decompositions and we investigated their performances.

A target application has been studied, which is the scalable compression, where the aim is to generate a bit-stream that can provide both SNR and geometrical scalability. Geometrical scalability refers to the spatial scalability in case the of images and to the spatio-temporal scalability in the case of video sequences.

The SNR or quality scalability is related mainly to the nature of the greedy approaches and the embedded quantization and coding. To do so, we designed a rate allocation algorithm that

offers a progressively refinable bit-stream, based on the subsets approach. On the other hand, the geometrical scalability is fulfilled thanks to the structure of the dictionaries, \mathcal{D}_s and \mathcal{D}_{st} , which are designed using parametric fashion.

Comparisons with state-of-the-art scalable and non-scalable codecs illustrate the good performance of the proposed compression techniques at low rates.

Version Abrégée

La représentation compacte des images ou des séquences d'images est une étape primordiale pour accomplir les tâches usuelles de traitement d'images ou de la vidéo telles que la compression, le débruitage, l'analyse de contenu, etc... L'efficacité d'une telle représentation est déterminée par le critère de "sparsity", qui est défini comme le nombre de termes utilisés dans une telle représentation, c.à.d. la meilleure représentation est celle ayant moins de termes, tout en tolérant une certaine erreur. Pour ce faire, on désire souvent de décomposer l'image en une combinaison linéaire de peu de primitives ou de composantes visuelles choisies parmi une large collection de fonctions bi-dimensionnelles, appelée le dictionnaire. Ces primitives sont habituellement conçues de telle manière que les représentations obtenues ont de bonnes performances d'approximation, tout en supposant une classe des fonctions pour les images en question. La classe des fonctions la plus récurrente pour la modélisation des images est celle qui englobe toutes les fonctions bi-dimensionnelles ayant une certaine régularité tout au long des contours et aussi dans des régions géométriques délimitées par ceux-ci.

Cependant, il a été démontré que les transformations de nature séparable, telle que les ondelettes isotropes, ne sont pas optimales pour capturer la géométrie des images. Pour remédier à cela, deux dictionnaires redondants ont été conçus pour la décomposition des images et des séquences d'images (désignés par \mathcal{D}_s et \mathcal{D}_{st} respectivement) à partir des fonctions de type "ridge-like", dilatées suivant des paramètres anisotropes, et orientées dans différentes directions. Concernant le dictionnaire spatio-temporel \mathcal{D}_{st} , un opérateur de déformation W a été défini, qui permet d'aligner ces fonctions le long des trajectoires de mouvement afin d'exploiter la nature de l'évolution temporelle du signal vidéo.

Afin d'obtenir une expansion compacte du signal visuel dans tels dictionnaires, seuls les primitives ou les *atomes* qui caractérisent mieux les composantes du signal devraient être choisis. Néanmoins ce problème a été bien étudié dans la théorie d'approximation, et est connu comme le problème d'approximation "sparse" ou parcimonieuse dans des dictionnaires non-restreints, et dont la solution optimale est en général de type *NP-Hard*. Des approches "greedy", basées sur le Matching Pursuit (MP) ou l'Orthogonal MP, ont été proposées pour trouver des solutions sous-optimales d'une manière itérative, où à chaque fois, l'atome qui est mieux corrélé au signal résiduel est sélectionné. Cependant, ces deux techniques ont une complexité de calculs assez considérable ce qui restreint leur champ d'application. Pour alléger un peu ce problème, nous avons introduit un algorithme d'approximation gourmand, appelé M-Term Pursuit (MTP), dont les performances

d'approximation sont légèrement inférieures à celles liées au MP. Les deux algorithmes (MP et MTP) ont été étudiés pour les décompositions d'images et de séquences vidéo.

Une application a été considérée, qui est la compression scalable, où le but est de générer un bit-stream qui peut offrir la scalabilité de qualité et géométrique (spatiale et temporelle). La scalabilité de qualité ou SNR est liée principalement à la nature des approches greedy et à la quantification et du codage. Ceci a été accompli en concevant un algorithme d'attribution de bit-rate pour avoir un bit-stream progressif. Par ailleurs, la scalabilité géométrique est réalisée grâce à la structure paramétrique des dictionnaires \mathcal{D}_s et \mathcal{D}_{st} .

Des comparaisons avec les codecs récents, qu'ils soient scalables ou non-scalables, ont été réalisées et cela a permis de souligner l'efficacité des approches proposées à bas débit.

Contents

Acknowledgments	iii
Abstract	v
Version Abrégée	vii
Contents	ix
Notations	xv
List of Figures	xix
List of Tables	xxi
1 Introduction	1
1.1 Image and Video Representation	1
1.2 Scalable Compression of Images and Video Sequences	2
1.3 More on Scalability Features	2
1.4 Main Contributions	3
1.5 Outline of the Dissertation	4
I State-of-the-Art	5
2 Image and Video Representation	7
2.1 Introduction	7
2.2 Image Representation	7
2.2.1 Regularity	8
2.2.2 Wavelet Bases	8
2.2.3 Adaptive Geometric Basis	12
2.2.4 Ridgelets and Curvelets	15
2.2.5 The Human Visual System (HVS) Primitives	15
2.2.6 Other Approaches	16

2.3	Video Representation	17
2.3.1	Modeling the Video Signal	17
2.3.2	The Predictive Feedback Paradigm	17
2.3.3	Three-dimensional Wavelet Basis	18
2.3.4	Lifting-Based Wavelets	19
2.3.5	Overcomplete Wavelets	19
2.3.6	Geometric Adaptive Basis	19
2.3.7	Spatio-temporal HVS Primitives	20
2.4	Dictionaries of Smooth Functions and Ridge-like Functions	20
2.5	Summary	20
 II A Tour of Sparse Approximation		23
 3 Sparse Approximation		25
3.1	Introduction	25
3.2	Mathematical Notions	26
3.2.1	The Dictionary	26
3.2.2	Vector Norms and Cost Functions	26
3.3	The Problem of Sparse Approximation	27
3.3.1	The Sparsest Representation of a Signal	27
3.3.2	Error-Constrained Approximation	28
3.3.3	Sparsity-Constrained Approximation	28
3.3.4	The Subset Selection Problem	28
3.4	The Computational Complexity of Optimal Solutions	28
3.5	Greedy Approaches	29
3.5.1	The Matching Pursuit MP	29
3.5.2	The Orthogonal Matching Pursuit OMP	31
3.6	Convex Relaxation Methods	33
3.6.1	The Sparsest Representation of a Signal	34
3.6.2	The Error-Constrained Approximation	34
3.6.3	The Subset Selection	34
3.7	Miscellaneous Approaches	35
3.8	Summary	35
 4 The M-Term Pursuit		37
4.1	Introduction	37
4.2	Mathematical Preliminaries	37
4.2.1	Sub-dictionaries	37
4.2.2	The Coherence	38
4.2.3	The Cumulative Coherence	38
4.2.4	Dictionary Partitions	38
4.2.5	Matrix Norms	39

4.2.6	Singular Values of The Gram Matrix	39
4.3	The M-Term Pursuit Algorithm in Finite Dimensional Spaces	40
4.4	The M-Term Pursuit in Infinite Dimensional Hilbert Spaces	44
4.4.1	The Approximation Algorithm	44
4.5	Approximation Performance	45
4.5.1	Comparison with MP	45
4.5.2	The Behavior of h_1 and h_{mp}	47
4.5.3	The Behavior of h_2 and h_{mp}	47
4.6	The Computational Complexity	48
4.7	Summary	48
III Visual Information Representation		53
5	Redundant Geometrical Dictionaries	55
5.1	Introduction	55
5.2	The Spatial Dictionary	55
5.3	The Spatio-Temporal Dictionary	58
5.4	The Warped Spatio-Temporal Dictionary	61
5.5	The Warping Operator W	64
5.5.1	The Optical Flow	64
5.5.2	Regularization	66
5.5.3	Matching Methods	66
5.5.4	The Motion Trajectories $(\mathbf{c}(t), t)$	69
5.5.5	Temporal Adaptivity	73
5.5.6	Inference of Forward Motion Fields	73
5.6	Boundary Atoms	75
5.7	Group Invariance Properties	76
5.8	Summary	76
6	Sparse Image Representation	79
6.1	Introduction	79
6.2	Matching Pursuit for Image Representation	79
6.2.1	The Algorithm	79
6.2.2	Convergence Rate	80
6.3	M-Term Pursuit for Image Representation	80
6.3.1	Dictionary Partitioning	80
6.3.2	The MTP Decomposition Algorithm	81
6.3.3	Convergence Rate	83
6.4	Approximation Performances	85
6.4.1	Threshold Parameters Effects	85
6.4.2	Comparison of MP against OMP	85
6.4.3	Approximation of Test Images	85

6.5	The Computational Complexity of MP and MTP Algorithms	88
6.6	Summary	93
7	Sparse Video Representation	95
7.1	Introduction	95
7.2	Matching Pursuit for Video Decomposition	95
7.2.1	The Algorithm	95
7.2.2	Convergence Rate	98
7.2.3	Choice for the Three-dimensional Dictionary	98
7.3	M-Term Pursuit for Video Decomposition	98
7.3.1	Dictionary Partitioning	98
7.3.2	The MTP Algorithm	99
7.3.3	Convergence Rate	101
7.4	Comparison of Approximation Performances	103
7.5	The Computational Complexity of the Pursuits Algorithms	104
7.6	Stability of the Pursuits Algorithms	105
7.7	Spatio-temporal Atoms as Visual Features	105
7.8	Summary	107
IV	Application to Scalable Compression	115
8	Scalable Compression	117
8.1	Introduction	117
8.2	Transform Coding	117
8.3	Scalability Requirements	118
8.3.1	SNR Scalability	118
8.3.2	Geometrical scalability	118
8.4	The Coding Schemes	119
8.4.1	An Overview	119
8.4.2	Problem Statement	120
8.5	Embedded Coding and Quantization	121
8.5.1	Scalable Coding	121
8.5.2	Coefficient Quantization	123
8.6	Rate Allocation	126
8.6.1	Problem Statement	126
8.6.2	The Rate Allocation Algorithm	127
8.6.3	The Procedure for Optimizing $J_i(\lambda_i)$	128
8.6.4	An Example of Rate Allocation	131
8.7	Summary	131

9	Evaluation of the Compression Schemes	135
9.1	Introduction	135
9.2	Evaluation of the MTP-based Image Coding Scheme	136
9.2.1	Rate Distortion Comparison	136
9.2.2	Visual Quality Comparison	136
9.3	Evaluation of the MP-based Video Coding Scheme (MP3D)	137
9.3.1	Rate Distortion Comparison	137
9.3.2	Visual Quality Comparison	138
9.4	Evaluation of the MTP-based Video Coding Scheme	139
9.4.1	Rate Distortion Comparison	139
9.4.2	Visual Quality Comparison	139
9.5	Evaluation of the SNR Scalability	140
9.5.1	SNR Scalability Performance	140
9.5.2	SNR Scalability Comparison	140
9.6	Evaluation of the Geometrical Scalability	140
9.6.1	Spatio-temporal Scalability Features	140
9.6.2	Spatio-temporal Scalability Comparison	142
9.7	Summary	142
V	Discussions	157
10	General Conclusions	159
10.1	Summary	159
10.2	Future Directions	161
VI	Annex	163
	Appendix	165
A.1	The entropy and the distortion in the coefficient subsets	165
	Bibliography	169
	Curriculum Vitae	181