TOWARD SPARSE AND GEOMETRY ADAPTED VIDEO APPROXIMATIONS

THÈSE N^{O} 3299 (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

Òscar DIVORRA ESCODA

Enginyer de telecomunicació, Universitat Politècnica de Catalunya, Barcelona, Espagne
et de nationalité espagnole

acceptée sur proposition du jury:

Prof. P. Vandergheynst, directeur de thèse
Dr M. Bierlaire, rapporteur
Dr M. Davies, rapporteur
Prof. M. Kunt, rapporteur
Prof. B. Macq, rapporteur

Lausanne, EPFL
2005

A la Rosa,

i als meus Pares.

L'escala d'aquesta vida,
l'has de pujar a poc a poc,
que si la puges de pressa
cauràs al primer escaló.

Popular Catalana

Acknowledgments

A PhD is, with no doubt, a unique experience. Not just because one does it, at most, once in a live (normally...), but because it is an extremely enriching experience. Through this PhD, I have had the exceptional opportunity to grow in all aspects. I have acquired knowledge, experience and know-how in research. Also, and may be even the most important, I have acquired a price-less live experience, I have widened my personal horizons and I have had the opportunity to meet wonderful people and make friends.

Non of this would have been possible without the worthful contribution of a myriad of persons and friends who deserve, at least (but never enough), to be acknowledged in this text. However, I hope that those that I may forget by mistake will forgive me for that.

First of all, I would like to thank Prof. Kunt for having welcomed me at the ITS, and for giving me the opportunity of doing a PhD at EPFL.

All my thanks go to Pierre (Prof. Vandergheynst): for having selected me as one of his PhD students, for the great opportunity he has given me to carry research in freedom, for all the discussions, for his support, for his trust, for all what I have learned (as a researcher and as a person) and for all the fun we shared.

An extremely important aspect of research is collaboration. Sharing knowledge and efforts with other fellows is of extreme synergy as well as it can be of uttermost fun. During these years I have had the chance to have very fruitful discussions with many people. I want to thank specially for this Prof. Macq, Dr. De Vleeschouwer, Prof. Frossard, Prof. Bierlaire, Dr. Flierl and Prof. Vandergheynst. I want to thank also Dr. Davies for his very valuable comments as a jury member of my PhD thesis. Moreover, the best and price-less experience has been the opportunity to discuss with many people at ITS and LTS2 (former F-Group) and to join efforts specially with *Dr. Figueras*, *Lorenzo G.*, *Ana*, *Lorenzo P.*, *Gianluca*, *Philipe*, *Patricia*. Thanks to you all!

Thanks also to the students that worked with me, I have learnt a lot with them: Florent, Gerard, Alvaro i Javier.

Fundamental research is, with no doubt, uncertain; things are not always as one expects or one would like. I want to thank Francesco and Julien for their great patience and friendliness.

This thesis has been source of experiences and trips. One of the best ones has been the time I spent in Belgium at TELE in UCL. I want to thank all the people from TELE for their nice welcome and company and specially the friends that shared with me some moments in there (as well as later!): Christophe, Stephanie, Laurent, Laurence and Marc.

I want to thank my office mate Ana, for all the discussions we have had as well as to my old office mates: Raphael and Meritxell, for their great company.

I want to thank Prof. Marqués for having suggested me to come to ITS, and to Dr. Clawin for having been the first to introduce me to the taste of research.

Apart from the academic experiences, this years in Lausanne would have not been the same without all the great time and fun I have spent with the past and present people at ITS, the catalan community of friends and all the people from here and there I have met.

I have had much joy hiking and skying in the mountains and doing many other things in this beautiful country. Thanks for all that to all those having shared their good humor and joy with me: Rosa, Patricia, Xavier, Raphael, Olivier, Andrea, Nicolas, Torsten, Elena, Elisa, Pierre, Stefan, Francesco, Vlad, Matheo...

These years would have not been the same without the presence of this huge community of estrange people that one may find everywhere: Catalans! Thanks to you all for having exported our very singular way of being. *Moltes gràcies a tots i totes per la vostra amistat i per ser com sou, els que encara hi sou com aquells que ja heu marxat:* Albert, Àngela, Sergi, Ferran, Francesc, Edu, Akira, Lluís, Yolanda, Xavi, Meritxell, Maria Eugènia, Ruth... *i molts d'altres que espero que no els hi sàpigui greu que no els pugui anomenar.*

Finalment, voldria agrair aquells que de forma molt més personal han fet possible que hagi arribat a la fi de tot plegat. Vull donar les gràcies a la Rosa, la meva companya i estimada. El seu ajut, paciència i suport han estat claus per poder tirar endavant aquesta tesi. També vull donar les gràcies als meus pares, pel seu suport, la seva confiança i per haver-me donat els valors i la força que m'han donat.

Abstract

Video signals are sequences of natural images, where images are often modeled as piecewise-smooth signals. Hence, video can be seen as a 3D piecewise-smooth signal made of piecewise-smooth regions that move through time. Based on the piecewise-smooth model and on related theoretical work on rate-distortion performance of wavelet and oracle based coding schemes, one can better analyze the appropriate coding strategies that adaptive video codecs need to implement in order to be efficient. Efficient video representations for coding purposes require the use of adaptive signal decompositions able to capture appropriately the structure and redundancy appearing in video signals. Adaptivity needs to be such that it allows for proper modeling of signals in order to represent these with the lowest possible coding cost. Video is a very structured signal with high geometric content. This includes temporal geometry (normally represented by motion information) as well as spatial geometry. Clearly, most of past and present strategies used to represent video signals do not exploit properly its spatial geometry. Similarly to the case of images, a very interesting approach seems to be the decomposition of video using large over-complete libraries of basis functions able to represent salient geometric features of the signal. In the framework of video, these features should model 2D geometric video components as well as their temporal evolution, forming spatio-temporal 3D geometric primitives.

Through this PhD dissertation, different aspects on the use of adaptivity in video representation are studied looking toward exploiting both aspects of video: its piecewise nature and the geometry.

The first part of this work studies the use of localized temporal adaptivity in subband video coding. This is done considering two transformation schemes used for video coding: 3D wavelet representations and motion compensated temporal filtering. A theoretical R-D analysis as well as empirical results demonstrate how temporal adaptivity improves coding performance of moving edges in 3D transform (without motion compensation) based video coding. Adaptivity allows, at the same time, to equally exploit redundancy in non-moving video areas. The analogy between motion compensated video and 1D piecewise-smooth signals is studied as well. This motivates the introduction of local length adaptivity within frame-adaptive motion compensated lifted wavelet decompositions. This allows an optimal rate-distortion performance when video motion trajectories are shorter than the transformation "Group Of Pictures", or when efficient motion compensation can not be ensured.

After studying temporal adaptivity, the second part of this thesis is dedicated to understand the fundamentals of how can temporal and spatial geometry be jointly exploited. This work builds on some previous results that considered the representation of spatial geometry in video (but not temporal, i.e, without motion).

In order to obtain flexible and efficient (sparse) signal representations, using redundant dictionaries, the use of highly non-linear decomposition algorithms, like Matching Pursuit, is required. General signal representation using these techniques is still quite unexplored. For this reason, pre-

vious to the study of video representation, some aspects of non-linear decomposition algorithms and the efficient decomposition of images using Matching Pursuits and a geometric dictionary are investigated.

A part of this investigation concerns the study on the influence of using *a priori* models within approximation non-linear algorithms. Dictionaries with a high internal coherence have some problems to obtain optimally sparse signal representations when used with Matching Pursuits. It is proved, theoretically and empirically, that inserting in this algorithm *a priori* models allows to improve the capacity to obtain sparse signal approximations, mainly when coherent dictionaries are used.

Another point discussed in this preliminary study, on the use of *Matching Pursuits*, concerns the approach used in this work for the decompositions of video frames and images. The technique proposed in this thesis improves a previous work, where authors had to recur to sub-optimal Matching Pursuit strategies (using Genetic Algorithms), given the size of the functions library. In this work the use of full search strategies is made possible, at the same time that approximation efficiency is significantly improved and computational complexity is reduced.

Finally, *a priori* based Matching Pursuit geometric decompositions are investigated for geometric video representations. Regularity constraints are taken into account to recover the temporal evolution of spatial geometric signal components. The results obtained for coding and multi-modal (audio-visual) signal analysis, clarify many unknowns and show to be promising, encouraging to prosecute research on the subject.

Version Abrégée

Le signal vidéo est une séquence d'images en mouvement, dont les images sont souvent modelées comme des signaux *réguliers par morceaux*. Ainsi, le signal vidéo peut être considéré comme un signal 3D *régulier par morceaux*, et composé de régions qui suivent un certain mouvement a travers le temps. La modélisation du signal vidéo *par morceaux* permet d'analyser en détail le comportement de différentes stratégies de codage, et ainsi de déterminer quelles sont les approches les plus appropriées pour maximiser le taux de compression. Afin de permettre un codage efficace de la vidéo, il est nécessaire d'utiliser des méthodes adaptatives de décomposition du signal. Cette adaptabilité doit être optimisée pour garantir une modélisation du signal avec un coût de codage minimum. La nature du signal vidéo est fortement liée à sa structure, avec une forte composante géométrique. Celle-ci inclut tant la géométrie temporelle (normalement représentée par l'information du mouvement) que la géométrie spatiale. La plupart des méthodes utilisées pour la représentation du signal vidéo ne tient pas compte de sa géométrie spatiale. De même que dans le cas des images, une stratégie prometteuse pour exploiter conjointement la structure géométrique spatio-temporelle est celle qui utilise des dictionnaires redondants avec une forte composante géométrique. Dans le contexte de la vidéo, les primitives géométriques 2D doivent suivre une évolution temporelle en formant des complexes primitives 3D qui ont la fonction de représenter, en même temps, les composantes géométriques spatiales et temporelles du signal.

Dans cette thèse de doctorat, plusieurs aspects concernant l'utilisation de méthodes adaptatives pour la modélisation du signal vidéo sont traités. Cette thèse traite particulièrement des aspects structurels de la vidéo ainsi que de sa nature géométrique.

La première partie du présent travail porte sur l'étude de l'utilisation de décompositions temporelles adaptatives dans des approches basées sur la décomposition de la vidéo en sous-bandes. L'influence de l'adaptabilité est notamment discutée pour deux stratégies de codage: les transformées en ondelettes 3D et le filtrage temporel avec compensation de mouvement. Les avantages de l'utilisation d'adaptabilité dans des représentations basées sur la transformée en ondelettes 3D sont démontrés à l'aide d'une étude théorique R-D ainsi que par des résultats expérimentaux. L'utilisation de l'adaptabilité dans le cadre du filtrage temporel avec compensation du mouvement est aussi étudiée en faisant une analogie entre le signal vidéo, avec le mouvement compensé, et les signaux *réguliers par morceaux* 1D. Cette analogie suggère l'introduction des transformées de longitude localement variable dans des schémas de décomposition par ondelettes bases sur des *lifting steps*. Cette modification permet une plus forte compression du signal grâce à une meilleure adaptation de la représentation des trajectoires de mouvement avec une longueur inférieure à celle du Groupe d'Images (GOP - en anglais -) ou quand l'erreur due à la compensation de mouvement est trop élevé.

Après l'étude d'adaptabilité temporelle, une deuxième partie de cette thèse se concentre également sur l'étude et la compréhension des concepts de base pour exploiter, conjointement, la struc-

ture géométrique spatio-temporelle du signal. Cette recherche se base sur des études précédentes qui tenaient compte de la géométrie spatiale de la vidéo, sans considérer son évolution temporelle (mouvement).

Afin d'obtenir des représentations flexibles et efficaces (parcimonieuses), avec des dictionnaires redondants, il faut utiliser des algorithmes de décomposition hautement non linéaires, tels que les algorithmes *gloutons* (*Greedy algorithms* et *Matching Pursuits* en anglais). L'utilisation de ces techniques est encore peu explorée. Pour cette raison, avant d'étudier de telles représentations, certains aspects liés à l'utilisation de ces algorithmes conjointement avec des dictionnaires cohérents pour l'approximation des images et de la vidéo sont étudiés.

Une partie de cette étude présente l'utilisation des modèles *à priori* dans des algorithmes non-linéaires comme les *Matching Pursuits*. En fonction du dictionnaire utilisé, et du signal, les *Matching Pursuits* peuvent avoir des grandes difficultés pour arriver à obtenir des expansions parcimonieuses optimales. Basé sur ce résultat, il peut être démontré, de manière théorique et expérimentale, que l'utilisation des modèles *à priori* (comme par exemple des modèles probabilistes) peut contribuer très significativement à l'amélioration des performances de ces algorithmes.

Une autre partie de l'étude préliminaire, pour l'utilisation des *Matching Pursuits*, concerne l'approche utilisée dans cette thèse pour la décomposition du signal vidéo et des images. L'approche proposé dans cette thèse améliore une méthode existante de *Matching Pursuits* utilisée dans le passé dans un but similaire. Cette méthode était basée, pour des raisons de complexité calculatoire, sur l'utilisation d'algorithmes génétiques. La méthode proposée dans cette thèse rend possible, d'une manière plus rapide et efficace, la substitution de l'algorithme génétique par une recherche exhaustive.

Finalement, l'utilisation des *Matching Pursuits* avec des modèles *à priori* pour la décomposition du signal vidéo est étudiée. Des critères de régularité ont étés imposés afin de capturer l'évolution temporelle des composantes géométriques 2D. Les résultats obtenus pour le codage de ces représentations, ainsi que les résultats issus des analyses multimodales (audio/vidéo) d'une séquence, permettent d'éclaircir une grand partie des points incompris jusqu'alors sur l'utilisation des dictionnaires redondants avec des *Matching Pursuits* pour les représentations géométriques adaptatives en espace et en temps du signal vidéo.

Resum

El senyal de vídeo és una seqüència d'imatges en moviment, les quals són sovint modelades com senyals *regulars per parts*. Tenint en compte això, el senyal vídeo pot esser considerat com un senyal 3D *regular per parts*, fet de regions que segueixen un moviment a través del temps. Gràcies a la modelització del senyal de vídeo com un senyal *per parts*, es pot analitzar amb detall el comportanent de certes estratègies de codificació, per tal de determinar quines són les més apropiades per a maximitzar la taxa de compressió. Per a una codificació eficient del vídeo, cal utilitzar mètodes adaptatius de descomposició del senyal. Aquesta adaptivitat cal que sigui optimitzada de manera que la modelització del senyal s'efectuï amb el mínim cost de codificació possible. La natura del senyal de vídeo, fortament lligada a la seva particular estructura, té un alt contingut geomètric. Aquest contingut geomètric contempla tant la geometria temporal (normalment representada per l'informació de moviment) com la geometria espacial. En tot cas, la majoria dels mètodes utilitzats per la representació del senyal de vídeo no tenen en compte la seva geometria espacial. De manera similar al cas de les imatges, una estratègia força prometedora per tenir en compte la geometria del senyal és mitjançant l'utilització de diccionaris de funcions redundants amb un alt contingut geomètric. Dins del contexte del vídeo, aquestes serien complexes funcions base en 3D capaces de modelar components locals del senyal amb contingut geometric espacial i temporal. O sigui, dit d'una altra manera, aquestes funcions descriurien elements de geometria espacial del senyal de video així com la seva evolució temporal.

Al llarg d'aquesta dissertació de doctorat es discuteixen diversos aspectes de l'utilització de mètodes adaptatius per la modelització del senyal de vídeo, donant una èmfasi especial a l'estructura *per parts* del vídeo i la seva natura geomètrica.

La primera part d'aquest treball estudia l'utilització de descomposicions adaptatives en temps dins dels mètodes de codificació de vídeo per subbandes. L'ús d'adaptabilitat és discutit principalment per a dues estratègies de codificació per subbandes: Representations *wavelet* 3D i filtratge temporal amb compensació del moviment. Els avantatges que aporta l'ús d'adaptabilitat en les representacions *wavelet* 3D son demostrats mitjançant un estudi teòric *R-D* així com un seguit de resultats experimentals. En aquest apartat, hom estudia també l'analogia entre el senyal de vídeo amb el moviment compensat, i els senyals 1D *regulars per parts*. Aquesta analogia motiva el fet d'introduir transformades de longitud localment variable dins dels esquemes de descomposició wavelet utilitzant *lifting steps* amb adaptació multi-hipotesi al moviment. Aquesta millora permet una major compressió del senyal. Això és gràcies a una millor adaptació de la representació en aquelles trajectòries de moviment que tenen llargada inferior a la de la unitat mínima de processament (Grup d'Imatges -*Group of Pictures* en anglès-), així com quan l'error degut a la compensació del moviment és massa elevat.

Després de l'estudi d'adaptabilitat temporal, la resta de la tesi està especialment dedicada a l'estudi i la comprensió de les bases necessàries per treure profit de l'estructura geométrica espacial i

temporal del vídeo. Aquesta investigació es fa sobre la base d'un estudi anterior que tenia en compte aquesta geometria espacial, però sense tenir en compte la seva evolució temporal (moviment).

Per tal d'obtenir representacions flexibles i eficients (esparses), amb diccionaris de funcions redundants, cal utilitzar algorismes de descomposició altament no linials, com ara són els algorismes *golafres* (*Greedy algorithms* i *Matching Pursuits* en anglès). L'utilització d'aquestes tècniques està encara força inexplorada. Per aquesta raó, i de manera prèvia a l'estudi de les representacions de vídeo, s'exploren certs aspectes de l'utilització d'aquests algorismes juntament amb diccionaris coherents per l'aproximació geomètrica d'imatges.

Una part d'aquest estudi presenta una investigació sobre l'utilització de models *a priori* en algorismes no linials com els *Matching Pursuits*. Depenent del diccionari utilitzat i del senyal a aproximar, els *Matching Pursuits* poden arribar a tenir seriosos problemes per obtenir expansions *esparses* del senyal. Tenint en compte això, es pot provar teòricament i experimentalment, que l'utilització de models *a priori* (per exemple, amb una formulació probabilística) poden conduir a una significativa millora del funcionament d'aquests algorismes.

Una altra part d'aquest estudi preliminar per l'utilització de *Matching Pursuits* tracta la tècnica utilitzada en aquesta tesi per la representació del senyal de vídeo i de les imatges. L'estratègia proposada en aquesta tesi millora un mètode *Matching Pursuit* sub-òptim utilitzat previament per d'altres investigadors. Aquest mètode es basava, per raons de complexitat de càlcul, en l'utilització d'algorismes genètics. El present algorisme ha estat estudiat i millorat per tal de fer possible, de manera ràpida i menys complexa computacionalment, la substitució de l'algorisme genètic per una recerca exhaustiva.

Finalment, s'ha investigat l'utilització de *Matching Pursuits* combinats amb models *a priori* en la descomposició del senyal de vídeo. S'han imposat condicions de regularitat en l'evolució temporal de components geomètriques 2D. Els resultats obtinguts en la codificació del senyal, així com en l'análisis multi-modal (àudio/vídeo) d'una seqüència, aclareixen moltes incògnites al respecte d'aquestes metodologies. A més a més, els resultats es mostren prometedors i encoratgen a proseguir la recerca en aquesta direcció.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

1.1.1 All is about Modeling

Freshly acquired digital signals are nothing but a bunch of samples that, depending on the nature of the recorded data, are structured (among other) in 1D sample vectors (e.g, sound), 2D Cartesian matrices (e.g, images) or 3D Cartesian matrices (e.g, video). These regular sets of samples are a sketch of the real world: they are the result of an uncountable number of physical events interacting among them. One looks toward processing the data measured from these events in order to extract the information it contains. However, digital raw data, as it is recorded, needs further processing.

Luckily for us, physicists have left us the evidence that physical events follow certain rules and can be, up to some degree, quite accurately modeled. Emulating our human senses, digital signal processing applications look into real world data searching for particular signal structures and particular events. Hence, basically, in order to efficiently process digital signals and to exploit their particular structures, appropriate signal models, adapted to applications and the underlying physics of the recorded events, have to be established. As depicted in Fig. 1.1 for a video processing system, a classic scheme applicable to many applications involves first a representation or approximation of the input signal by means of a model. This signal modeling stage has the purpose of reducing as much as possible the dimensionality of the signal and to help for efficient representation. The resulting simplified representation is then used as source of information by the application (e.g. segmentation, tracking, compression, indexing, etc...).

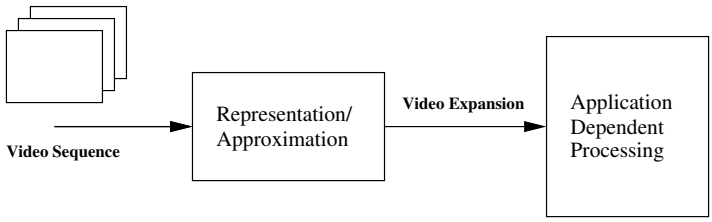


Figure 1.1: Building blocks of a generic video processing scheme.

A classic approach used to model signals is the construction of models based on a superposition

of basic waveforms. Formally this can be written as

$$\hat{f} = \sum_{\gamma \in \Gamma} c_\gamma \cdot g_\gamma, \tag{1.1}$$

where $\hat{f}$ is the model of the original signal f, each g_γ is one of the basic waveforms, c_γ are non-zero scalars that determine how relevant is waveform g_γ within the model and finally, Γ indicates which are the basic waveforms involved in the model (i.e, normally one has a pool of waveforms, sometimes called a basis, and one selects only a few of the available waveforms to construct the model).

This thesis considers, in first instance, signal modeling for the purpose of video compression. Thinking about compression, one can have the reflex of establishing a relation between compression capabilities and the use of low redundancy basis (or, in the limit, orthonormal ones) in order to construct the pool of available waveforms for Eq. (1.1). This has been typically done in order to limit the maximum number of coefficients to code when linear transforms are used (e.g, critically sampled wavelet transforms). However, this is far from being accurate. In a coding application, what counts is the number of bits required to code the signal model. This involves coding the set of used functions I_m and their corresponding coefficients. In fact, even if the pool of possible waveforms to use for signal modeling has an enormous size (with a number of functions much higher than the original number of signal samples) very efficient signal models, capable to well balance between approximation accuracy and coding cost, can be achieved.

1.1.2 Modeling Video and Signal Structure

Good video modeling requires, with no doubt, the use of building functions well adapted to video nature. Scientist realized very early that non-linear models were required for efficient video approximation. Consequently, they introduced the estimation of motion within their coding strategies. Indeed, since video signals are sequences of moving images, very particular phenomena and structures appear in the signal (see Fig. 1.2). In video temporal dimension, long anisotropies and elongated structures appear as the result of moving image regions. A classical model used to predictively ex-

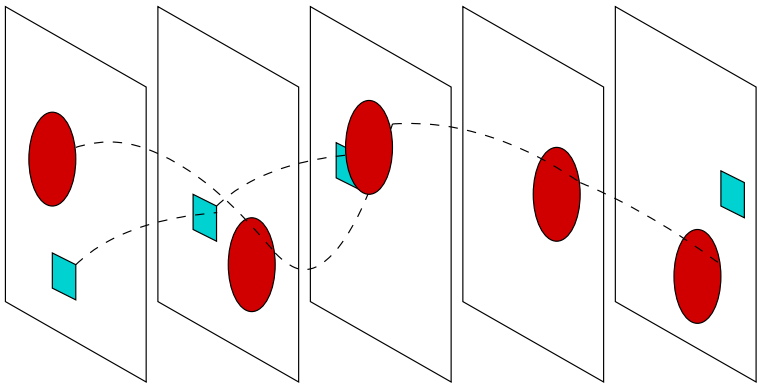


Figure 1.2: Video motion is the temporal geometric transformations of 2D images, i.e, from a 3D point of view, motion is the temporal geometric component of video signals like edge orientation is for 2D images.

tract motion information from video signals (i.e, in order to tune the shape of the basic waveforms in Eq. (1.1)) is block based motion compensation (see next chapter for further detail).

However, this video modeling approach often lacks of accuracy. Indeed, motion is not necessarily

uniform on square patches. Moreover, the arbitrary division of video frames in blocks of a determined size rarely respects the 2D structure of the signal.

Another approach is the one based on the use of temporal wavelet transforms. Temporal functions are oriented along motion trajectories (see Chapter 2). Indeed, video signals are modeled as a summation of multi-scale wavelet basis functions that capture smooth oriented features along motion trajectories. Accurate modeling requires motion adapted temporal wavelet functions to accurately follow motion and to handle appropriately the arbitrary length of object trajectories. However, this is not yet enough to model video signals appropriately.

Video modeling for coding purposes does not reduce to retrieve the best model to achieve compression. Indeed, modern video coding applications require to implement supplementary flexibility features in addition to compression efficiency. Present transmission requirements for network communications and the diversity of receivers having many different processing capabilities, impose to video coding schemes the need for features like:

- Progressive efficient transmission of the information, also known by scalability, permitting a receiver to decode partially video bit-streams such that, depending on the requirements, a downscaled version of the signal (in time or space) or a lower quality version of this may be decoded (see [135]),
- robust video streams allowing to privilege protection of critical data of the video model such that, when sent over a lossy network, the quality loss at reception due to stream corruption is minimized (see [111]),
- efficient representation for data mining such that, in addition to compression, one can also index data in a smart way, etc...

These additional characteristics, necessary for modern video coding schemes, require a well defined structuring of video data. For this purpose, models used for signal modeling must implement such a structure. Hence, going back to Eq. (1.1), the selection of basic waveforms as well as the pool of waveforms must be designed such that the signal structure is taken into account.

A clear example of this is the bad behavior of predictive video coding strategies in order to achieve efficient scalable video representations. For this purpose, structured multi-scale video representations have had to be adopted such as spatio-temporal wavelet decompositions (see Chapter 2), or approaches based in spatial Laplacian Pyramid decompositions [74].

Good spatio-temporal video modeling is also of key importance for general purpose video analysis. Accurate modeling of video structures is of great relevance to extract good video features that carry information about the visual scene. Apart from coding, computer vision applications may profit from this accurate modeling. These are, for example, spatio-temporal video segmentation, region retrieval, audio-visual multi-modal analysis, object extraction, etc...

Exploiting the temporal and spatial structure in video signals is very important. Most of present techniques exploit accurately the temporal geometry by using models capable to represent motion. Multi-scale spatial structure can be also exploited by means of wavelet like representations. However, what about spatial geometry? Apart from the multi-scale structure of images, and video frames, edges surrounding regions have a high geometric content, which is normally not exploited by models. Spatial geometry in images and video requires a special attention. Most of spatial information can be found under the form of geometry. Hence, signal modeling strategies capable to exploit this need to be studied.

1.1.3 Spatial Geometry in Images and Video

During the last years, a great effort has been devoted by the scientific community to understand how 2D geometric data contained in images may be efficiently modeled. Indeed, even though dyadic discrete wavelet basis have very nice properties to represent piecewise-smooth signals [121] (like images), they lack of the capacity to represent efficiently singularities of one or higher dimensions (e.g, edges).

Classic dyadic separable wavelets, widely used in image and video compression, are unable to see the regularity along edges in 2D (and higher dimensions) signals. As described by many authors, wavelets are only able to see 1D singularities, i.e. delta singularities. Hence, to describe an edge, a high number of 2D wavelet basis is required. Fig. 1.3 illustrates very clearly this effect. In the left side of the figure, an edge is approximated using a set of separable wavelet basis functions. The spatial support of each one of the functions activated by the edge are represented, in the figure, by the different squares. One can see that the required number of wavelets increases very rapidly with the approximation refinement. In the contrary, the edge is much more efficiently modeled by a set of elongated basis functions where geometry is taken into account (see again Fig. 1.3, right).

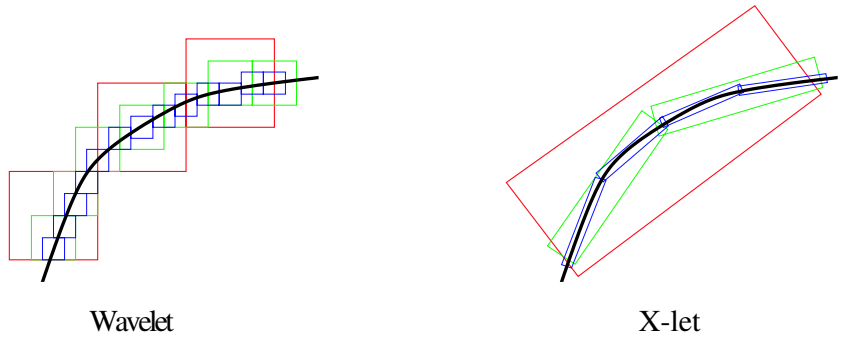


Figure 1.3: Non-geometric vs. geometric representation of a 2D edge: a clear difference of behavior can be appreciated between classic isotropic dyadic wavelets and a geometry oriented dictionary (x-let) [52].

Many approaches have been proposed in order to supply the necessary flexibility for geometry based signal models. During the last years, an uncountable number of bases (or frames [121]) and other non-linear approaches have been proposed to supply such adapted representations. These are well known and can be easily recognized by their "lets" ended name: curvelets [27], contourlets [52], ridgelets [26], wedgelets [53], bandelets [118], etc... Apart from these strategies, some investigations have been carried on the use of redundant, highly dense, geometric dictionaries together with highly non-linear decomposition algorithms like Matching Pursuits [50, 61, 66, 78, 141, 178]. Highly non-linear decomposition algorithms may achieve much sparser signal representations over redundant dictionaries than those achieved by frame based strategies. Indeed, frame based signal decomposition strategies are not sparsity preserving [28].

Finally, just to notice that apart from the mathematical proofs that justify the use of geometry based approaches to represent images and video signals, it seems that the human visual system would use similar analysis strategies. Investigations carried by neuroscientists seem to indicate that the brain cortex incorporates geometric based visual information modeling strategies. Physiological studies have shown that the receptive fields in the visual cortex have a localized, oriented and band-pass response. Moreover, research carried to identify sparse components in images [136] and video signals [137] has demonstrated the astonishing resemblance between the retrieved sparse components

and the response of the cortex receptive fields. These studies also suggest that brain analyzes images and video signals based on approaches that resemble more to highly non-linear decomposition algorithms than to frame based ones [137].

1.2 Contribution of the Thesis

This thesis mainly targets the study of adaptive video approximations for efficient video modeling with application to compression and other applications.

We first tackle the problematic, from a theoretical and practical point of view, of optimal temporal decompositions of video in the framework of subband video coding. The MCTF extension of H.263++ [75] is taken as basic platform to test the approach suggested by theoretical investigations. Results and conclusions are general and mainly reflect properties of wavelet approximations theory. Hence, they are valid to be ported into other video standards and wavelet based coding strategies.

Second, sparse fully geometry adapted video approximations, for coding purposes and other applications, are investigated in detail. For this purpose, apart from a practical study on an particular video decomposition scheme, the theoretical basics concerning the signal decomposition algorithms and strategies, for any signal decomposition using redundant dictionaries, are analyzed as well. To this aim, this work proposes:

- a model based Rate-Distortion (R-D) theoretical analysis of different wavelet decomposition strategies for motion compensation free subband video coding. This analysis gives a better understanding of coding performances of non-linear video approximations with 3D separable wavelet bases. A locally adaptive temporal decomposition strategy is suggested in order to improve the R-D performance of coding applications,
- a piecewise-smooth model concerning the temporal behavior of motion compensated video data. According to this, and the theoretical background on R-D performance of wavelet based and oracle based coding schemes, an intra-adaptive scheme of the MCTF extension of H.263++ is proposed. This allows for a more flexible modeling of video signals such that a better R-D performance is achieved.
- a detailed theoretical analysis on the influence of using accessory *a priori* models in highly non-linear signal expansion algorithms together with coherent dictionaries. Practical examples validate the findings and show the relation between sparse solutions, signal structures and the role of the dictionary in the representation of signal features,
- a feasible efficient full search matching pursuit strategy for image decompositions using 2D geometric dictionaries. The proposed algorithm, substitutes previous sub-optimal strategies based on genetic algorithms, improving approximation results and reducing computational complexity,
- an *a priori* based predictive greedy strategy to extract 3D primitives from video signals,
- a spatio-temporal geometric video representation scheme based on the aforementioned 3D primitives. The obtained representation, when used for signal scalable compression, shows to be, in average, more efficient than some *state of the art* techniques. The same representation approach shows to be interesting, as well, as a source of video features for multi-modal data fusion [68],
- a global analysis on the algorithmic requirements for sparse video approximations on using geometric redundant dictionaries.

Valuable aspects on the use of highly non-linear algorithms like, for example, Matching Pursuits, are underlined in this thesis. On one hand, efficient sparse approximations of natural signals require highly coherent dictionaries. On the other hand, Matching Pursuits do not behave properly with such coherent dictionaries. The proposed Bayesian paradigm, for highly non-linear signal approximation algorithms, tries to fill this gap by making decomposition algorithms to be signal adaptive. These results are then applied to extract 3D geometric structures from video signals. The results obtained with the investigated geometric video representation strategy clarify many unknowns and show to be promising, encouraging to prosecute research on the subject.

1.3 Outline of the Thesis

The thesis is organized as follows. In the next two chapters, we focus on the usage of adaptive representations on present state of the art video coding techniques.

Chapter 2 situates the problematic of video approximation by presenting an overview of most common strategies used for video coding. In this chapter, video approximations for video coding are presented as a modeling problem. We review modeling approaches from linear transform techniques up to the most sophisticated non-linear, motion based, video representation strategies. This is performed by describing, at the same time, the video compression paradigms appeared during the last years as well as their features and purposes.

Chapter 3 studies the use of locally adaptive temporal transforms within the framework of subband video coding. Two main wavelet decomposition schemes video representations are considered in this investigation. We take into account the cases of motion compensated free 3D wavelet decompositions as well as Motion Compensated Temporal Filtering (MCTF). Considering video as a piecewise-smooth video signal, a R-D study is presented to justify the necessity to include adaptive temporal transforms in both approaches. Assumptions and the theoretical modeling of the problem are, then, validated by a detailed experimental study based on the MCTF scheme proposed by *Flierl* [75]. For this purpose, we further develop this scheme in order to include the needed temporal transform flexibility.

As most video modeling techniques do not take into account an essential particularity of video signal structure, i.e., video signal has a high spatial geometric content. For this reason, efficient models for video signal approximation should use basic building pieces capable of exploiting such a characteristic. The remaining of this thesis presents a series of studies that expose, little by little, necessary tools for video approximations using redundant geometric dictionaries. After these studies, a particular approach toward sparse and geometry adapted video approximations is presented and investigated.

In **Chapter 4**, a general overview of common concepts in sparse signal representations and approximations is presented. This exposes the general background to understand the problematic of sparse approximations when redundant libraries of functions are used. We examine, as well, some of the most relevant non-linear algorithms used to supply feasible solutions to sparse problems. Some hints are given about the capacity of some of such algorithms to supply optimal solutions to the sparse problem.

Matching Pursuit (MP) appears as one of the few non-linear algorithms for sparse approximations able to handle very big dictionaries of functions. However, very coherent dictionaries pose problems to MP algorithms (as well as other approaches) to find efficient sparse representations of signals. We propose in **Chapter 5** a Bayesian variation of Matching Pursuits as well as Basis Pursuits. From this point of view, this chapter shows theoretically and empirically that *a priori* based adaptive non-linear decomposition algorithms can do significantly better than classic ones. Indeed, signal adaptive

non-linear algorithms show to be a key element in sparse signal representations and approximations with highly redundant dictionaries.

Chapter 6 studies a particular case of geometric image approximations using a redundant dictionary and Matching Pursuits. Given the size of the dictionary, previous attempts [65, 78] of using this approach were obliged to use a suboptimal formulation of the MP algorithm. In this chapter, a feasible optimal MP solution is proposed. The proposed approach is analyzed and evaluated with regard to achieved improvements and complexity costs.

Based on the background of the three previous chapters, **Chapter 7** studies the approximation of video signals by the superposition of 3D spatio-temporal geometric features. These 3D spatio-temporal functions are intended to capture, at the same time, spatial geometry as well as temporal motion. In order to extract the temporal evolution of 2D geometry components (i.e. 2D geometry components obtained with the approach proposed in chapter 6), an *a priori* model is used to help MP in this task. Results are evaluated in a coding framework as well as by using the investigated video representation as a source of video features for multi-modal audio-visual analysis of sequences. The results obtained clarify many unknowns and show to be promising, encouraging to prosecute research on the subject.

Finally, **Chapter 8** concludes with a summary and an outlook of future research.

Chapter 2

Video Representations: A Coding Perspective

2.1 Introduction

There is no doubt that video is a very complex and particular kind of multi-dimensional signal. Through the years, research has put into evidence the strong need of fine analysis of video signals for efficient coding and compression. Video is composed of components with a extremely relevant structure. Indeed, the physical nature of video signal implies that most of its features and characteristics may be modeled as following a given set of laws and behaviors. Efficient compression requires to exploit the redundancy underlaying all that can be specified in terms of a model. One looks for the appropriate modeling of video signals, such that the number of degrees of freedom in the selected representation is significantly reduced with respect to raw video data. Usually, the fewer the parameters needed to model a video approximation, the lesser the information required to code the retrieved parameters.

Fig. 2.1 depicts the main building blocks in a typical coding scheme (applicable, actually, to any kind of signal). A first stage performs the signal modeling such that dimensionality can be reduced, i.e. at this stage *structural* signal redundancy is exploited. Many modeling strategies may be used, these can be classified as linear or non-linear:

- Linear approaches have usually low complexity computational requirements. However, their modeling capacity is very limited.

- Non-linear approaches are often much more powerful for efficient signal modeling but with significantly higher computational cost. These are able to mix different signal description modalities and combine low and high level features description.

For coding purposes, parameters configuring a certain signal model (e.g. coefficients of a linear transform, motion vectors, etc...) can not take any value. That is, their precision must be limited. To achieve that, parameters must be quantized by some means; using the so called scalar quantization or the more sophisticated vector quantization [92, 93]. Finally, entropic coding of quantized parameters takes care of the remaining *statistical* redundancy contained in the quantized signal representation.

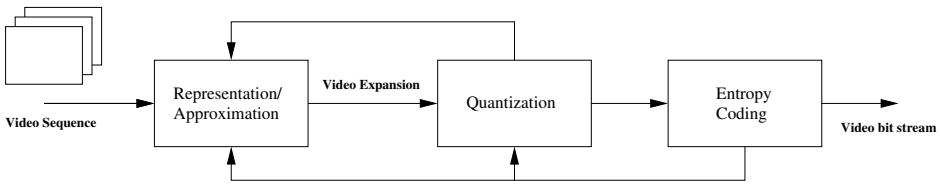


Figure 2.1: Building blocks of a generic compression scheme.

The reader may observe, in Fig. 2.1, some feed-back arrows from the final coding stage toward the quantization and transformation stages. This implies that, mainly for non-linear approaches, signal representations and quantization may be adapted to achieve an optimal rate vs. distortion performance.

Pure signal compression is not the only required feature in modern video coding systems. Indeed, new transmission requirements of the widely used Internet, additional reliability constraints of access networks with packet losses (e.g. Wireless Networks) and the tendency to use, the more and more, networks of sensors to make distributed acquisition of data (sound, video, environmental data, etc...), arise new flexibility and functional necessities to be included in new coding strategies.

With regard to video coding, these can be classified mainly as:

- Video Scalability [135];
- Distributed Video Coding [86];
- Multiple Description Coding [184];
- Robust Video Delivery [111].

Each one of these very emerging research topics is a vast field in itself. Since scalability is the feature that has motivated the main transformation of present coding strategies, this is shortly described in the following. For further details, and for a description of the remaining itemized coding paradigms, the reader is referred to check the suggested references.

Video coding scalability involves the capacity to provide interoperability between different services and to flexibly support receivers with different display, computation and/or reception bandwidth capabilities. Receivers either not capable or not willing to reconstruct the full resolution of a coded signal, can receive and decode subsets of the layered bit stream to display video at a different spatial or temporal resolution from the original or at a lower quality. Hence, main kinds of scalability are:

- PSNR scalability: Signal reconstruction from a truncated version of the bit-stream such that a lower quality in the reconstructed signal is achieved with respect to the original.
- Temporal scalability: Signal reconstruction from a truncated version of the bit-stream such that the displayed frame rate is reduced.
- Spatial scalability: Signal reconstruction from a truncated version of the bit-stream such that the displayed spatial resolution is reduced.

To efficiently achieve such coding properties together with compression efficiency, video *signal structure* requires to be exploited for a maximum flexibility.

This chapter is structured as follows: First, in Sec. 2.2, common strategies of video representation for video coding are reviewed. The strategies described are: predictive video representations,

transform based representations and video expansions on redundant dictionaries. Sec. 2.3 discusses different strategies to exploit temporal geometry (i.e. motion) in video signals. Next, integration of motion description in predictive video coding schemes and transform-based coding schemes is described in Sections 2.4 and 2.5. Finally, conclusions are drawn in Sec. 2.6.

2.2 Video Representations

2.2.1 Predictive Video Representations

Predictive video representation is the first strategy ever used to exploit the temporal redundancy and structure of video signals. It is as early as *1929* that one may find intellectual property documents that describe methods to transmit only changing regions in video coding applications [112].

Predictive video representations intend to exploit temporal video structure by modeling future frames in terms of previous temporal frames. Often, 2D video redundancy is also exploited by means of the addition of a 2D signal transformation within the loop of the predictive signal representation.

Fig. 2.2 shows the structure of a closed-loop prediction scheme. Assuming an initial zero state at time 0, all frames following the first one can be put in terms of the previous ones. This is performed by the predictor, which is in charge of modeling, using a limited set of parameters, the future frame of the sequence. Predictor parameters may be tuned in order to reduce the prediction error variance, i.e. the mean square error of the predictor. The video representation is, thus, composed by the prediction parameters and the residual error. For coding purposes, quantization and dequantization are inserted after the 2D transformation and before the inverse 2D transformation respectively (see Sec. 2.4).

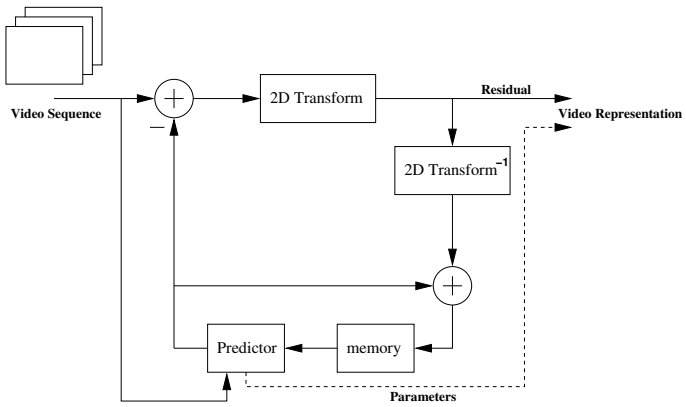


Figure 2.2: Block Diagram of a Predictive Video Representation Scheme.

Given the closed-loop structure, and the consequent infinite feed-back of the residual error, an N-th order linear predictor is considered to perform better than a length N Karhunen-Loeve Transform [108] if the predicted signal is a stationarity stochastic process. However, the widely used assumption of stationary, for images and video, is not accurate and is far from being able to exploit the real piecewise structure of signals. This is the reason why highly non-linear predictors have been introduced for video coding (see Sec. 2.4). In ulterior chapters, non-stationarity in video signals is addressed in more detail in the context of motion compensated temporal filtering (MCTF) and geometric video representations.

2.2.2 Transform Based Representations

Like for any other signal, a strategy for video modeling is transform based representations. In the video milieu, transform representations have often been associated, in the past, to 2D or 3D versions of the DCT block transform [162]. These methods assume the signal to be stationary into each partition block. They do not exploit the structure of the signal unless an adaptive tree-based partition of the signal is used.

Transform based representations concern any modeling of the kind:

$$f = A\mathbf{b}, \tag{2.1}$$

where, without loss of generality, f is a 1D signal, A is a matrix where each column is a basis vector and $\mathbf{b}$ is a coefficients vector. Even if, in here, f is a 1D vector, this can be associated to a raster scanned array of samples from a video Group of Pictures (GOP).

Typically, A is a $m \times m$ matrix with all columns linearly independent. This defines a determined system with equal number of equations and variables. Hence,

$$\mathbf{b} = A^{-1} f,$$

which for orthonormal bases is such that $A^{-1} = A^T$.

A part from DCT, the linear transform formulation of Eq. (2.1) is also the case of 3D wavelet based video representations [37, 110, 113, 146]. Discrete wavelet transform (DWT) has appeared to be a valuable tool for efficient representation of non-stationary signals. Moreover, the multi-scale structure of common dyadic wavelet representations has shown some ability in adapting to human visual characteristics [121]. The local character of wavelet basis functions is of key importance for efficient representation of sudden signal discontinuities like edges. Locality together with the tree-based parent-child structuring of wavelet basis is what allows, in non-linear approximations, a better representation of non-stationary signals compared to block DCT or KLT [60, 121]. Another of the wavelet virtues is their capacity to approximate well polynomial signals, depending on the number of vanishing moments of the wavelet basis in use. From a filter point of view, the number of vanishing moments of a wavelet basis is equal to the number of *zeros* at frequency π of the filter used to generate the wavelet coefficients in an iterative, two channel, filter bank [121, 181]. As seen later in Chapter 3, images and videos are well approximated by piecewise-polynomial models. Hence, 3D wavelets representations may be used to exploit that (with limitations).

Fig. 2.3 depicts the generation of a typical 3D wavelet transform for video representations. A Haar wavelet transform is used to decompose the temporal dimension of a GOP of 8 frames. At each temporal decomposition level, high frequency temporal bands (H, LH, LLH) and the lowest temporal band (LLL) are further decomposed with a 2D wavelet spatial decomposition. The separable 3D wavelet decomposition applied to the video GOP ends up with the well known subband scheme of Fig. 2.4.

The schematic description of a 3D wavelet decomposition depicted above clearly states the easiness of extending 2D wavelet decompositions to 3D domain. This is achieved thanks to the possibility to generate 3D wavelet kernels by the tensorial product of 1D kernels. From a computational point of view, this is a great advantage.

Video is composed by smooth regions that move smoothly trough time and that are surrounded by local singularities like edges. Hence, 3D wavelets can obtain a good representations with additional interesting properties like their multi-scale structure. Wavelet representations have, however, their limitations. Besides their advantages, they present a set of drawbacks that limit the performance of applications where they are used. This can be enumerated as follows:

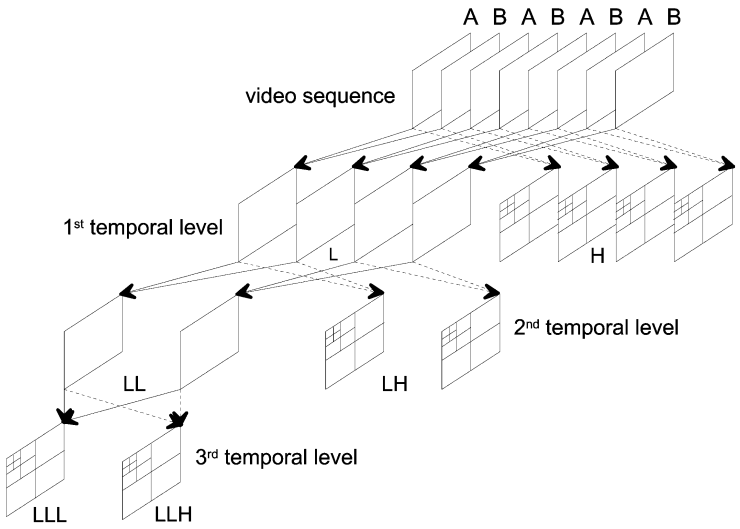


Figure 2.3: 3D Wavelet Video decomposition. In the scheme, the 2D isotropic spacial DWT follows a temporal Haar based decomposition. Due to the linearity of the operations their order can be instinctively swapped [135].

- Independently of the dimensionality, wavelet decompositions do not exploit the correlation among coefficients of different scale subbands issued from the same signal singularity (e.g. and edge).
- In 2D and above, separable wavelets are not capable to exploit smooth structures like image edges and motion trajectories in video sequences. Indeed, separable wavelet based representations are unable to capture signals geometry. During the last years, however, intensive works have been carried out in order to include motion information in 3D wavelet video representations, as exposed later in this chapter.

3D Wavelet Subband Video coding

Popular examples of 3D wavelet coding schemes are 3D-SPIHT [29], ESCOT [189] (which is the 3D version of EBCOT, used to compress images with the JPEG2000 standard [8]) or the 3D extension of GTW [103] which intents to be a generalization of EBCOT. This schemes exploit the 2D+t wavelet representation (the reader may see the subband decomposition scheme in Fig. 2.4) to exploit temporal redundancy of static areas in a video sequence.

Most popular schemes are 3D-SPIHT and ESCOT. These are based on the efficient coding of positions and amplitude of wavelet quantized coefficients. Both supply embedded, bit-plane coding based, bit-streams capable to supply spatial and SNR scalabilities. However, their working principle is significantly different. SPIHT exploits the parent-child relationship among coefficients of different subbands with the same spatial location. This scheme supplies an efficient way of coding significant coefficients position, as well as trees of zero coefficients. The ESCOT approach supplies a quite different coding strategy. It does not exploit multi-scale parent-child relationships. To the contrary, ESCOT supplies a powerful context adaptive arithmetic coding that exploits a large range of coefficients neighborhood models.

Although such approaches do not exhibit very high performances in video coding due to the lack of motion compensation, they are attractive for some applications due to their scalability capabilities, and their low computational complexity needs.

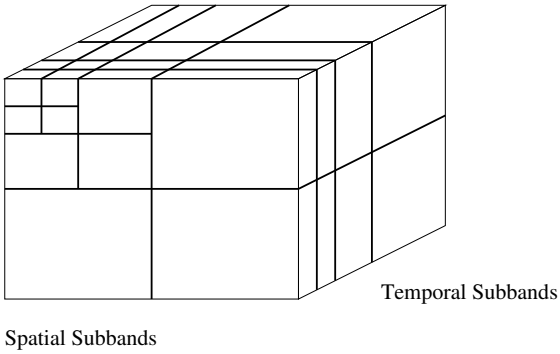


Figure 2.4: Subband scheme of the 3D Wavelet (2D+1D) transform often used in wavelet video coding.

2.2.3 Use of Redundant Dictionaries

The need to exploit geometry in image and video signals has encouraged the exploration of alternative transform schemes. One can establish a similar formulation to Eq. (2.1) but with A not being, this time, a square matrix anymore. In order to introduce enough basis functions capable to model the large variety of geometric characteristics of image and video features, redundant basis need to be considered. A turns out to be an $m \times n$ matrix such that $m < n$, and where each column corresponds to a basis function of the overcomplete dictionary in use. This formulation establishes an under-determined equations system with more than one possible solution. Good solutions are those considered to minimize the number of non-zero coefficients. Hence, a sparsity criteria is often used to retrieve an appropriate signal representation. Exact signal representations (i.e. Eq. (2.1)) do not always satisfy the requirements of certain applications based on natural signals modeling (e.g. denoising or compression). To the contrary, signals approximation is more appropriate for denoising or compression applications.

In this scope, the scientific community has been investigating the use of very overcomplete dictionaries of basis functions with high geometric meaning. While in the image field, numerous research studies have been carried out to exploit spatial geometry (the reader may find a small sample of these in [27, 51, 66, 118, 141, 182]), in the video field this remains quite unexplored.

An approach to video geometric representations is the one proposed by *Frossard* in [78]. In his work, video GOPs are approximated by the superposition of a small set of 3D functions belonging to a vast overcomplete dictionary. In that case, dictionary basis functions were able to model efficiently signal components with edge-like characteristics. Indeed, dictionary functions had an elongated form in the spatial domain, able to adapt to edge lengths and angular orientations. However, temporal geometry (i.e. motion) was not exploited in his approach. The temporal dimension of 3D functions was formed of a length adaptive Gaussian shape. This was intended to exploit temporal redundancy of static, non-moving regions.

Fig. 2.5 depicts the scheme proposed for video signal decomposition. Each GOP is decomposed as the summation of a set of 3D dictionary functions. This decomposition is carried out iteratively by a Matching Pursuit (MP) algorithm [122] (see Chapter 4 for some details on this algorithm). Video representations are composed by the index parameters that determine the 3D selected functions by MP and the respective projection coefficients.

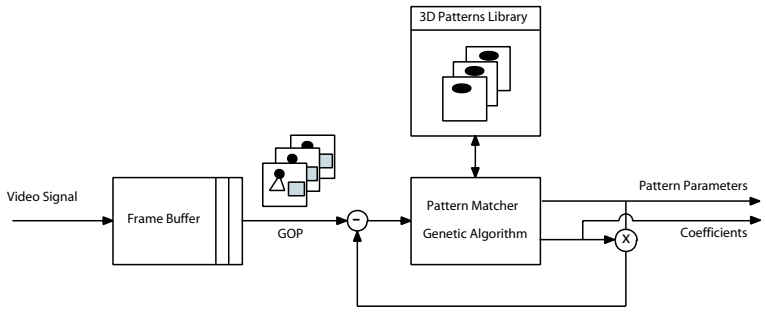


Figure 2.5: 3D Matching Pursuit video Representation [78].

3D Matching Pursuit Video Coding

Matching pursuit decompositions translate video signals into a set of parameters and coefficients, which are in charge to represent in a more or less accurate way each one of the GOPs of a sequence. 3D Matching Pursuit video coding relays on coding the description supplied by parameters (which in [78] determine positions, scaling and rotation of basis functions) and coefficients. For this purpose, efficient quantization of parameters and coefficients is required. Basis functions parameters quantization is, normally, already performed during the definition of the over-complete dictionary. *A posteriori* quantization of parameters, once selected by MP, tends to introduce very important and undesired geometric distortions (see [78]). Coefficients quantization may be performed in the MP decomposition loop (i.e. re-injecting the quantization error to the residual signal under approximation) or *a posteriori*. Studies have demonstrated that in-loop quantization reduces the distortion introduced in the final compressed representation [44]. However, efficient out-of-loop (or *a posteriori*) quantizations may be of help for flexible and progressive, fine grain, scalable coding [66, 80].

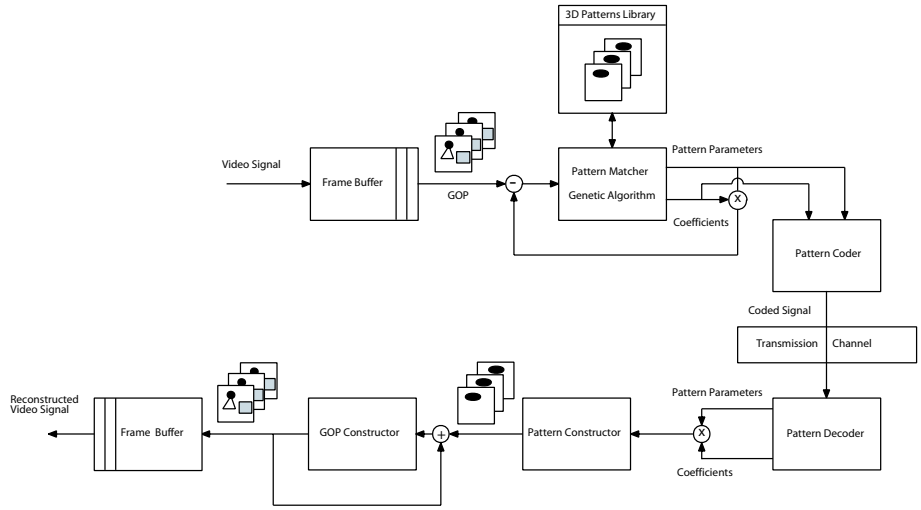


Figure 2.6: Building blocks of a Matching Pursuit video compression scheme [78].

Fig. 2.6 depicts a complete coding/decoding system based on 3D MP video decompositions. The used dictionary is available at coding and decoding sides. The parametric description, jointly with the quantized coefficients, of every GOP is entropy coded and send progressively (according to the MP decomposition order) to the receiver. At the receiver, the compressed bit-stream is progressively

decoded and the GOP is progressively reconstructed using the primitives selected by MP at the coder side. The reader will notice that coding complexity is much higher than decoding complexity. This is a clear example of an asymmetric coding/decoding scheme. Very simple terminals may be able to decode with no effort the coded video stream [78].

2.3 Exploiting Video Temporal Geometry: Modeling Motion

The temporal dimension of video signals is subject, excluding occlusions or appearing effects, to the relative motion of scene objects with respect to the camera. This motion can normally be represented using models with a limited number of parameters, leading in this way to exploit temporal redundancy.

Scene objects have in the real world three dimensions. Video cameras just capture the 2D projection of objects. This projection affects in the same way the 3D motion suffered by objects. See, for example, the well known rigid 3D rotation model [183]:

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = \mathbf{R} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + \begin{bmatrix} D_x \\ D_y \\ D_z \end{bmatrix}, \tag{2.2}$$

where (X, Y, Z) are the initial coordinates of the object, (X', Y', Z') are the newly mapped positions, (D_x, D_y, D_z) are the 3D translation and

$$\mathbf{R} = \begin{bmatrix} \cos\theta_y \cos\theta_z & \sin\theta_x \sin\theta_y \cos\theta_z - cos\theta_x \sin\theta_z & \cos\theta_x \sin\theta_y \cos\theta_z + \sin\theta_x \sin\theta_z \\ \cos\theta_y \sin\theta_z & \sin\theta_x \sin\theta_y \sin\theta_z + \cos\theta_x \cos\theta_z & \cos\theta_x \sin\theta_y \sin\theta_z - \sin\theta_x \cos\theta_z \\ -\sin\theta_y & \sin\theta_x \cos\theta_y & \cos\theta_x \cos\theta_y \end{bmatrix},$$

where $(\theta_x, \theta_y, \theta_z)$ represent the rotation angles with respect to the 3D axes. The motion produced by this 3D model projected on a camera can be fully represented by the 8-parameter *projective mapping* [124] if one of the following two conditions is satisfied: i) there is no translational motion in the camera-object axis ii) the imaged object has a planar surface ($aX + bY + cZ = 1$, a, b, c constants). The 2D *projective mapping* motion is fully described by :

$$x' = \frac{a_0 + a_1 x + a_2 y}{1 + c_1 x + c_2 y}, \quad y' = \frac{b_0 + b_1 x + b_2 y}{1 + c_1 x + c_2 y}, \tag{2.3}$$

with motion field

$$\begin{bmatrix} d_x \\ d_y \end{bmatrix} = \begin{bmatrix} x' \\ y' \end{bmatrix} - \begin{bmatrix} x \\ y \end{bmatrix}, \tag{2.4}$$

which dependents on the 5 3D motion parameters and the three plane parameters of the object surface.

Normally, objects are not planar surfaces, but may be approximated locally by planar surfaces. Hence independent *projective mapping models* (which are mappings between two arbitrary quadrilaterals) may accurately represent, with constant parameter values, sufficiently small parts of the 2D projected motion.

Sufficiently small areas of 2D projected motion can also be represented by simpler motion models than *projective mapping*. The interest of using simpler models lays on the purpose of using fewer parameters and/or to use models that have not a fractional form (unlike in Eq. (2.3)). Let us review rapidly some of these models in the following point.

2.3.1 Motion Models

Affine Motion

Orthographic projection of motion of a planar surface gives the so called affine motion model. This is a simpler model than *projective mapping* discussed previously in this section. The affine model is build up from a set of simpler transformations. These are, translation (2 parameters), rotation (1 parameter), shearing (1 parameter) and anisotropic scaling (2 parameters) [109]. Combinations of subsets of these can be considered also to build simpler motion models which are nothing but particular cases of the affine one. The affine model can be described by the following mapping equation:

$$\begin{bmatrix} d_x \\ d_y \end{bmatrix} = \begin{bmatrix} a_0 + a_1 x + a_2 y \\ b_0 + b_1 x + b_2 y \end{bmatrix} = \begin{bmatrix} a_1 & a_2 \\ b_1 & b_2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} a_0 \\ b_0 \end{bmatrix}, \tag{2.5}$$

where (a_0, b_0) represent translation and the remaining of the parameters take care, in a joint way, to describe the rest of transformations. The affine motion can be visualized, in practice as the deformation of a triangle into another, by moving the corners of the triangle [183]. Therefore, an affine motion can also be parametrized by the MVs of the three corners of a triangle.

Bilinear and Polynomial Motion

More accurate approximations of *projective mapping* exist based on *polynomial mapping* formulations. An example of this is the Bilinear motion model. This can be visualized as the warping of a quadrilateral into a curvilinear quadrilateral [98, 183].

$$\begin{bmatrix} d_x \\ d_y \end{bmatrix} = \begin{bmatrix} a_0 + a_1 x + a_2 y + a_3 xy \\ b_0 + b_1 x + b_2 y + b_3 xy \end{bmatrix}. \tag{2.6}$$

To achieve better approximations of motion mapping between two arbitrary quadrangles (i.e. *projective mapping*), higher order polynomials are needed:

$$\begin{bmatrix} d_x \\ d_y \end{bmatrix} = \sum_{\substack{0 \le i,j \le N \\ i+j \le M}} \begin{bmatrix} a_{i,j} \\ b_{i,j} \end{bmatrix} x^i y^i, \tag{2.7}$$

where N and M are arbitrary natural numbers that determine the degree of the used polynomial. However, if too many parameters are required to tune a *polynomial model*, even if *projective mapping* has a fractional form, it may be worth using directly the latter.

2.3.2 Motion Estimation Techniques

In order to determine motion mappings between two arbitrary frames, a large range of motion estimation techniques exists. All these techniques try to relate regions with similar characteristics between video frames. Ideally, one looks forward to finding the correspondences among moving regions between two frames.

This problem can be addressed in several ways. Depending on the application, one will be more appropriate than others. The different flavors of motion estimation techniques can be classified depending on the paradigm they use to model and obtain motion information. Two main approaches can be found in motion estimation.

- *Global Motion Estimation*: One may assume that the set of motions between two frames from a sequence can be characterized by a single parametric model like those reviewed in Sec. 2.3.1. In the real world, this is typically suitable when scene motion is basically due to camera movements.

- *Local Motion estimation*: When the underlying motion cannot be characterized by a simple model, then *local motion estimation* is required in order to identify the motion model that applies to each image location.

In real world scenes, local motion estimation is of key importance in order to well capture temporal video geometry.

Optical Flow

A large family of *intensity-based* approaches is the one relaying on the so-called *Optical Flow* paradigm. In this techniques, intensity changes in time are assumed to be associated to motion. Hence, temporal variations of spatial image gradients are used to compute local velocity. Optical flow techniques can supply pixel-wise motion estimations which are not subject to any limiting motion model. However, their measurement is often biased and subject to ambiguities caused by the well known *aperture problem* [183]. In [15], one can find a large review of most common optical flow techniques, together with a comparative evaluation of their reliability and performance.

Region Based Techniques

Most used techniques, mainly within the framework of video coding, are those *intensity-based* approaches that assign to selected spatial areas one of the 2D motion models discussed above. Often, this is done by dividing a frame in a regular set of partitions (e.g. in small image blocks) and by assigning to each partition a motion model and a set of parameters that determine to which region, of a reference frame, the current frame partition corresponds.

The simplest example of such an approach is the well known Block Matching (BM) algorithm. As depicted in Fig. 2.7, the so-called *Anchor Frame* is divided in blocks and a translational motion model is assumed. Then, for each one of the blocks, a motion vector is assigned by searching in a selected *Reference Frame* the block that best matches with each one of the frame partitions. Depending whether the reference frame for a particular block is in the past or in the future with respect to the anchor frame, motion estimation will be referred respectively as *backward motion estimation* or *forward motion estimation*.

Apart from the purely translational motion based BM algorithm, a whole family of Block-based algorithms can be derived by using more complex motion models. Indeed, in addition to translations, all sorts of deformations may be considered leading to a *generalized* block matching algorithm [161].

Depending on the application, independent optimization of each one of the partitions may be not suitable. This is because continuity in the motion mapping between anchor and reference frame is not guaranteed. Indeed, with no additional constraint, a best match strategy can confer very different motions to two neighboring blocks. Depending on the scene, this may be according to physical reality or, to the contrary, far from this. For some applications, the mismatch with physical reality may be a problem. In order to solve this, some regularity between neighboring frame regions needs to be imposed. For this purpose, two main approaches can be observed:

- A first approach is to constraint the commonly used least squares matching criteria to some regularity measure between motion parameters of neighboring blocks. This can be done, for

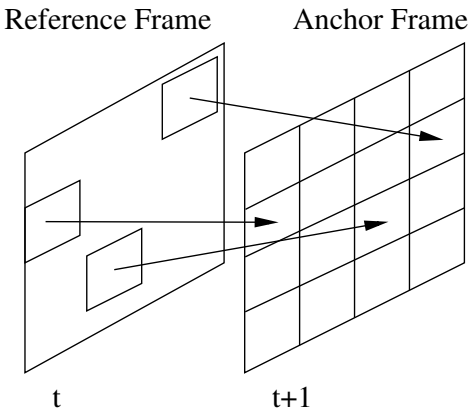


Figure 2.7: Frame at time $t+1$ is predicted by means of translated frame t pieces.

example, by means of a Bayesian formulation of the problem. One may model motion parameters as vectors of variables belonging to a Markov Random Field (MRF). Such formulation often appears when solving joint motion estimation and segmentation problems [81, 149]. Regularization constraints are also often used in optical flow estimations to solve, among others, the aperture problem [104].

- A second approach is the one known as mesh-based motion estimation. This imposes that partition boundaries shared between neighboring regions (or blocks), in an anchor frame, must remain to be shared boundaries in the reference frame as well. This may be seen from a node point of view: motion, in mesh based approaches, is represented by tracking the position of mesh nodes from the anchor frame to the reference one, keeping, at the same time, all nodes connectivity and taking care that no link between nodes crosses over another link. Typical frame partitions are based on triangular or squared elements (depending on the desired motion model: affine or bilinear). The imposed regularity poses problems in handling motion discontinuities in natural video sequences although it solves the problem of blocking effects of simple BM approaches [10, 125].

A part from intensity-based matching criteria, other measures may be considered in order to be independent of the luminance changes of the scene. These are, for example, covariance minimization or phase correlation techniques [69, 117].

A major concern of the techniques presented in here is that they are used to determine motion on discrete data. This implies that a limitation in spatial resolution exists. Matching methods have, by nature, a maximum accuracy of ± 1 pixel. To solve that, Fractional-sample-accurate motion compensation was introduced [21, 84, 85] by means of interpolation methods. Typically, bilinear filters are used to interpolate in order to achieve higher precisions of up to the 8th of a sample [186].

Another classic enhancement of nowadays motion estimation techniques is the local subdivision of regions or blocks into smaller ones in order to allow higher precision in the representation of local motion. When the motion model in use is too simple (e.g. translational motion), lower estimation errors can be achieved if motion models are assigned depending on a variable block size criteria. Hierarchical divisions of blocks in BM or triangles in meshes is a technique that is used quite often [83, 169].

Many aspects can be refined and, indeed, as can be found in literature, have been refined in order to achieve efficient motion representation in order to maximize performance in applications [170].

However, the reader may notice that most current motion estimation techniques are mainly focused on the recovery of motion fields in a pixel based representation of video signals. Indeed, motion based video representations rarely profit or take into account the spatial or spatio-temporal structure of video signals. Motion representation approaches decompose video frames in arbitrary partitions that do not respect signal structure. Efficient video representations should not uncouple spatial decompositions from motion estimation. Both aspects are linked within the video structure. Robust and flexible video signal representations require the understanding of efficient joint representation of both: spatial and temporal video features.

2.4 Motion Compensated Predictive Video Representations and Related Coding Schemes

Highly non-linear predictors are used in order to adapt the representations as much as possible to the structure of video signals. As a major tool, motion compensation is used to capture and represent efficiently temporal video geometric changes. The need for exploiting temporal geometry for efficient video representations was discovered as early as 1970s [172], but it was not until almost ten years later that was introduced the way it is used in nowadays video codecs [107]. Often in video signals, few motion parameters are able to model frame to frame changes (up to some accuracy) and, thus, supply good frame approximations that generate small residual error when used within hybrid predictive video representations.

Commonly, simple translational models together with block matching are used in predictive video coding (nevertheless, a lot of research has also been done using other kinds of representations like mesh-based representations, e.g. [125]). For each block, a non-linear R-D optimization criteria is used to determine the most appropriate model to predict the signal. Subtle modeling strategies have been introduced in predictive video coding to model accurately the signal. In addition to the MC refinements exposed in the precedent section, relevant enhancements are:

- Appropriate handling of boundaries in video frames [168].
- Multi-reference motion compensation [187], where any past frame from the GOP may be used as reference frame for video prediction.
- Multi-hypothesis weighted motion compensation [76]. Several blocks from different reference frames may be used to jointly represent as a linear combination of this a given anchor frame block [71] (This is in some way an example of the well known bi-directional prediction, which as can be seen in the next section is very related with Motion Compensated Wavelet Transforms and the use of lifting schemes).

In addition to rate constrained motion representation techniques, rate-distortion performance of modern video compression schemes is the result, also, of using:

- Intra-picture prediction techniques: which are able to somehow capture some of the 2D geometric structure in intra-frames;
- Waveform coding differences: which is in charge of the transformed representation of the residual error. This is typically done using DCT, separable wavelets or in order to exploit some geometric structure in error signals, by using redundant dictionaries together with matching pursuits [5, 132];
- Waveform coding of various refreshed regions (or transform coding of intra data).

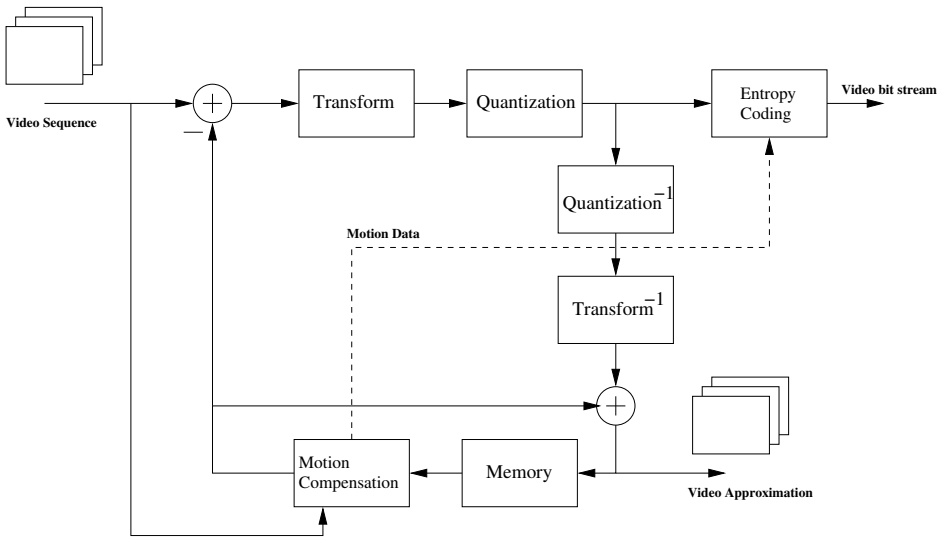


Figure 2.8: Basic block diagram of a simple predictive video coding scheme.

The need for a common framework in industry to develop video transmission and storage applications have lead the video community to create a large number of video coding standards. These have evolved through several decades thanks to technology advances, although constraint to the unavoidable economical interests of the moment. The modern basic structure for predictive video coding was first standardized in ITU-T Recommendation H.261 [1] (see Fig. 2.8) and has absolutely conditioned the structure of its successors MPEG-1 [2], H.262|MPEG-2 [3], H.263 [4], MPEG-4 Part 2 [5], and H.264/AVC [6].

One can see, here again, that common predictive techniques operate video signal in a pixel based fashion, without considering its spatio-temporal structure and partitioning it with no respect of its multi-scale structure. In the following section, spatio-temporal decompositions of video signals, based on motion compensated wavelet representations, are reviewed. Although these do not exploit spatial geometry yet, multi-scale signal structure is taken into account. This allows to combine efficient video coding with properties like scalability for progressive video streaming (see Sec. 2.1).

2.5 Motion Compensated Wavelet Transforms and Related Video Coding Schemes

Most efforts of present video coding research for efficient compression are being done toward the promising combination of linear transforms and motion compensation for non-linear video approximations. The use of motion compensation within the wavelet temporal representation, based on the lifting scheme [39, 171], has the purpose of performing the lifting filtering in the direction of motion. This motion oriented filtering drastically reduces the number of significant wavelet coefficients generated in the transform. Indeed, in this way, multi-scale redundancy can be exploited not only from those regions that remain unchanged in a period of time, but also those objects subject to motion. In the following we briefly review the well known lifting schemes used to generate one decomposition level of the motion compensated Haar and 5/3-wavelet transforms. Then, most popular video coding schemes based on motion-compensated temporal wavelets are shortly described.

2.5.1 Motion Compensated Lifted Wavelets For Video Coding

Motion-Compensated Lifted Haar Wavelet

MC lifted Haar step is widely used for video coding (e.g. [144, 158]). Fig. 2.9 depicts the ladder scheme that carries out the Haar transform in the video signal along motion trajectories. As can be seen, the prediction/update steps of this structure are performed using the samples that motion vectors connect. In the figure, $\mathbf{s}_{2\kappa}$ and $\mathbf{s}_{2\kappa+1}$ represent respectively, the even and odd input samples to the lifting scheme (in video these correspond to even and odd video frames). $\mathbf{l}_\kappa$ and $\mathbf{h}_\kappa$ represent the low pass and high pass samples at the output of the lifting scheme. The factors appearing next to the lines are normalizing factors applied once operations of prediction and update steps are performed. In [70, 75], the authors use, for the update step, the negative version of the motion vector retrieved in the prediction step.

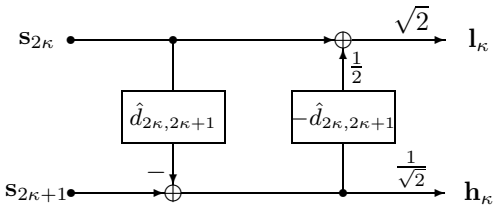


Figure 2.9: Haar transform with motion-compensated lifting steps. Both steps, prediction (with motion vector -MV- $\hat{d}_{2\kappa,2\kappa+1}$) and update (with MV $-\hat{d}_{2\kappa,2\kappa+1}$), utilize block-based motion compensation. The update steps use the negative motion vectors of the corresponding prediction steps.

Motion-Compensated Lifted 5/3 Wavelet

At this point, the MC 5/3 wavelet lifting scheme is reviewed [158]. Akin to the Haar scheme, motion vectors are decided during the prediction stage and reused in its negative form for the update step. This scheme uses a multi-hypothesis MC prediction step that looks for the two most optimal vectors that achieve the most suitable linear combination for the prediction step. This scheme performs better than the Haar one, i.e. the associated wavelet has one additional vanishing moment. See Fig. 2.10 for a schematic description.

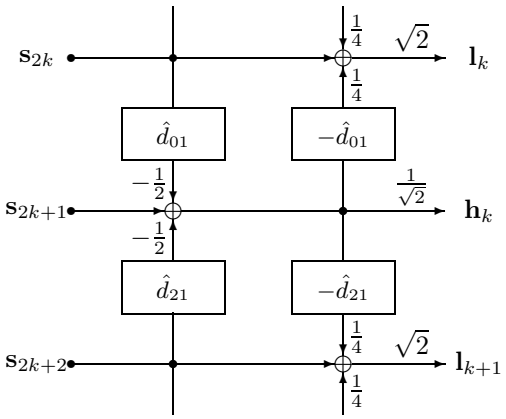


Figure 2.10: Lifted 5/3 wavelet with motion compensation. Both steps, prediction (with MVs $\hat{d}_{01}$ and $\hat{d}_{21}$) and update (with MVs $-\hat{d}_{01}$ and $-\hat{d}_{21}$), utilize block-based motion compensation. The update steps use the negative motion vectors of the corresponding prediction steps.

As one can see, the better performance Bi-directional MC and its generalization in multi-hypothesis MC, can be explained from a wavelet signals approximations point of view. Kernels build from a higher number of hypothesis are more efficient representing higher order polynomial variations in signal, and higher order polynomials approximate smooth signals better (see Chapter 3 for an analogy among MC video signals an piecewise-smooth signals).

Beyond Classic Motion Compensated Lifting Steps

Multi-hypothesis MC, used in the generation of the different MC wavelet kernels based on the lifting scheme, can be naturally extended, as in the case of classical predictive video coding, to use any even frame as reference frame in a multi-reference MC way. This was introduced by *Flierl* in [70, 75]. A short review on the subject may be found in Sec. 3.8.1.

2.5.2 Motion Compensated Temporal Filtering (MCTF)

A coding scheme, which has appeared to be successful in exploiting temporal redundancy, is based on motion compensated temporal wavelets [30, 134]. This has been possible thanks to the use of the flexible lifting scheme, that allows the inclusion of non-linear, non-invertible, operations into its ladder structure such as quantization (e.g. integer to integer transforms) or motion compensation, being still the whole scheme invertible.

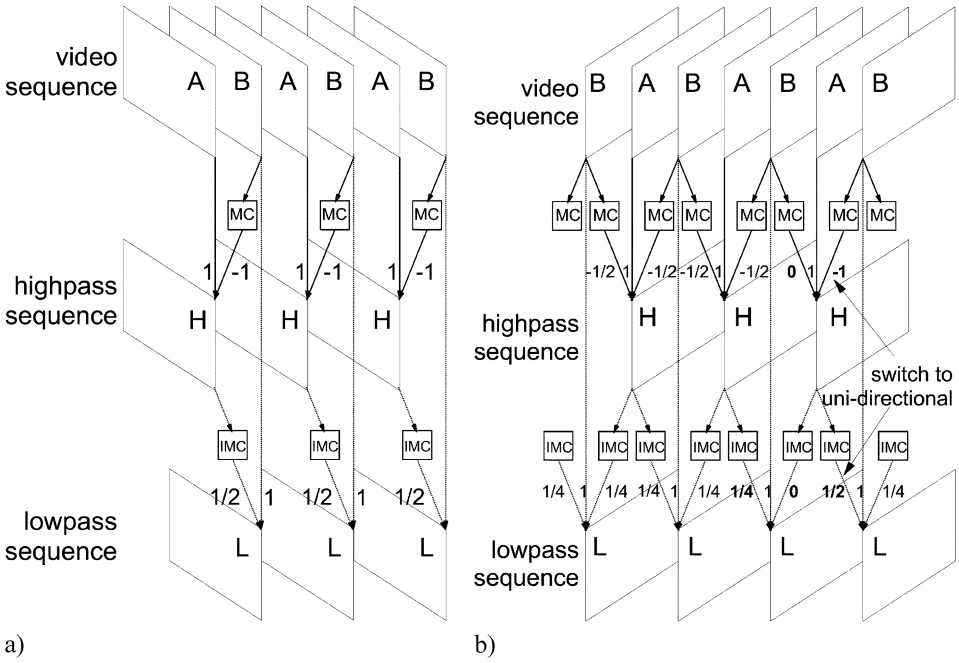


Figure 2.11: Basic motion compensated temporal wavelet transform schemes. a) Depicts the MC Haar Transform. b) Depicts the MC Daubechies 5/3 Wavelet Transform [135].

Fig. 2.11 depicts schematically the construction of two well known MC wavelet decompositions. One can see how MC is inserted in the operators that generate low and high frequency subbands in a level by level fashion. To generate high frequency bands, odd frame components are predicted with a linear combination of even frame components. In the other way round, low frequency bands are generated, as established by the lifting scheme, updating even frames with a linear combination

of the generated high frequency bands. Prediction and update steps, may benefit from most of the non-linear optimization techniques developed for efficient motion representation in predictive MC video approaches.

MCTF (t+2D): The Coding Scheme

Many works have studied the use of MCTF representations in the framework of video coding. MCTF temporal transform together with a posterior 2D wavelet transform (see Fig. 2.12) form a 3D (t+2D) representation of video signals able to capture, up to some degree, temporal video geometry and multiscale structure jointly with the multi-scale structure of temporal subbands. Video representations are composed by an hybrid mix of data, formed by motion vectors (issued from any motion estimation technique: BM, mesh-based...) that represent temporal signal geometry, and transformed coefficients that take care of capturing all the remaining components of the signal. An important detail in here is that quantization, as observed in Fig. 2.12, is performed, unlike in the predictive case, after the whole spatio-temporal video representation. The direct consequence is that progressive scalable, fine grain, coding is possible without drift problems. The interested reader may find a wide review of the approach in [73, 140, 159].

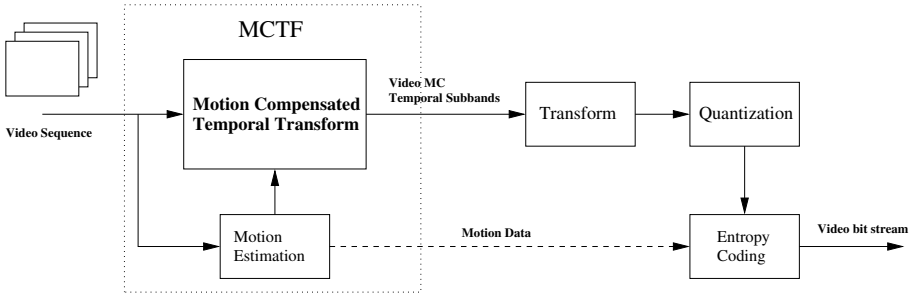


Figure 2.12: MCTF video encoder building blocks.

The MCTF coding scheme is quite flexible to allow PSNR scalability, temporal scalability and some spatial scalability. PSNR scalability is achieved thanks to a progressive transmission of video signal; depending on the transmitted amount of information, a higher or lower video quality will be displayed at the receiver. This is commonly achieved by progressive transmission of coefficients amplitude using bit-plane coding approaches. t+2D video representations require the transmission of motion vectors with full accuracy independently of the desired reception quality. This poses a problem in terms of scalability at very low bit rates. Indeed, the fixed amount of motion information overhead imposes a lower limit in the possible transmission bit-rates. The main limitation of the t+2D MCTF scheme is its lack of spatial scalability in terms of motion vector coding. Indeed, motion compensation and temporal representation is done in the space domain. Reconstruction of different resolutions at the decoder do not profit from lower requirements in motion representation accuracy.

In-band Motion Compensated Temporal Filtering (2D+t)

Another approach concerning MCTF wavelets is the one that performs, in the first place, the 2D wavelet transformation and, later, the motion compensated temporal wavelet. This approach, studied recently by several authors (see for example [12, 160]), supplies a novel framework to solve some of the limitations of the (t+2D) scheme. Indeed, it reorders the decomposition scheme such that a spatial multi-scale representation of frames is performed, with no direct influence of blocky

effects or distortions induced by MC (although this effects do not affect spatial decomposition, they may appear in the final signal reconstruction). Then, multi-scale structured 2D data is transformed in order to capture motion and temporal multi-scale structure. However, motion is not that simple to capture this time. In order to integrate spatial resolution requirements and progressive motion accuracy transmission, motion estimation is performed in the transformed wavelet domain (see Fig. 2.13).

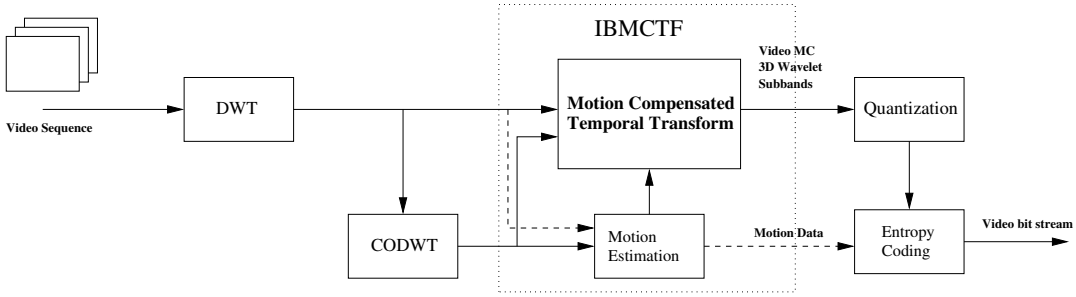


Figure 2.13: IBMCTF video encoder building blocks.

As one may expect, the use of critically sampled discrete wavelet transforms in space imposes a problem of representation variance with signal translation (i.e. the same signal slightly translated generates a whole collection of different coefficients [121]). In order to solve that, motion compensation is performed considering this variance. Over-complete, non-decimated wavelet frame representations are used to retrieve the appropriate reference coefficients for temporal lifting steps updates and predictions.

Data structuring allows, in this case, the embedded transmission of motion and spatio-temporal signal components, leading to a fully embedded stream of information that allows lower bit-rates than in the t+2D MCTF case.

2.6 Conclusions

In this chapter a brief introduction to present strategies for video representations and video coding is given. The fact that signal adapted representations are cruzial for flexible coding with efficient *R-D* performance is underlined. In order to exploit as much as possible signal structure, video signal decompositions have evolved up to 3D decomposition approaches where video is represented by the superposition of 3D primitives which are able to jointly model information about motion and spatio-temporal scale signal components (i.e. spatio-temporal motion oriented components). However, most of the reviewed approaches ignore a very relevant aspect in structural redundancy. Spatial geometry, one of the most relevant structural components in images and video, is not taken into account. Only one of the reviewed approaches takes spatial geometry into account. This is the approach based on Matching Pursuits. Although it does not exploit temporal geometry, it is capable to supply fine grain PSNR scalability as well as spatio-temporal scalability with better performances than non-motion compensated 3D wavelets and some predictive scalable approaches.

Chapter 3

Introducing Adaptivity in Wavelet Video Codecs

3.1 Motivation

Interest in sub-band video coding has been motivated during the last years due to its suitability for certain video streaming applications. Scalability, low computational cost with a reasonable R-D performance and the possibility to set robust delivery on lossy channels are among its features. Popular examples of 3D wavelet based coding schemes are 3D-SPIHT [29], ESCOT [189] (which is the 3D version of EBCOT, used in the 2D case to compress images with the JPEG2000 standard [8]) or the 3D extension of GTW [103] which intents to be a generalization of EBCOT.

Progressively, motion-compensation has also been included in wavelet based approaches as they combine excellent compression efficiency with embedded representation capabilities. A relevant contribution to the field has been motion-compensated lifted wavelet transforms [134, 144]. To improve compression efficiency, adaptive lifting schemes have been investigated for motion-compensated temporal filtering (MCTF). In particular, frame-adaptive motion-compensated lifted wavelets [70] and multi-hypothesis motion-compensated lifted wavelets [75] have been proposed in the framework of a MCTF extension of H.263++ [72]. Further improvements have been accomplished with the MCTF extension of H.264/AVC [157].

In the case of a simple separable wavelet transform scheme, some R-D performance enhancement is possible if a non-linear decomposition scheme is used. In this chapter, a locally adapted temporal transform is theoretically and empirically analyzed based on a piecewise-smooth model of video signals. A locally adapted temporal approach may reduce, up to some degree, the number of wavelet coefficients needed to represent a singularity. Although it will not dramatically change the decay of distortion with rate, this solution gives some additional degrees of freedom to displace the D-R curve and obtain better performances than with standard dyadic wavelets. Unlike fixed length wavelet transforms, *best basis* like transforms are able to adaptively set the analysis scale (or window length) that better suits the signal to be represented. Long smooth pieces are grouped into long transformed segments while fast signal variations tend to be localized. Temporal adaptivity is also useful in the case of Motion Compensated wavelet representations. In this chapter, an analogy is also presented between MC video sequences and piecewise-smooth 1D signals. Theoretic R-D results of oracle based methods, to code piecewise-smooth 1D signals [148, 180], justify our proposed Intra-

adaptive approach for MC lifted wavelets for video coding. A large number of results validate this analogy and demonstrate, in practice, the benefits of using temporal adaptivity in MCTF.

This chapter is structured as follows: First, a set of simple examples are presented in Section 3.2 to illustrate the temporal adaptivity problem. After, we review in Section 3.3 the theoretical background of non-linear approximations of 1D piecewise-smooth signals based on wavelet transforms. Based on this and using a synthetic deterministic model of a moving edge, R-D bounds are derived for different 3D wavelet transform schemes in Section 3.4. These theoretical results together with several practical experiments presented in Section 3.6 illustrate the influence of temporal adaptivity within simple 3D wavelet video coding. After this first analysis, intra-adaptive wavelets are investigated within the MCTF framework. In Section 3.7, an analogy between motion compensated coding of video signals and coding of piecewise-smooth 1D signals is established. Adaptive motion-compensated lifting schemes can be extended by the use of intra macroblocks, as proposed in Section 3.8. They allow separate encoding of intervals with smooth motion trajectories. Frame-adaptive motion-compensated lifted wavelets [70] permit a flexible encoding of motion trajectories as long as sufficient candidate reference frames are available. The number of efficient reference frames decreases not only due to encoding constraints but also due to scene changes and frequent object occlusions. If frame-adaptive motion-compensated wavelets are not able to provide the desired flexibility, intra macroblocks are used to permit separate encoding of intervals with smooth motion trajectories. A detailed analysis of the effect on R-D performance of intra-adaptivity is presented in Section 3.9. Finally, conclusions are drawn in Section 3.10.

3.2 Coding: To Join or not to Join?

Wavelet decompositions present an optimal R-D behavior when applied to approximate polynomial signals if wavelets with enough vanishing moments are used [121]. However, singularities between polynomial pieces generate many wavelet coefficients that are costly in terms of coding rate.

Coding improvements are introduced, here, for some simple examples when the number of decomposition levels is adapted. In the following examples, sequences of two consecutive images are coded. Hence, only one level of temporal dyadic wavelet decomposition can be considered at most. In all cases, a 5 level dyadic wavelet spatial decomposition is performed by means of the Daubechies-9/7 filters [14]. We compare between a Haar transform in the temporal dimension (i.e. a one level Haar transform) and no temporal transformation. Uniform dead-zone quantization is applied to wavelet coefficients. Coding cost is measured by means of their Shannon entropy.

Temporally aligned video pixels that, due to the effect of some motion, belong to different objects, have often no relation among them. If these are modeled as being a set of IID Gaussian random variables, then there would be no change in the coding efficiency with or without applying an orthonormal transform. However, this is not the case for the image sequences presented in here. They have a spatio-temporal structure and temporal changes (e.g. produced by motion) follow a model that does not correspond to a set of IID Gaussian random variables. Although signals are often modeled by jointly Gaussian stochastic models, real images and video have a quite different behavior. In fact, deterministic piecewise-constant models [148] suit better our purposes.

As theoretically proved, from a 1D point of view (see Sec. 3.3.2), if a temporal pixel has an edge (important gray level change in the temporal dimension), wavelet transform coefficients will be spread throughout all subbands. Hence, coding will be inefficient in terms of R-D. On the other hand, if a temporal pixel stays more or less unchanged, energy will be compacted and the pixel will be efficiently coded.

3.2.1 Coding Very Different Pictures

Consider the two images of Fig. 3.1 as if they were a sequence. This situation can be seen as an extreme case of a scene cut. We see that there is apparently no significant redundancy to exploit between them. In Fig. 3.2, one can appreciate that coding without orthonormal Haar transform

Figure 3.1: Lenna (left) and Barbara (right) 256x256 pictures.

is more efficient in terms of R-D. It can be easily seen that, having such different images, joined temporal coding implies a significant loss in efficiency. This behavior is observed in a large range of coding bit-rates. Nevertheless, at very low bit-rates (very high distortion) the difference becomes negligible due to the very high quantization noise.

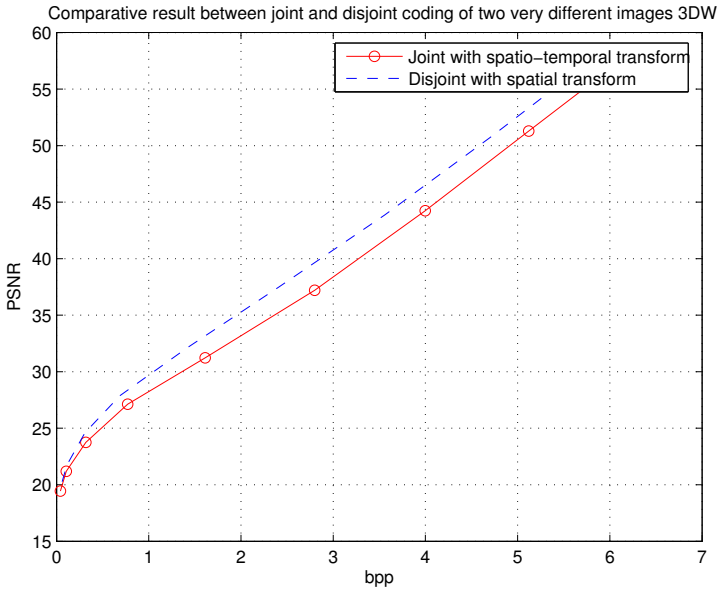


Figure 3.2: Demonstration of an artificial scene cut: Coding the images Barbara and Lenna jointly with a Haar transform or independently.

3.2.2 Coding a Scene Cut

Let us consider, now, a more common video sequence: the table tennis sequence (Fig. 3.3). Two very similar frames can be efficiently decorrelated by a wavelet transform (let us forget, for the moment, about the benefits of using in addition motion information). This is underlined by the left

Figure 3.3: Table tennis sequence frames 129, 130 and 131.

graphic in Fig. 3.4, where results of jointly coding frames 129 and 130, using a Haar transform are shown. This is more R-D efficient when compared to the case where no temporal transformation is performed. In the right plot of Fig. 3.4, another particular example is analyzed. This time a change of scene is taking place (frames 130 to 131) and this gives rise to a situation similar to the *Lenna/Barbara* sequence (Sec. 3.2.1). Again, as the reader can from the curves, joint coding of both pictures is not R-D efficient. In order to be optimally adaptive in the temporal wavelet

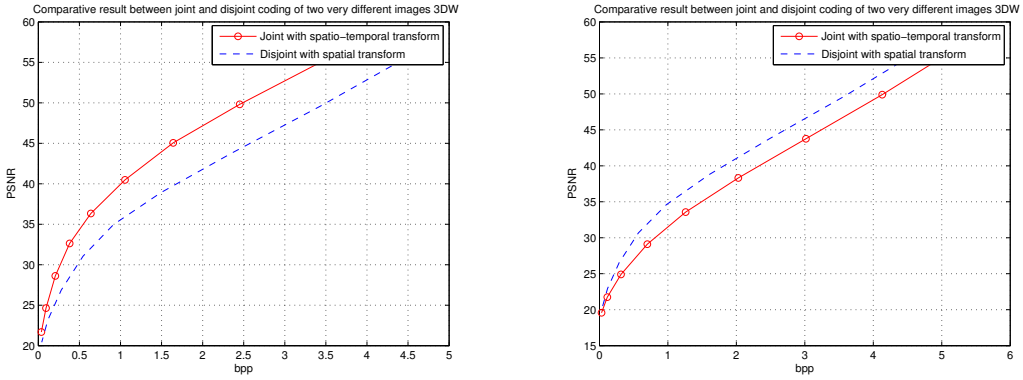


Figure 3.4: R-D efficiency of one level of temporal wavelet decomposition for two different scene events from the sequence table tennis. Left: No change of scene (frames 129 and 130). Right: Change of scene (frames 130 and 131).

decomposition depth, the number of wavelet subbands should locally adapt in space. In this way, only when local unpredictable changes appear in a sequence of images, the wavelet decomposition used to jointly code them can be optimally adapted for that particular scene and still profit from a high number of wavelet decomposition levels in the remaining spatial locations.

3.3 Deterministic Signal Models for R-D Analysis

This section recalls the use of deterministic models for R-D performance analysis of non-linear approximation algorithms. An emphasis is put on edge modeling and their influence on wavelet

based coding. A piecewise-constant simple model for images is reviewed and an extension for moving edges in natural video sequences is proposed.

3.3.1 Use of Deterministic Models for R-D Analysis

Signals are often associated to stationary jointly Gaussian stochastic models independently of their nature. However, in many cases, these models do not match the real nature of signals and prevent from properly exploiting their characteristics. In this direction, a lot of works have been done in the field of non-linear signal approximations and their relation to signal compression. A performance analysis of non-linear approximations needs to be performed on the particular class of signals to be compressed. For this purpose, deterministic models representing the kind of signal events to study are considered. This approach was taken in [32, 148, 180] for piecewise-smooth signals, non-linear approximation on wavelet bases and other oracle based approaches. A detailed rate distortion study about the appropriate way of coding edges is described in [148] for the 1D case.

3.3.2 Edge Modeling in 1D Signals: Piecewise-x Models and Wavelet Coding

There are many examples [31, 51, 148] where signals can be associated to deterministic signal models allowing a better analysis of the properties of their representations within the framework of a given application, e.g. for coding purposes. The deterministic model that better fits the behavior of signals with discontinuities is the so called *piecewise-smooth* signal model [31] (see Fig. 3.5).

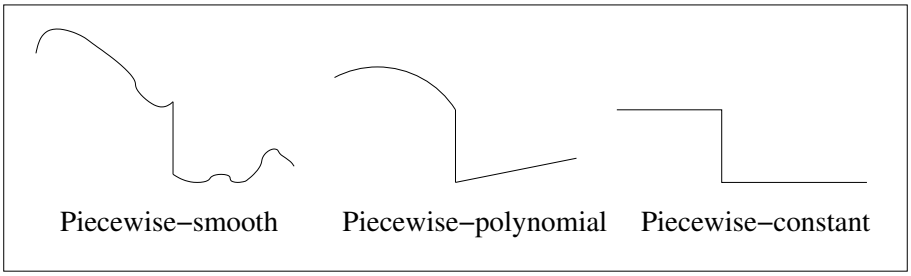


Figure 3.5: Example of 1D piecewise-smooth, piecewise-polynomial, piecewise-constant signals.

Fig. 3.5 depicts an example of a piecewise-smooth 1D signal, as well as two more examples that represent two particular cases of the piecewise-smooth class of signals. These particular cases are the piecewise-polynomial and the piecewise-constant ones. As depicted, these signals are composed, respectively, by polynomial and constant signal intervals separated by singularities. This class of signals has been sown to be efficiently represented by wavelets, due to their locality which capture well abrupt signal changes. Moreover, smooth or polynomial parts are efficiently represented by coarse approximations obtained from the wavelet basis scaling functions. Nevertheless, wavelet transforms do not exploit the correlation among wavelet coefficients from different subbands generated by an edge (see Fig. 3.6). Thus, even though the wavelet transform is well suited for representing discontinuities, as discussed in the remaining of this section, it reveals to be suboptimal in terms of R-D.

Fig. 3.6 illustrates the response of a wavelet transform to a step function. If a wavelet with sufficient vanishing moments [121] is used, each polynomial area can be represented with coefficients belonging to the low frequency band (the scaling functions). In such a case, the only part of the signal that generates non-zero wavelet coefficients are discontinuities. To accurately code the step using a

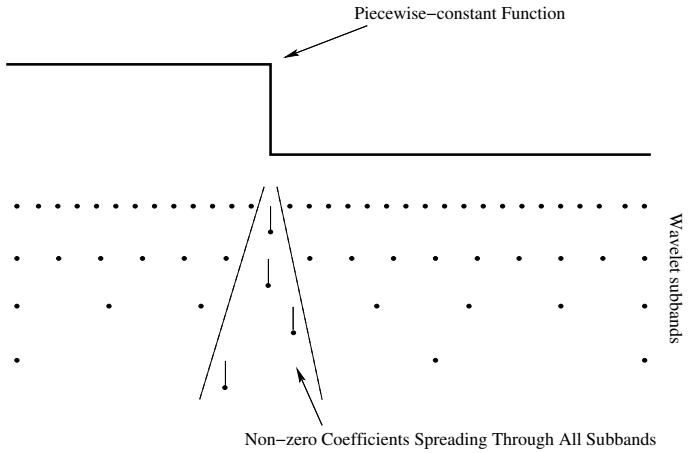


Figure 3.6: Spreading of coefficients through the wavelet subbands of a 1D piecewise-constant signal representation.

wavelet transform, non-zero coefficients amplitudes and positions need to be coded. On the other hand, Fig. 3.7 shows a non-linear approach widely discussed in approximation theory [32, 148, 180]. This intends to find out a more efficient representation of piecewise signals in general, assuming the existence of an oracle that tells where switching points among smooth pieces are located. If this is the case, since very efficient approximations of smooth intervals and efficient coding of discontinuity locations can be achieved, it is possible to obtain a better R-D behavior than in the case where only wavelets are used. It is, indeed, more efficient to separately code discontinuity locations and smooth parts. For instance in Fig. 3.7, in order to locate the edge and to set its size, it is just necessary to supply one position plus one amplitude. Moreover, the use of an independent representation in each constant interval will not generate additional information to code, i.e., consider a Haar wavelet that is used in each one of these (with an appropriate handling of boundaries and scaling functions coefficients, see Sec. 3.8.3 for an example). Then, no non-zero coefficients will be generated. To the contrary, in the simple 1D wavelet case, the number of locations and amplitudes to code is proportional to the number of decomposition subbands.

In oracle based coding of piecewise-polynomial signals, the asymptotic behavior of distortion (D) at high rates is described as a function of rate (R) [148]. This can reach the bound:

$$D_O(R) \sim 2^{-B \cdot R}, \tag{3.1}$$

where B is a positive constant. In the case of wavelet coding, the asymptotic behavior at high rates is worse:

$$D_W(R) \sim \sqrt{R} \cdot 2^{-A\sqrt{R}}, \tag{3.2}$$

where A is a positive constant. Unlike in (3.1), distortion decreases with exponent $\sqrt{R}$, which corresponds to a slower decay with the rate.

Notice that even if the asymptotic R-D behavior is analyzed at high rate, it is sufficient to motivate the use of adaptive coding [148, 180], and to understand the coding efficiency of different approximation approaches (e.g. [51, 148]).

3.3.3 A Model for R-D Analysis of Edge Compression in Natural Images

Images are typically modeled as the composition of a piecewise-smooth component plus a texture component. Most of the image information is carried in edges within the piecewise-smooth part

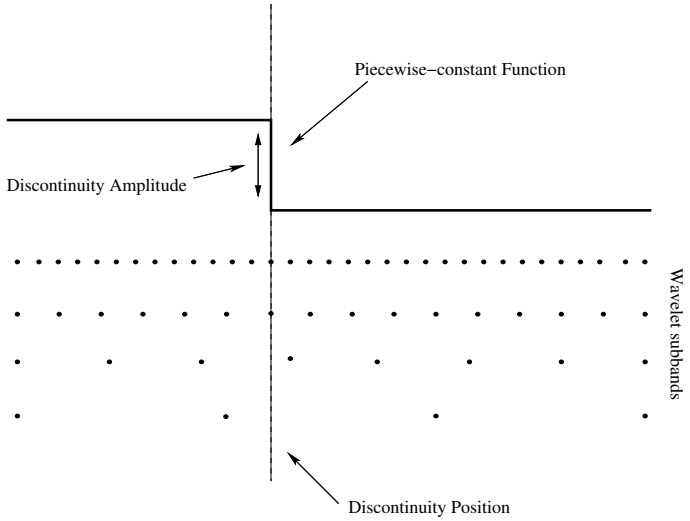


Figure 3.7: There are no wavelet coefficients to code from a 1D piecewise-constant signal representation when using an oracle to code the discontinuities, i.e. position and amplitude.

[129, 179]. As in the 1D case, toy models are proposed to analyze different coding strategies based on non-linear approximations. In the case of images with edges, these are piecewise-constant synthetic images with smooth C^p edges or piecewise-polynomial edges [51, 164]. A well known example is the *Horizon* model. This was used in [53] and [51] to justify the need for exploiting geometry in natural images. Although geometrical representation of video signals is not addressed until ulterior chapters, we use the *Horizon* model in the R-D analysis presented here.

The *Horizon* model is an image $f(x, y)$ defined on the unit square $[0, 1] \times [0, 1]$, i.e. $x, y \in [0, 1]$, such that

$$f(x,y) = \begin{cases} 1 & y \geq b(x) \\ 0 & \text{otherwise} \end{cases}, \tag{3.3}$$

where $b(x)$ has finite length inside the unit square and belongs to the class of functions that are p-times continuously differentiable (i.e., $b(x) \in C^p$). See Fig. 3.8 for a graphical example.

3.3.4 A Moving Edge Model for Video Sequences

In this section a piecewise-constant deterministic model of a video sequence is introduced. In some essence, video objects are smooth regions surrounded by edges. To represent moving edges with a simple model, we use the *Moving Horizon*. That is, a 3D signal composed of a sequence of *Horizon* images where the edge moves from frame to frame at a given speed. Fig. 3.9 depicts six sample frames of the whole *Moving Horizon* sequence.

3.4 3D vs 2D+1D Temporal Adaptive Wavelet Transforms for Video Coding

In this section, a theoretical evaluation of the influence of temporal adaptation on moving edges is presented. For this purpose, the deterministic piecewise-smooth model for natural video signals is used to measure the R-D performance of different 3D wavelet decomposition schemes.

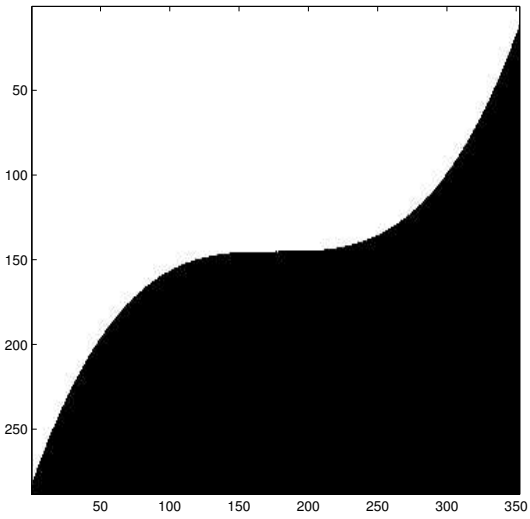


Figure 3.8: Example of Horizon toy model image [51].

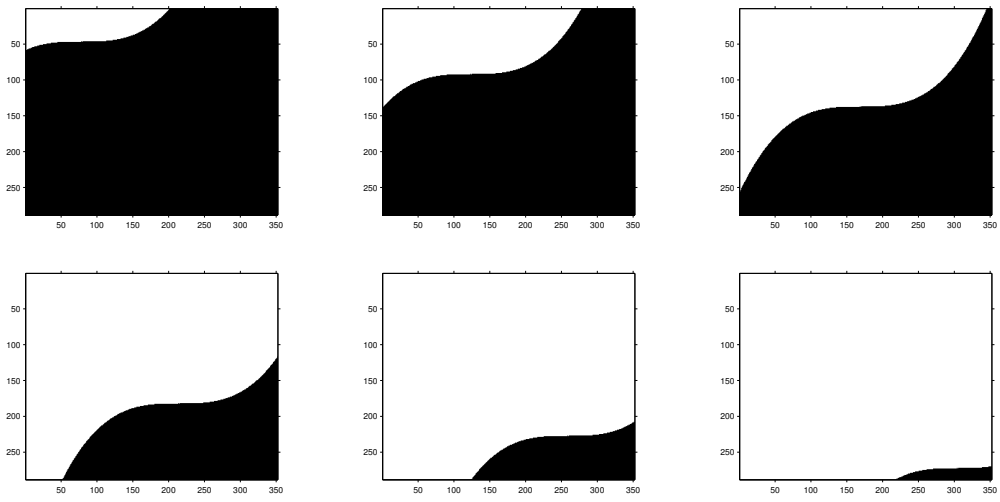
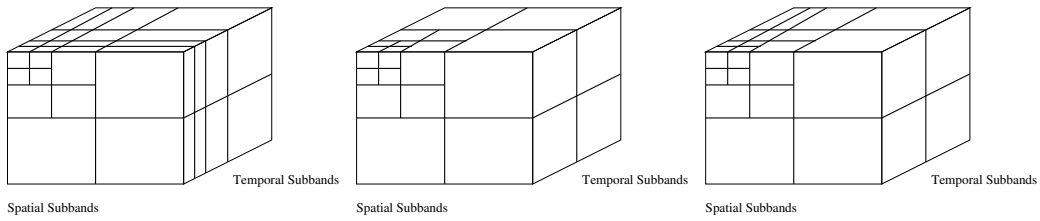


Figure 3.9: Sample of the moving horizon sequence. A smooth curve is moving through time with a given motion vector. The temporal sequence goes from left to right and from top to bottom.

3.4.1 Adaptivity and Time-Frequency Tiling for Video Decompositions

The classical 3D wavelet decomposition scheme used for subband video coding (see the left graphic in Fig. 3.10) is mainly conceived for exploiting the temporal redundancy of static areas. This scheme is inefficient to represent moving edges since temporal discontinuities generate many coefficients over all temporal subbands. Temporal discontinuities get little coding profit from being decomposed on a wavelet basis. Hence, they should be decomposed as little as possible while non-moving areas are projected on a wavelet basis with as many subbands as possible. These opposed requirements impose the use of an adaptive scheme in order to globally optimize the R-D performance of the coding scheme. Temporal discontinuities need to be represented by means of high level signaling (oracle based method), more appropriate to indicate which are their locations. This adaptive temporal representation may be seen as a kind of *Best Basis* decomposition where the R-D performance is optimized [153, 154].



(a) Classic **(b)** Isotropic **(c)** Unequal

Figure 3.10: Three possible different subband schemes for video representation. (a) Classical scheme where the video signal is assumed to remain mostly static. (b) The 3D isotropic, tree structured subband scheme (extension of the common 2D one used for image compression [8]). This scheme is adapted for general 3D volume figures. (c) 3D subband scheme adapted for the representation of the moving horizon toy model of Fig. 3.9.

In order to evaluate how well (in terms of R-D performance) a moving edge can be represented, we analyze the performance of different decomposition schemes (see Fig. 3.10). These fixed schemes exploit the structure of particular 3D signals. Hence, they do not behave in the same way with the *Moving Horizon* model. To guarantee a precise comparison, in the following, upper bounds have been computed for schemes that are better for coding moving edges while a lower bound on the R-D is derived for the classic separable wavelet scheme for video coding.

3.4.2 Isotropic 3D Wavelet Coding of the Moving Horizon Model

The isotropic 3D Wavelet decomposition scheme (see middle of Fig. 3.10) is typically used to decompose 3D volumes when no particular direction predominates in the signal. This particular choice, as seen in the following, is due to the better R-D behavior of isotropic separable wavelet transforms compared to other linear wavelet decomposition schemes when applied to this kind of signals.

The isotropic 3D wavelet scheme is characterized by having a fixed number of subbands per 3D isotropic decomposition level: $2^3 - 1$. Thus, the total number of subbands (including the lowest frequency band) for J decomposition levels is

$$N_{sb} = 1 + \sum_{j=0}^{J} (2^3 - 1) = 7J + 8.$$

For example, 5 wavelet spatio-temporal decomposition levels generate a total of 36 spatio-temporal subbands.

Considering this decomposition scheme and the *Moving Horizon* model, the following can be stated:

Theorem 3.1 *Assume the use of a coding scheme based on a 3D isotropic wavelet decomposition, scalar quantization and efficient coefficients position coding. Then, its R-D performance (D(R)) achieved when coding the* Moving Horizon *model can be upper bounded as:*

$$D(R) \lesssim \frac{1568 \cdot 2^{1/3}}{3} \sqrt{\frac{W(E(R))}{E(R)}} - \frac{262144}{3} \sqrt{\frac{W(E(R))^3}{E(R)^3}}, \tag{3.4}$$

(where $E(R) = 1024/21 \log(2) 2^{2/3}(56 + 3R)$ and $W(x)$ is the principal branch of the Lambert W-function [35, 185]) whose asymptotic behavior at high rates is:

$$D(R) \sim \sqrt{\frac{\log(R)}{R}}. \tag{3.5}$$

The reader will notice that Eq. (3.5) clearly reflects the lack of efficiency of separable wavelets to represent singularities along edges (see [51] for a comparison in the case of 2D signals). This must be addressed by some approach that takes into account geometric information. An example is the use of motion compensation to better exploit geometry in the temporal dimension (see Sec. 3.8).

Proof: To prove Theorem 3.1, let us consider the use of non-overlapping Haar wavelets. At each decomposition level j, the number (n_j) of non-zero wavelet coefficients, generated by the 3D transform, corresponds to the number of basis functions that intersect with the moving edge. As the *Moving Horizon* model is a finite length signal, the number of non-zero coefficients is a function of the size of the support of the wavelet basis functions (2^{-j}). Since the signal to represent is a 2D manifold, a total of 2^{2j} wavelet basis functions per subband is required to cover the surface drawn by the moving edge. Hence, given that each decomposition level is composed of seven subbands,

$$n_j \sim 7 \cdot 2^{2j} \tag{3.6}$$

wavelet basis functions are needed at the jth decomposition level. Consider the whole 3D separable transform (including the coarse scale approximation), the total number of non-zero coefficients may be estimated as:

$$N_J \sim \sum_{j=0}^{J} n_j + 1 = \sum_{j=0}^{J} 7 \cdot 2^{2j} + 1 = \frac{28}{3} \cdot 4^J - \frac{4}{3}, \tag{3.7}$$

where J is the finest scale used to approximate the *Moving Horizon* model (linear approximation). Since both, Distortion (D) and Rate (R), depend on N_J, an estimate of each one may be easily derived.

Non-zero wavelet coefficients, generated by a 3D separable, dyadic and normalized basis, have an amplitude decay at the jth level behaving like

$$|c_{j,\mathbf{k}}| \sim 2^{-\frac{3j}{2}}, \tag{3.8}$$

where $\mathbf{k}$ indexes coefficients within each subband [51, 121]. Indeed, the finer the scale of analysis is, the smaller the amplitude of coefficients becomes. For a R-D scalar optimized quantization of normalized wavelet transform coefficients, all subbands are quantized with the same quantization step size $\Delta_j = \Delta \ \forall j \in [0, J]$. If one intends to code all coefficients up to level J, the quantization step size has to be small enough, i.e. $\Delta \sim 2^{-\frac{3J}{2}}$. That is,

$$R_{c_{j,\mathbf{k}}} \sim \log_2 \left(\frac{1}{\Delta} \right) = \frac{3J}{2} \tag{3.9}$$

bits are necessary to independently code each wavelet coefficient with a distortion of

$$D_{c_{j,\mathbf{k}}} \sim \Delta^2 \sim 2^{-3J}. \tag{3.10}$$

Now, total R and D may be estimated in terms of the finest detail decomposition level J.

The total rate is composed by the cost of coding the whole set of non-zero coefficients N_J plus the total number of bits required to specify the location of non-zero coefficients in the subbands issued from the 3D transform. For example, considering a classical oct-tree organization of coefficients, where they are associated in a parent-children manner [31, 113], a total of eight additional bits per coefficient needs to be considered (in the case of the low frequency band, these bits are not required). The use of an oct-tree allows an efficient description of children sub-trees significance. Thus,

$$\begin{aligned} R &\sim N_J \cdot R_{c_{j,\mathbf{k}}} + (N_J - 1) \cdot 8 \\ &= \left(\frac{28}{3} \cdot 4^J - \frac{4}{3}\right)\frac{3J}{2} + \left(\frac{28}{3} \cdot 4^J - \frac{7}{3}\right) \cdot 8 \\ &= \left(14J + \frac{224}{3}\right)4^J - 2J - \frac{56}{3}. \end{aligned} \tag{3.11}$$

The total distortion results from the non-linear approximation introduced by coefficients quantization plus the truncation of the wavelet expansion at the Jth level. According to this, two main terms appear in the distortion expression:

$$\begin{aligned} D &\sim N_J \cdot D_{c_{j,\mathbf{k}}} + \sum_{j=J+1}^{\infty} n_j \cdot |c_{j,\mathbf{k}}|^2 \\ &= \left(\frac{28}{3} \cdot 4^J - \frac{4}{3}\right)2^{-3J} + \sum_{j=J+1}^{\infty} 7 \cdot 2^{2j} \cdot 2^{-3j} \\ &= \frac{49}{3}2^{-J} - \frac{4}{3}8^{-J}. \end{aligned} \tag{3.12}$$

Finally, an estimate of the asymptotic $D(R)$ behavior is obtained by combining (3.11) and (3.12).

First, solving J from (3.11) implies a transcendental equation. Thus, we upper bound the rate, which for large values of J and given the exponential nature of the expression, turns to be an accurate approximation:

$$\begin{aligned} R &\sim \left(14J + \frac{224}{3}\right)4^J - 2J - \frac{56}{3} \\ &\lesssim \left(14J + \frac{224}{3}\right)4^J - \frac{56}{3}. \end{aligned} \tag{3.13}$$

This can be turned into:

$$J \gtrsim \frac{1}{6\log(2)}\left(3\ \mathrm{W}\left(\frac{1024}{21}\log(2)2^{2/3}\left(56 + 3R\right)\right) - 32\log(2)\right), \tag{3.14}$$

where $x = \mathrm{W}(y)$ is the principal branch of the Lambert W-function [35, 185], i.e. the solution to the inverse function of $y = xe^x$ that is analytic at 0 [35, 185]. A lower bound on the rate-distortion behavior of the isotropic 3D wavelet coding scheme is derived by replacing (3.14) in (3.12):

$$D(R) \lesssim \frac{1568 \cdot 2^{1/3}}{3}\sqrt{\frac{\mathrm{W}\left(E(R)\right)}{E\left(R\right)}} - \frac{262144}{3}\sqrt{\frac{\mathrm{W}\left(E(R)\right)^3}{E\left(R\right)^3}}, \tag{3.15}$$

where $E(R) = 1024/21\log(2)2^{2/3}(56 + 3R)$.

Under the high rate assumption, Eq. (3.15) may be characterized by a simpler expression. This will help us to understand the general R-D behavior of the present coding scheme and to compare with the rest of 3D decomposition approaches in the following. $\mathrm{W}(E(R))$ can be accurately approximated by $\log(E(R))$ for big R. Moreover, at such high rates, the approximation performed in (3.13) becomes negligible, as well as the influence of the second term in Eq. (3.15) with respect to the first one. Hence, together with the linearity of $E(R)$ with respect to R, the $D(R)$ high rate asymptotic behavior of the isotropic 3D wavelet decomposition scheme to code the *Moving Horizon* model is:

$$D(R) \sim \sqrt{\frac{\log(R)}{R}}, \tag{3.16}$$

which proves Theorem 3.1.

■

3.4.3 Classic 3D Wavelet for Video Coding of the Moving Horizon Model

In this subsection the Classic 3D Wavelet packet used in video coding (see left of Fig. 3.10) is analyzed in the particular case of the *Moving Horizon* Model. It is well known that this scheme is optimized to profit from static signal regions. Hence, being more efficient to code static areas, a poorer R-D behavior is, thus, expected for those signal parts made of moving edges.

This 3D transform may be seen as a 2D+1D wavelet decomposition scheme. First, a 2D isotropic dyadic decomposition is applied on each video frame. Then, these are further transformed in the time direction by means of a 1D dyadic wavelet transform.

Compared to the isotropic 3D wavelet decomposition, there is an increase in the number of decomposition subbands. Indeed, there are $3 \cdot (J+2)$ spatio-temporal subbands at every spatial decomposition level and $J+2$ temporal subbands at the spatial scaling function level. This makes a total of

$$N_{sb} = (J+2) + \sum_{j=0}^{J} 3 \cdot (J+2) = 3J^2 + 10J + 8.$$

In the case of 5 temporal subbands are performed, and up to 5 spatial wavelet decomposition levels, a total number of 96 subbands are generated.

Considering all that, the following can be stated:

Theorem 3.2 *Assume the use of a coding scheme based on a Classic Packet 3D Wavelet used in video coding, scalar quantization and efficient coefficients position coding. Then, its R-D performance ($D(R)$) on the coding of the* Moving Horizon *model can be lower bounded as:*

$$\begin{aligned} D(R) \quad \gtrsim \quad & \frac{32}{3\log(2)} \frac{\left(3\,W(E(R)) - \dfrac{11}{\log(2)}\right) 2^{2/3}\, W(E(R))}{E(R)} \\ & + \frac{424 \cdot 2^{2/3}}{3} \frac{W(E(R))}{E(R)} + \frac{8192}{3} \left(\frac{W(E(R))}{E(R)}\right)^3 \\ & - 512 \cdot 2^{1/3} \left(\frac{W(E(R))}{E(R)}\right)^2, \end{aligned} \tag{3.17}$$

(where $E(R) = \frac{4}{3}\log(2)\left(3 \cdot 2^{2/3} + \sqrt{2}\sqrt{2^{1/3}\,(-7+6R)}\right)$) whose asymptotic behavior at high rates is:

$$D(R) \sim \frac{\log^2(R)}{\sqrt{R}}. \tag{3.18}$$

A proof may be found in Appendix A.1.

Compared to the bounds and asymptotic behavior derived for the 3D isotropic case, we can see that the present scheme behaves worse for coding moving edges. A more graphical evidence of this is presented in Fig. 3.11.

3.4.4 Unequal Temporal Partition 3D Wavelet Coding of the Moving Horizon Model

Keeping the 2D+1D structure of the classic packet wavelet, another decomposition scheme may be explored in order to get closer in performance to the 3D isotropic one. We name the present 3D transform scheme, the Spatially Unequal Temporal Partition 3D Wavelet Scheme (see right of Fig. 3.10). It is characterized by the adaptation of the number of temporal subbands depending on the spatial wavelet subband.

In effect, for each spatial decomposition level j, there are $3\,(J+2-j)$ spatio-temporal subbands. Furthermore, $J+2$ temporal subbands appear at the lowest spatial scaling function level. Hence, the total number of subbands for J decomposition levels is

$$N_{sb} = J + 2 + \sum_{j=0}^{J} 3\,(J+2-j) = \frac{3}{2}J^2 + \frac{17}{2}J + 8. \tag{3.19}$$

In the case where a scheme of 5 temporal subbands is used, and up to 5 temporal wavelet decomposition levels are allowed, a total of 66 subbands is generated.

Considering all that, the following can be stated:

Theorem 3.3 *Assume the use of a coding scheme based on a Spatially Unequal Temporal Partition 3D Wavelet decomposition, scalar quantization and efficient coefficients position coding. Then, its R-D performance ($D(R)$) coding the* Moving Horizon *model can be upper bounded as:*

$$D(R) \lesssim 304 \cdot 2^{5/6} \frac{\sqrt{W\left(2\log(2)E(R)^2\right)}}{\sqrt{\log(2)}E(R)} - 2048 \cdot 2^{2/3} \frac{W\left(2\log(2)E(R)^2\right)}{\log(2)E(R)^2} \tag{3.20}$$

(where $E(R) = 16/3(2^{1/3} + \sqrt{17 \cdot 2^{2/3} + 2 \cdot 2^{2/3}R})$), whose asymptotic behavior at high rates is:

$$D(R) \sim \sqrt{\frac{\log(R)}{R}}. \tag{3.21}$$

See Appendix A.2 for a proof.

Although the bound, in this case, is slightly worse than in the isotropic decomposition case (see Fig. 3.11), the asymptotic behavior turns out to be the same. This scheme exhibits much better performances, when coding the *Moving Horizon* model, than classic wavelet for video.

3.4.5 2D+1D Temporally Adaptive Subband Coding

Fig. 3.11 depicts the coding performance of each one of the wavelet decomposition schemes analyzed in the previous subsections. Curves correspond to the $D(R)$ lower bound for the classic scheme and the upper bounds of the alternative decompositions. Their respective $D(R)$ expressions are those exposed in Theorems 3.1, 3.2 and 3.3. It is clear how the coding performance of the isotropic and the unequal based decompositions outperforms the classic one. This difference becomes even bigger at high rates. In effect, the usual classic decomposition, which presents the typical temporal ringing effects, is not at all appropriate to code moving edges.

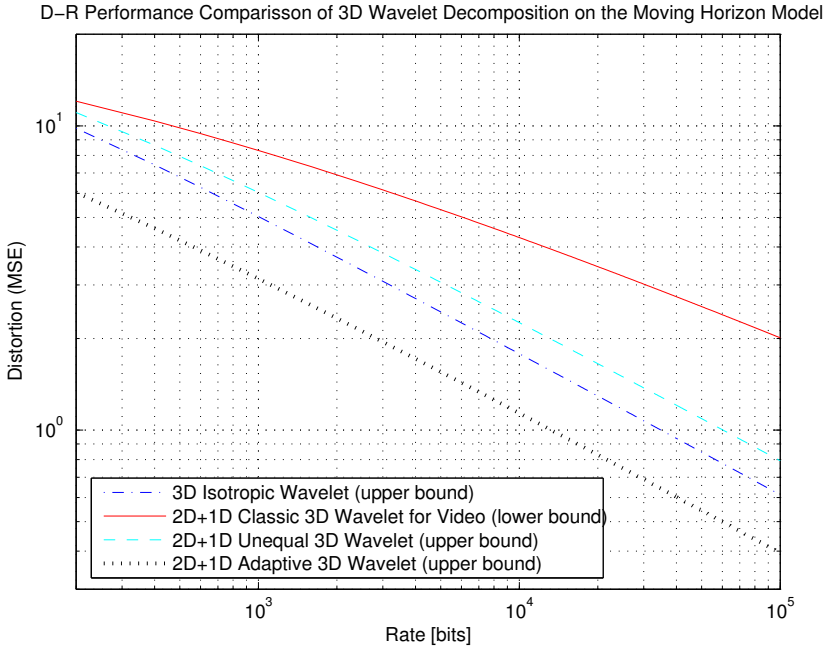


Figure 3.11: Theoretical R-D Performance of different 3D Wavelet decomposition schemes applied on the *Moving Horizon* model. The classic 3D Wavelet curve represents a lower bound on the R-D, whereas the other two are upper bounds on the R-D performances of the isotropic and the unequal decompositions, respectively.

However, video is not only composed of moving edges; static regions are common and are fancily mixed with edges in the same scene. This heterogeneous nature, as one may expect, requires the use of adaptation in the representation. In the framework of 3D wavelet video coding, temporal transforms need, thus, to be adapted depending on the spatial location. In this way, classic wavelet schemes may be exploited for static scene regions while wavelet decomposition depth may be adapted within moving areas.

If we have a closer look at the best decomposition schemes of the previously analyzed, one can see that, when coding moving edges, there is still some counterproductive temporal filtering. Some side signaling should describe the temporal filtering adaptation such that a large variety of decompositions are enabled: from no temporal transformation at all up to full Group of Pictures (GOP) length.

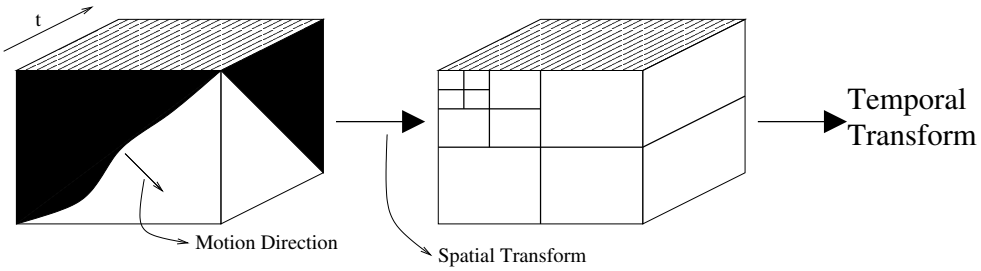


Figure 3.12: Model taken for the theoretical performance estimation.

Fig. 3.12, depicts the fact that, after a first spatial transformation of each frame, an adaptive temporal transformation is applied instead of the fixed length one. The reader will notice that using spatially local adaptivity in the temporal transform involves a non-linear decomposition scheme. This implies that, unlike in the separable linear case, the order between spatial and temporal transformations shall not be exchanged.

A way of spatially adapting the different temporal decompositions is to divide spatial subbands into subband dependent macroblocks. Each spatial macroblock forms together with the temporal dimension a spatio-temporal tube. In each of these temporal *tubes*, a different temporal transform may be applied (see Fig. 3.13).

Temporal partitions can be set by means of a combinatorial approach (as in [148]) or, more computationally efficiently, by using a prune-join* binary tree approach [164]. To efficiently represent the data contained in each partition (whose size may differ from that of a dyadic partition) one can use frame adapted lifting steps [70] (see also the example of Sec. 3.8.3).

The R-D bound:

Let us now derive an upper bound on the R-D behavior of an adaptive transform based coding scheme applied on our toy video signal. For this particular example, rate coding costs of a prune (and not a prune-join) binary tree are considered. This is done for two main reasons:

- The coding efficiency, at high rates, of a simple prune binary tree approach is lower than that of a prune-join binary tree. Hence the R-D bound computed with the cost of a full depth prune binary tree can be considered as an upper bound on the distortion for a given rate.

*The need of a prune-join approach instead of a simple prune binary tree is due to the need to optimally handling 1D piecewise-smooth discontinuity events. This can be for example: sudden motion of a region that was static, or sudden stop of a moving edge. Simple moving edges are more like delta singularities once the 2D transform has been applied. For these simple prune binary trees are enough to describe them.

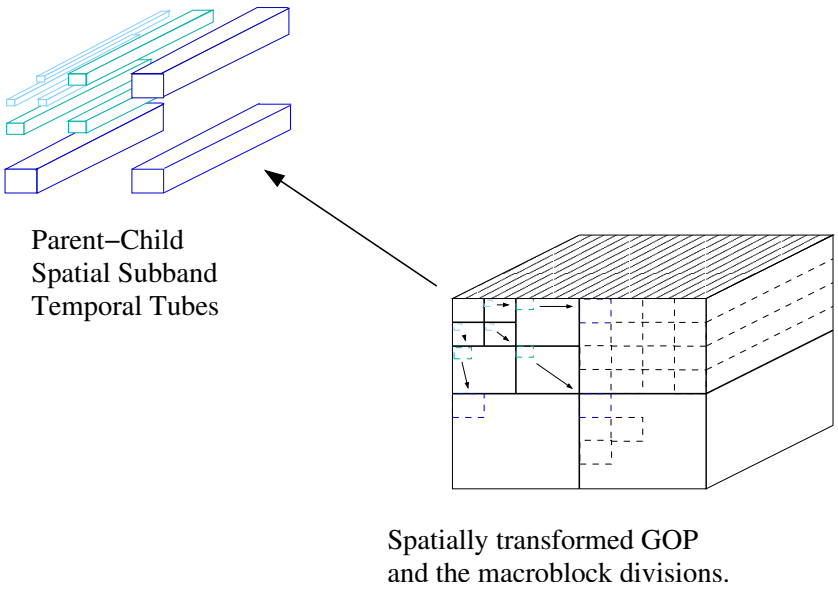


Figure 3.13: A possible spatially local implementation of temporal adaptivity.

- The resulting partitioning coding tree of one of the temporal macroblocks intersecting the *Moving Horizon* edge, would correspond, locally, to a full depth tree partition (see the left figure in Fig. 3.14). This means that, for each frame where the edge is crossing a given macroblock, there will be a leaf of the partitioning tree of that macroblock. Hence, locally, the resulting tree structures from both schemes are equivalent.

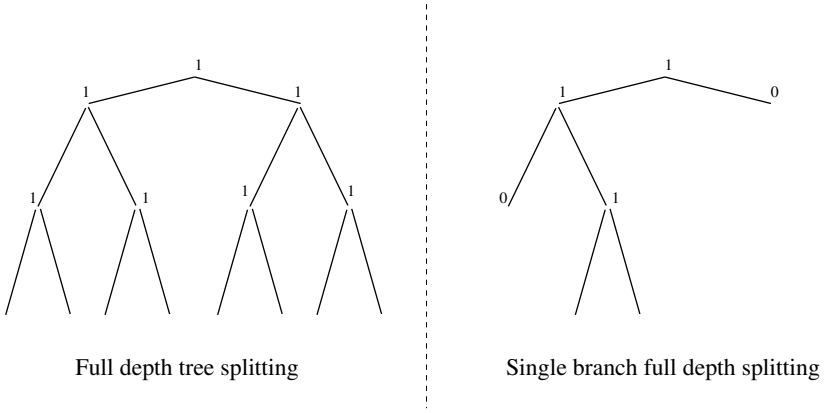


Figure 3.14: Binary tree used in the temporal partition.

In the present analysis, we assume for simplicity purposes, that only a single tree is used per spatial macroblock. This means that all macroblocks at the same spatial location are partitioned using the same tree. Without loss of generality, this is done to simplify the calculation of the side information cost to derive a $D(R)$ bound for the temporally adaptive coding scheme. Indeed, even if the scale of analysis is different depending on the spatial wavelet subband, the optimal partitioning trees for several parent-child related areas are expected to be quite similar in practice.

Considering all that, we can state the following:

Theorem 3.4 *Assume the use of a coding scheme based on a 3D macroblock adaptive wavelet decomposition, scalar quantization, efficient coefficients position coding and efficient partition coding. Then, its R-D performance ($D(R)$) coding the* Moving Horizon *model can be upper bounded as:*

$$D(R) \lesssim \frac{2^{5/6}208\sqrt{W(2\log(2)E(R)^2)}}{\sqrt{\log(2)}E(R)} - \frac{2^{2/3}1024\ W\left(2\log(2)E(R)^2\right)}{\log(2)E(R)^2}, \tag{3.22}$$

(where $E(R) = 16/3 \cdot 2^{1/3} + 16/3 \cdot 2^{1/3}\sqrt{37+4R}$) which has an asymptotic behavior at high rates such that:

$$D(R) \sim \sqrt{\frac{\log(R)}{R}}. \tag{3.23}$$

We observe that, in the adaptive case, the asymptotic R-D behavior is the same than that of the 3D isotropic wavelet and that of the unequal partition case. However, given the full adaptation of the approach, one can see in Fig. 3.11 how the R-D upper bound is even lower than the rest.

Proof: The proof follows the same schema than the previous ones. Modeling of the different involved elements is described shortly in the following.

The number of coefficients per spatial decomposition level given the full depth partition tree is:

$$n_j \sim 3 \cdot 2^j 2^J.$$

For each 2D subband, the number of non-zero coefficients is proportional to 2^j (where j indicates the scale level), for the 2^J involved temporal samples.

The total number of coefficients is, thus,

$$N_J \sim \sum_{j=0}^{J} n_j + \frac{n_0}{3} \sim 6 \cdot 4^J - 2^J 2.$$

Distortion is the addition of the error introduced by quantization and scale truncation. Quantization and spatio-temporal truncation remains the same as in previous schemes. Hence, distortion can be computed as:

$$D \sim 2^{-J}13 - 4^{-J}2. \tag{3.24}$$

The main change appears in the rate where the additional factor containing the cost of coding the partition tree appears. The coding cost is proportional to the number of nodes. That is, for the case of a full depth tree, $R_{tree} \sim 2^J - 1$. Thus,

$$R \sim N_J \cdot R_{c_{j,\mathbf{k}}} + R_{tree} + (N_J - 1) \cdot 8 = 4^J (9J + 48) - (3J + 15)\, 2^J - 9.$$

The reader will notice that the partition tree information, some extra bits in efficient coding of the coefficients oct-tree could be also saved. Nevertheless, computations are done as if a total of eight bits per coefficient were needed as well.

R may be upper bounded as

$$R \lesssim 4^J 9 \left(J + \frac{48}{9}\right) - 2^J 3\sqrt{J + \frac{48}{9}} - 9.$$

Following the method of previous sections, J is easily derived as:

$$J \gtrsim \frac{3\mathrm{W}\left(2\log(2)E(R)^2\right) - 32\log(2)}{6\log(2)},$$

where $E(R) = 16/3 \cdot 2^{1/3} + 16/3 \cdot 2^{1/3}\sqrt{37 + 4R}$. This, combined with (3.24), yields the final result of Theorem 3.4.

■

3.5 2D+1D Temporally Adaptive Decomposition and Coding Scheme

In this section, a short review of the implementation realized to check the hypothesis and conclusions of sections 3.3 and 3.4 is given. As previously seen, better performances are achieved by approaches based on setting an adaptive temporal transformation instead of the classic fixed one typically used in 3D wavelet video coding. Previous to the temporal expansion, the spatial transform is obtained by means of the classic 2D isotropic dyadic wavelet used for image coding [8]. In our case, the usual Daubechies-9/7 biorthogonal filters [121], known to perform very well, have been selected.

3.5.1 R-D Adaptive Temporal Subband Decomposition Using the Lifting Scheme

The lifting scheme (see Chap. 2 for an introduction) has introduced into wavelet transforms a wide range of possibilities thanks to its flexibility. It allows non-linear operations in the prediction and update steps while allowing a perfect reconstruction of the original data. In motion compensated subband video coding (e.g. MCTF video coding), the use of the lifting scheme has allowed to introduce motion information into the prediction/update steps. Moreover, very flexible implementations permit to adaptively change the kind of wavelet in use [70, 73] as well as to introduce breakpoints in the wavelet decomposition to locally adapt the number of wavelet subbands [46].

For the sake of simplicity, the implementation of our experiment is based on the Haar transform (in what concerns the temporal dimension). The transform is based on a rate-distortion optimization, in which the Haar lifting step or a void step (see sec. 3.5.1) are selected.

1D Haar Wavelet Transform Using the Lifting Scheme

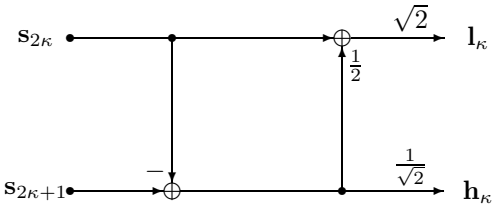


Figure 3.15: Haar transform lifting steps.

Let us briefly recall the basic structure of a lifted Haar step, which is one of the most used ladder structures to illustrate lifting schemes [39, 171]. Fig. 3.15 depicts the ladder scheme that carries out the Haar transform applied to the signal along time. Even samples are used to predict odd ones thanks to a first order predictor. Then residual error is re-injected to the original even sample in order to orthogonalize both outputs (hi-band (h_k) and low-band (l_k)). Output scaling factors take care of the energy normalization of coefficients.

Adapting Wavelet Decomposition Depth within the Lifting Scheme

In order to tackle the problem of temporal adaptivity, the so-called intra refresh may be introduced into the lifted wavelet scheme.

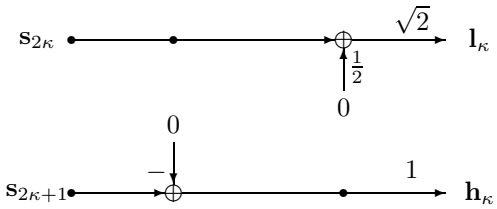


Figure 3.16: Broken lifting step, neither prediction nor update is performed. The use of the ladder scheme would reduce R-D performance when applied to a piecewise-smooth signal. Both, prediction and update, may be inhibited on a picture macroblock level. Notice the change on the scaling factors to control the noise at the quantization stage.

Intra refresh corresponds to the adaptive insertion of void lifting steps. These do not perform further temporal filtering on the signal. They implement the necessary breakpoints in the wavelet decomposition for an efficient R-D signal approximation. This special "lifting" mode is depicted in Fig. 3.16. With proper selection of this lifting mode, an approximation to the desired temporal wavelet decomposition with a local adaptation of the number of subbands is obtained.

The reader will notice that in the absence of prediction and update steps, scaling factors of the output signals have been modified. This is in order to adapt the h_k output signal to the fixed step size quantizer used for all subbands. Regarding the low band l_k, the $\sqrt{2}$ scaling adapts the dynamic range of the signal to fit that of the next scale level for further decomposition.

3.5.2 The Coding Scheme

The coding scheme used in our experiments is based on an intra-adaptive approach, similar to the one used in the MCTF extension of H.263++ [46], but applied to the 2D wavelet transformed frames. This works on a 16x16 macroblock (MB) basis and motion compensation is switch off. To get the best coding performances, macroblock sizes should be adapted according to the 2D spatial subband scale (as discussed in Sec. 3.4.5). For simplicity, however, a fixed macroblock size implementation has been selected, which is largely enough for the proof of concept. The encoder chooses for each macroblock the best type of step in a rate-distortion sense, i.e. minimizing the R-D Lagrangian cost of the high band that issues from the ladder scheme. Haar-type and void-type are the two candidates to encode each macroblock. For simplicity, there are no further MBs subdivisions for a finer edge adaptation, hence non-linear adaptive partitions are associated to MBs of size 16x16. Temporal transform breakpoints are estimated at each decomposition level depending on the "low" frequency subband obtained at the lower level. All subband macroblocks generated by the temporal wavelet transform are encoded in a wavelet intra-frame fashion. For this purpose, a block by block raster scanning is used. All coded subbands are quantized using a dead-zone uniform quantizer. The same quantization step-size is used for all them. Finally, the whole generated information is encoded with Huffman codes. GOPs of size 32 are used in our experiments (up to five wavelet decomposition levels), shorter length wavelet transforms are provided by the temporal adaptation introduced by Intra MBs (void-type lifting steps).

3.6 Results: Performance of Local Adaptation of Temporal Transform's Length

The same sequence model used to derive our theoretical results is used to generate our validating results. A natural video sequence with similar motion characteristics as the ones theoretically studied is also taken into account. This corresponds to the first GOP of *table tennis*, where some piecewise-smooth objects are moving on a static scene.

3.6.1 A Synthetic Scene: The *Moving Horizon* Model.

The *Moving Horizon* sequence has been generated using the function of Fig. 3.8. Two versions of the sequence described in Fig. 3.9 have been obtained. One intends to test which is the smallest gain one may obtain using temporal adaptivity to code a translational edge. In effect, for a static scene, classic video decompositions and temporally adaptive decompositions should behave almost the same (a slight difference could be found due to side information of the adaptive case). Hence, a very small motion is tested: the horizon model moves with a $(1, 1)$ displacement vector.

In order to underline the influence of motion speed into R-D performance of temporal adaptivity, a faster moving version of the *horizon* is also used. In effect, a $(5, 5)$ displacement vector is considered too. Fig. 3.17 depicts the overall R-D performance gain introduced by temporal adaptivity. The left

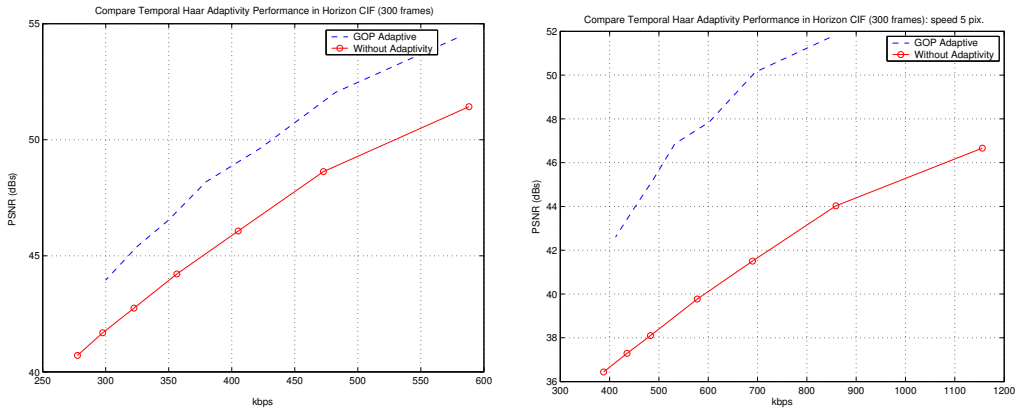


Figure 3.17: R-D performance comparison between the fixed subband decomposition scheme and the temporally adaptive. Left: (1,1) displacement vector speed. Right: (5,5) displacement vector speed.

graphic of Fig. 3.17 compares the different R-D performance in the slow motion case, whereas the right graphic illustrates the fast motion case. Both graphs evidence what was already expected: the higher the motion is, the more beneficial locally temporal adaptive wavelet decompositions become.

To have a better idea of adaptivity contribution frame by frame, a comparison of the temporal distortion evolution for a determined coding rate is shown in Fig. 3.18. This shows that the improvement of performance due to adaptivity is constant in average and affects the whole sequence. The overall gain associated to the use of adaptive transforms can also be evidenced if one visually examines the obtained residual error after signal compression (see Fig. 3.19). In fact, fixed dyadic temporal wavelet decomposition generates a much higher number of coefficients. In *classic* 3D wavelet decomposition, more coefficients are used to describe each of the salient moving signal structures (edges). A higher amount of quantization noise is, thus, introduced during the non-linear

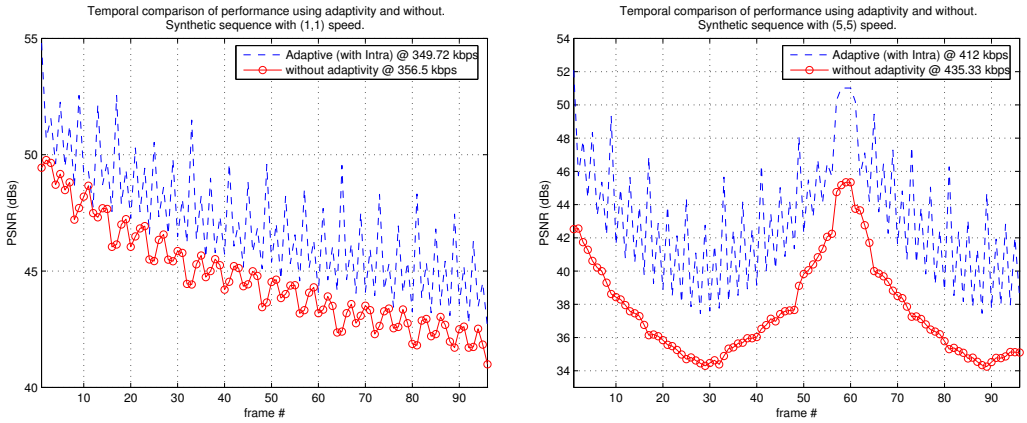


Figure 3.18: Temporal performance comparison between the fixed subband decomposition scheme and the temporally adaptive. Left: (1,1) displacement vector speed. Right: (5,5) displacement vector speed.

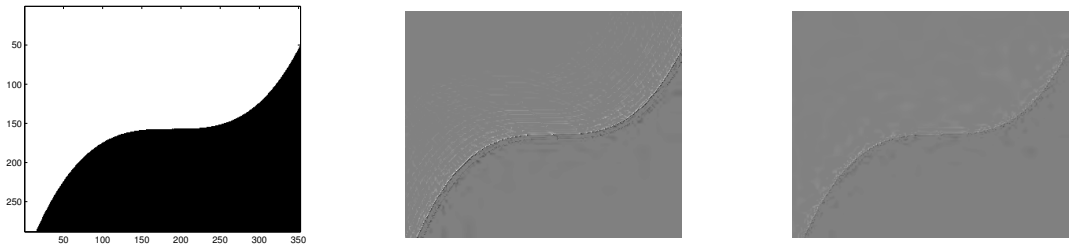


Figure 3.19: Visual comparison between the residual error after compression of the 30th frame of the synthetic sequence *Horizon* (see left picture for the original frame) with speed vector of (5,5) pixels. In the middle, we see the residual signal that corresponds to the fixed decomposition structure and compression rate of 435.3 kbps. At the right, the residual generated by the adaptive scheme at 412 kbps can be seen.

approximation stage of compression. In non-linear wavelet approximations of piecewise-smooth signals, quantization noise produces the well known ringing effect. The reader will easily see in Fig. 3.19 that the amount of residual signal in the temporally adaptive decomposition is significantly lowered.

3.6.2 A Natural Scene: Table Tennis

Let us now check the behavior of the temporally adaptive approach with a scene having the ingredients of the synthetic videos analyzed above (moving edges) but with a natural origin. Fig. 3.20 illustrates a fraction of the natural sequence used for the tests. The selected frames correspond to an equally spaced sample of the first 32 frames GOP of *table tennis*. The example is composed of a fixed background where smooth objects move at different speeds. The ball moves fast while the arm has a moderate speed. Akin to the synthetic example, the use of temporal adaptivity introduces an

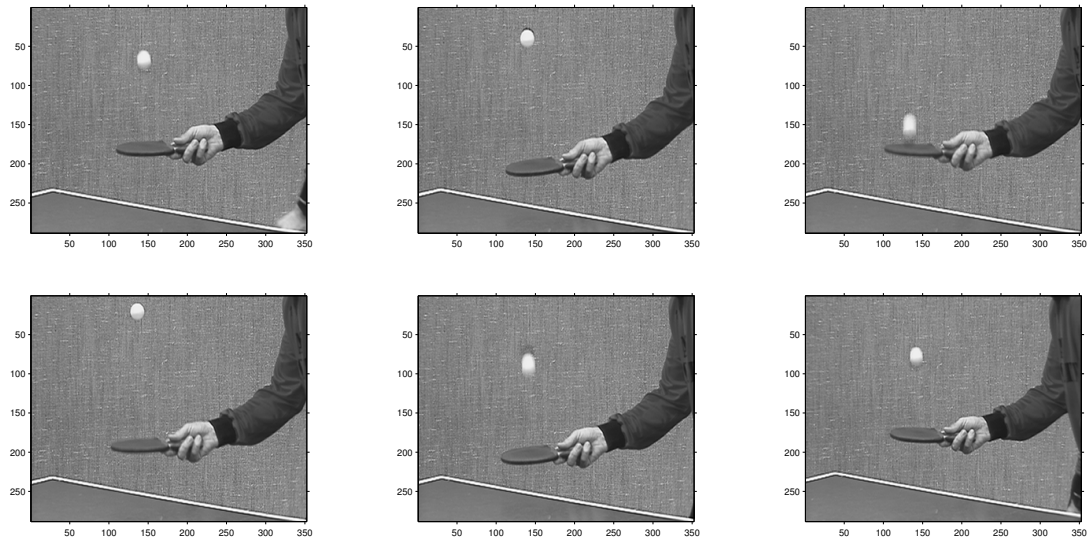


Figure 3.20: Sample of the table tennis sequence. Objects are delimited by smooth curves moving through time. The temporal sequence is from left to right and from up down.

overall R-D performance gain. However, the higher complexity of the signal (i.e. motion and edge geometry are much more diverse) and the lower contrast between regions make the gains, in this case, less significant (see Fig. 3.21). In the frame by frame evolution of PSNR (Fig. 3.21), we see a significant distortion reductions for certain frames. Notice also that, even if many frames exhibit a very similar distortion, the compression rate is 23 kbps smaller in the adaptive case.

Distortion decrease has an impact in the reduction of spatio-temporal ringing as depicted by Fig. 3.22. In it, both approaches (fixed dyadic and adaptive) are compared for the 25th frame. For a clearer visual analysis, the reader may see in Fig. 3.23 the error introduced by both approaches during compression. The ringing effect and the lack of edges preservation are much more evident when the fixed length dyadic wavelet decomposition has been used. Spatio-temporal ringing is reduced in the adaptive case.

The simulations and the generation of results for this section, have show that the average signal complexity of natural scenes makes somehow more difficult to achieve improvements in the R-D performance by means of using our particular adaptive transform implementation. In fact, there

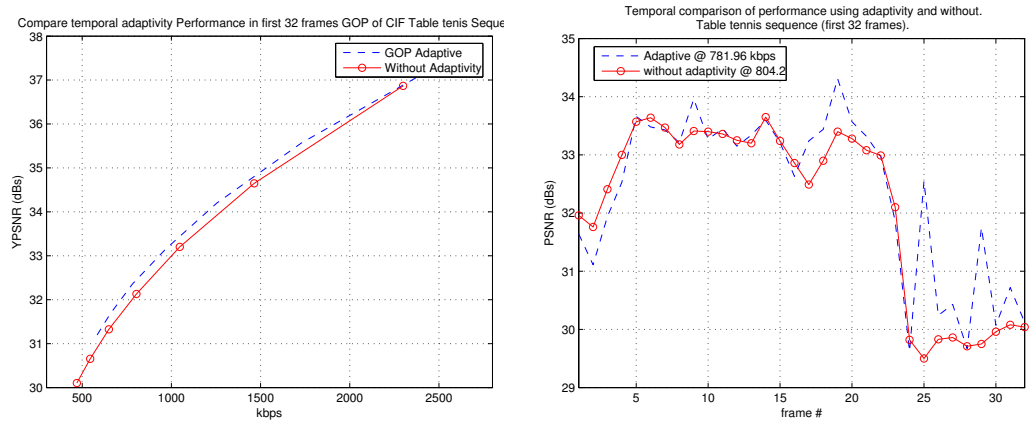


Figure 3.21: left: R-D performance comparison between the fixed subband decomposition scheme and the temporally adaptive for the first 32 frames GOP of the sequence Table tennis (CIF format). right: Temporal performance comparison between the fixed subband decomposition scheme and the temporally adaptive.

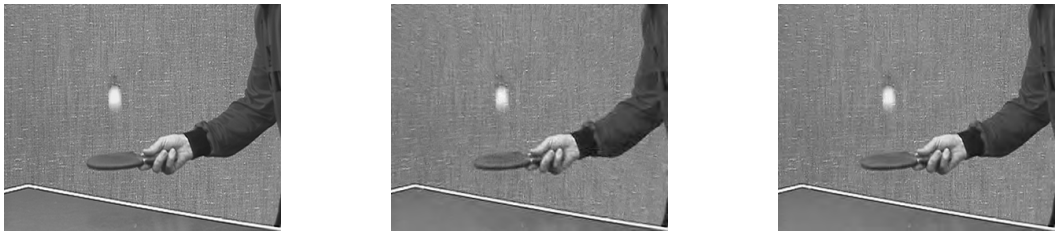


Figure 3.22: Visual comparison between both versions of the 25th frame of the first GOP of the table tennis sequence (left: original frame, middle: non-adaptive, right: adaptive). This example illustrates how in the most evident cases, the contours of moving objects are better preserved when adaptivity is in use. Ringing is lighter around areas nearby contours.

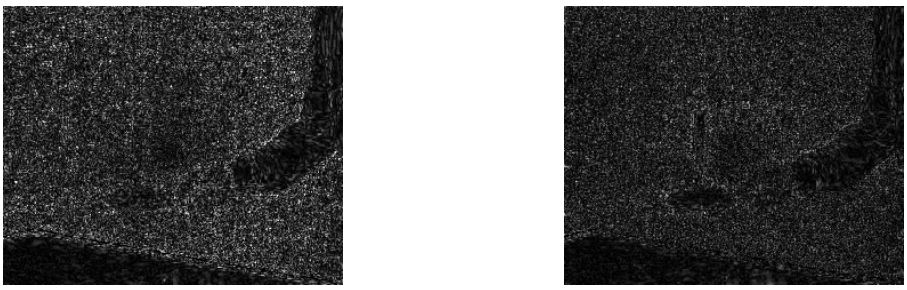


Figure 3.23: Visual comparison between the absolut value of the residual error after compression of the 25th frame of the Table tennis sequence. On the left, we see the residual signal that corresponds to the fixed decomposition structure and compression rate of 804.2 kbps. On the right, the residual generated by the adaptive scheme at 781.98 kbps can be seen.

is an additional factor that may help understanding this behavior and how to avoid it. The algorithm used to implement the temporal adaptive transform is based on an adaptive version of the lifting scheme. As discussed previously, this selects, based on a R-D criteria, the best of the steps of Sec. 3.5 for every spatial location (each macroblock in our basic implementation). The generation of the lifted wavelets and the emplacement of lifting splitting steps are done in a level by level fashion. Like in the classic lifting scheme, once an adaptive lifted wavelet level has been generated, the next wavelet decomposition level is obtained by applying again the lifting procedure to the result issued from the previous level. In our case, however, the preliminary splittings from a previous level may pose some problem to the subsequent decomposition levels. The fact that each lifting decomposition level is optimized in a greedy fashion without taking into account the rest of levels implies that the final solution, in most of the cases, is suboptimal. An additional point to consider, regarding sub-optimality, is that R-D optimization is based on the heuristic that only R-D performance of the high bands is taken into account. As seen with the practical examples, improvements due to adaptivity may be sufficiently good enough such that, despite the suboptimal algorithm, a global increase of performance is registered. However, in the case of many natural scenes, we have found this sub-optimality to be an important handicap. As suggested in Sec. 3.4.5, a solution to the suboptimal implementation issue is to formulate the problem using a more complex and computationally expensive dynamic programing approach [148]. To find the best solution, a combinatorial problem that takes into account all possibilities and all subband levels should be formulated. Otherwise, a prune-join binary tree based algorithm may find a good solution with a reasonable computational cost [163]. Arbitrary length transforms may be obtained by introducing frame-adaptive lifting schemes [70].

3.7 Adapting Wavelet Expansions in MCTF Video Coding

3.7.1 Is MC the Solution for Video Representations?

The use of temporal adaptivity in 3D wavelet video coding has been discussed until here. We have underlined the importance of exploiting geometry in the representation of signal manifolds like those generated by moving edges in a video sequence. In this sense, motion compensation based techniques have proved to be successful for video representation. As discussed in Chap. 2, MC is intended to capture video geometric changes through time. Current researches aim to combine the advantages of linear temporal transforms and efficient motion compensation. A promising scheme for exploiting successfully temporal redundancy is based on motion-compensated temporal wavelets implemented in the lifting scheme.

If we assume that, as far as there is motion to track, the temporal motion-compensated wavelet transform is perfectly aligned with this (even if this may be seen as an ideal situation), then fixed length temporal subband representations (as used for video coding in [73]) are suboptimal. This is due to the fact that abrupt transitions in the signal (e.g. sharp termination of a motion trajectory) may generate many non-zero wavelet coefficients. Hence, here arises again the need for temporally adaptive partitions for the wavelet based representation. The remaining of this chapter discusses the use of adaptive temporal representations within the well known MCTF framework.

3.7.2 Motion Compensated Video and Piecewise-x Signals

Sec. 2.5.2 discusses how motion oriented filtering can drastically reduce the number of significant wavelet coefficients generated in the transform. The effect of including motion compensation, from the transformation point of view, is to smooth out the signal that is going to be transformed with

the temporal wavelet. As long as the motion of the scene can be accurately estimated, the temporal signal processed by the wavelet transform will be smooth or even constant if no local temporal illumination changes occur in the scene. When motion cannot be correctly estimated, or simply when there is an occlusion or an appearing object, the signal seen by the wavelet transform presents a step in amplitude. This step issues from the mismatch between the best signal sample candidate for prediction found by MC and the signal sample being predicted. Again, we can associate the temporal behavior of each of the trajectories in MC video with the *piecewise-smooth* model. In Sec. 3.3.2 we reviewed the rate distortion consequences of that with regard to wavelet representations as well as the benefits of oracle based representation methods.

To address the needed adaptivity, we consider, as in Sec. 3.5 the use of a lifting wavelet based approach. More particularly, the frame-adaptive lifting scheme proposed in [70] (see Sec. 3.8) is modified by introducing an additional lifting mode (Intra mode) in order to allow a more adaptive, oracle like, method. The length of wavelet transforms (i.e. the number of decomposition subbands) is adapted to cover smooth areas while avoiding wavelet kernels to cross edges. As in the previous case where MC was not used, this transform adaptivity is, in some sense, a *Best Basis* (see for example the works by *Ramchandran* [153, 154]). However, unlike in the work by *Ramchandran*, we are luckily not constrained here by the rigid structure of simply pruned binary trees. This non-dyadic flexibility is what allows to obtain significantly better results than simple wavelet based representations.

3.8 Intra-Adaptive Motion-Compensated Lifted Wavelet Transforms

Our analysis and results of Sec. 3.9 are based on the multi-hypothesis, frame-adaptive motion compensated lifting scheme proposed in [70, 75] enhanced with Intra-adaptive steps [47]. The Multi-hypothesis frame-adaptive scheme already introduces some temporal adaptivity allowing a free selection of reference frames for MC within a GOP. Moreover, it allows an adaptive selection of the most suitable lifting step (Haar or 5/3) for a minimum distortion at a given rate. Nevertheless, it still forces a fixed number of multi-scale subbands in the MC temporal wavelet decomposition of the signal and limits temporal adaptation dealing with temporal discontinuities and motion misalignments as discussed in Sec. 3.7. In the following we review the scheme presented in [70] and [75], we discuss its implicit temporally adaptive properties and propose the inclusion of Intra macroblocks as an additional mode within the lifting scheme in order to allow further flexibility.

3.8.1 Frame-adaptive Motion-Compensated Lifting Scheme

Frame-adaptive MC Lifting schemes are a very flexible approach that allows a representation to be adapted (up to a certain degree) to the signal. It helps to overcome occlusion effects, changes of scene or to palliate the effects of deficient motion compensation at fine scales. Figs. 3.24 and 3.25 show the way frame adaptivity is implemented in the lifting scheme for the Haar and 5/3 wavelet cases. As it can be seen, in this two particular examples, the fixed and classical structure of the lifting scheme is broken such that for every instance of this, every even frame in the GOP can be used to select the best prediction signals. The update step is performed accordingly and following the complementary scheme to the one determined in the prediction step.

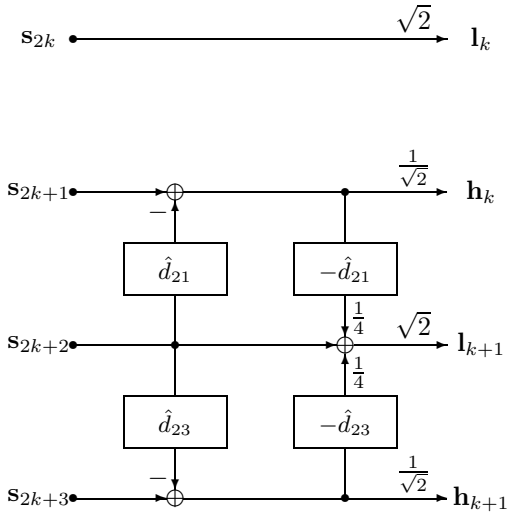


Figure 3.24: Example of the first decomposition level of the Haar transform with frame-adaptive motion-compensated lifting steps. The frame $\mathbf{s}_{2k+2}$ is used to predict frame $\mathbf{s}_{2k+1}$

.

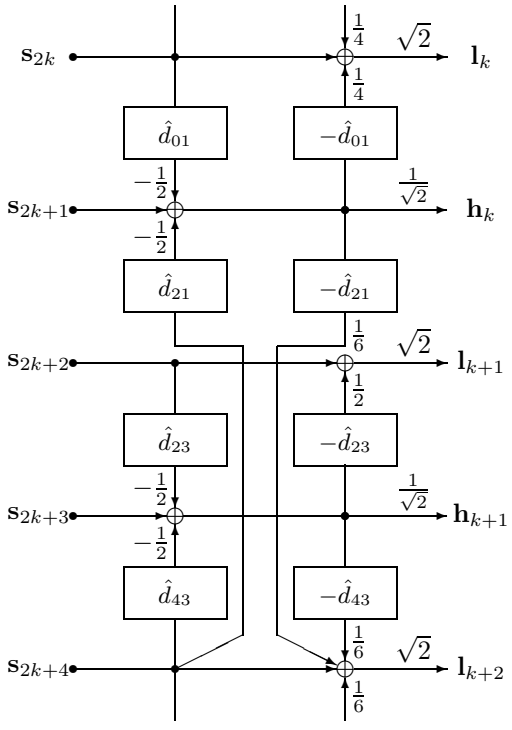


Figure 3.25: Example of the first decomposition level of the 5/3 transform with frame-adaptive motion-compensated lifting steps. The frames $\mathbf{s}_{2k}$ and $\mathbf{s}_{2k+4}$ are used to predict frame $\mathbf{s}_{2k+1}$

3.8.2 Intra-Adaptive Scheme

The frame-adaptive scheme does not change the default number of wavelet decomposition subbands nor considers alternative methods for MC at a particular location. Several works have introduced in MC wavelet based video coding ([128, 177]) the so called intra refresh (widely used in classic predictive video coding approaches [7]). See also a recent work published as a proposal for the new scalable video coding standard [157].

In the framework of MC lifted wavelets, intra refresh corresponds to the adaptive insertion of void lifting steps. These do not perform further temporal filtering on the signal. They implement the necessary breakpoints in the wavelet decomposition for an efficient R-D signal approximation (as discussed in Sec. 3.7 and Sec. 3.3.2). This special "lifting" mode is depicted in Fig. 3.16. With a proper selection of this lifting mode, an approximation to the desired temporal wavelet decomposition with a local adaptation of the number of subbands is obtained.

Due to the absence of prediction and update steps, the scaling factors of the output signals have been modified. We adapt the output signals h_k to the fixed quantizer step size used for all subbands. Regarding the low band l_k, the scaling by $\sqrt{2}$ adapts the dynamic range of the signal to fit that of the next scale level to be further decomposed.

Both frame- and intra-adaptivity are suitable to handle discontinuities of motion-trajectories in order to achieve an optimal R-D behavior (see Sec. 3.3.2). If suitable reference frames are available, frame-adaptivity will be the best choice in order to reduce the energy in the detail subbands. But if suitable reference pictures are out of reach, the intra mode is required. Moreover, the final level of the dyadic decomposition offers only one reference picture. In that case, the intra mode is the sole alternative.

3.8.3 Toy Example of an Intra-Adaptive Decomposition

Let us consider a simple example in here. Fig. 3.26 depicts the adapted transform that would better suit a piecewise-constant signal for an optimal R-D at middle and high rates. The adapted configuration of the ladder steps depends on the R-D trade-off, where coefficients coding and side information are taken into account. If the required rate is sufficiently low, the best configuration may simply turn out to be a simple linear wavelet transform. In effect, for very low bit-rates, the average approximation of the whole signal becomes more efficient in terms of R-D than using oracle information.

3.9 Results of Intra-Adaptivity in MCTF

In this section, we analyse the effect of introducing local spatio-temporal adaptivity into the lifting scheme used for motion compensated temporal filtering. This adaptivity is introduced in practice by inserting Intra macroblocks in the lifting scheme (as described in Sec. 3.8). We evaluate the benefits of using this temporal adaptivity and compare improvements in terms of R-D supplied by the use of Intra MBs in the lifting scheme. Tests are performed using four different test sequences in order to supply different signal characteristics to the coder. These sequences are in QCIF format (176x144 at 30 Hz) and are identified as *cnn*, *football*, *foreman* and *table tennis*. Furthermore, results are presented as well to illustrate the statistics on the usage and selection of Intra Macro-Blocks by the coding algorithm.

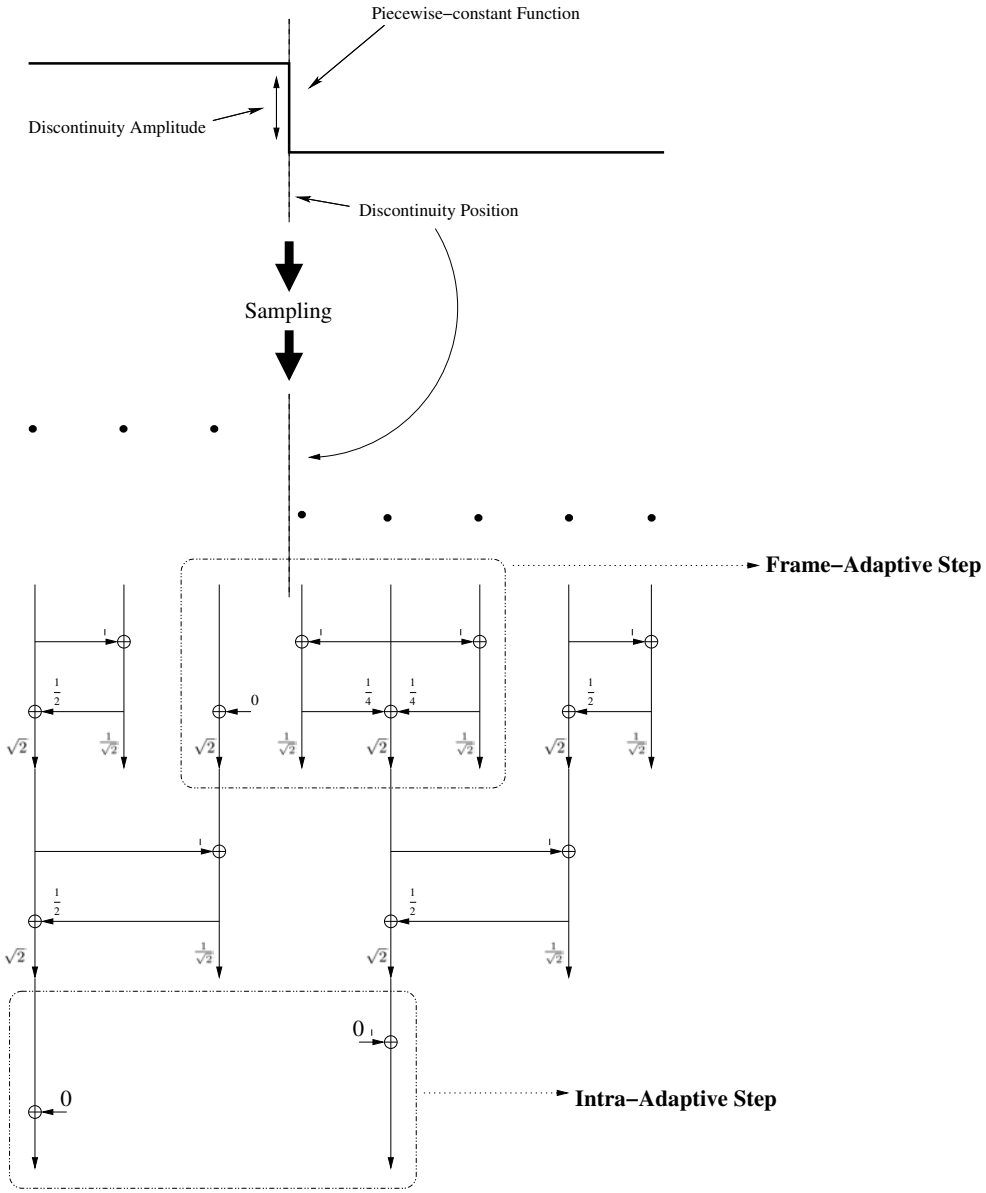


Figure 3.26: Decomposition example for a piecewise-constant function.

3.9.1 The Coding Scheme

The coding scheme used to test the intra-adaptive approach is, again, a MCTF extension of H.263++ [70, 72, 75]. Unlike in Sec. 3.5.1, no previous wavelet transform has been applied on frames, motion is taken into account and frame-adaptive lifting steps are used. The encoder chooses for each macroblock the best type of lifting scheme in a rate-distortion sense. This selection is carried out macroblock by macroblock minimizing the Lagrangian costs of the high band of the lifting scheme. Haar-, 5/3-, and void-type structures are used as candidates to encode each macroblock. Due to the flexibility of the frame adaptive scheme, $M = 1$ or $M = 2$ reference frames may be used by the algorithm to code each macroblock. Since the goal here is the experimental verification of the referenced theoretical concept, there is no further subdivision of macroblocks for motion compensation purposes. Motion vectors are obtained by block-based rate-constrained motion estimation jointly optimized with the lifting mode selection. The motion information is estimated at each decomposition level depending on the results of the lower level by using half-pel accurate motion compensation. All subband macroblocks generated by the temporal wavelet transform are encoded with the H.263 8x8 DCT codec. All intra-frame coded subbands are quantized using a uniform dead-zone quantizer with the same quantization step size. Huffman codes are used for entropy coding and motion vectors are predicted from spatial neighbors. GOPs of size 32 are used in our experiments (up to five decomposition levels). Shorter wavelet transforms are provided by the intra macroblocks.

The reader must notice that in the present MC lifting approach, lifting modes are decided level by level without taking into account the interaction among them. Akin to the previously studied case of adaptive 3D Wavelet video representations, this greedy aspect of the algorithm induces suboptimality in the result. As suggested in Sec. 3.4.5, a solution to the suboptimal implementation issue is to formulate the problem using a more complex and computationally expensive dynamic programing approach [148]. To find the best solution, a combinatorial problem that takes into account all possibilities and all subband levels should be formulated.

3.9.2 Global R-D Performance of Local Temporal Transform Length Adaptation

The usage of shorter instances of the lifted wavelet scheme than the maximum allowed GOP length of 32 commonly contributes locally to areas where object trajectories are shorter than 32 frames. Hence, the benefit is going to be of local nature when the particular characteristics of the sequence require it. Fig. 3.27 shows how using this additional coding mode in the MC lifting scheme, introduces a moderate overall gain to the whole R-D performance of a coded sequence. Average improvements range from 0.2 to 0.5 dBs in middle and low motion sequences with some change of scenes and local fast motion. However, in highly moving sequences, like football, the improvement of introducing the Intra adaptation through the use of Intra MBs is of higher relevance as new information is efficiently coded.

3.9.3 R-D Performance of Local Temporal Transform Length Adaptation

In Fig. 3.28, the PSNR of the adaptive wavelet decomposition length is depicted over time. For *cnn*, *table tennis* or even *foreman*, a very strong panning occurs at a particular moment of the sequence. Due to the strong motion appearing in the sequence *football* a significant overall improvement can be observed for each frame in the upper right chart of Fig. 3.28.

When using exclusively one reference frame for the prediction/update steps, the use of the Intra-adaptive scheme contributes the achieve a slightly better global R-D improvement with respect to

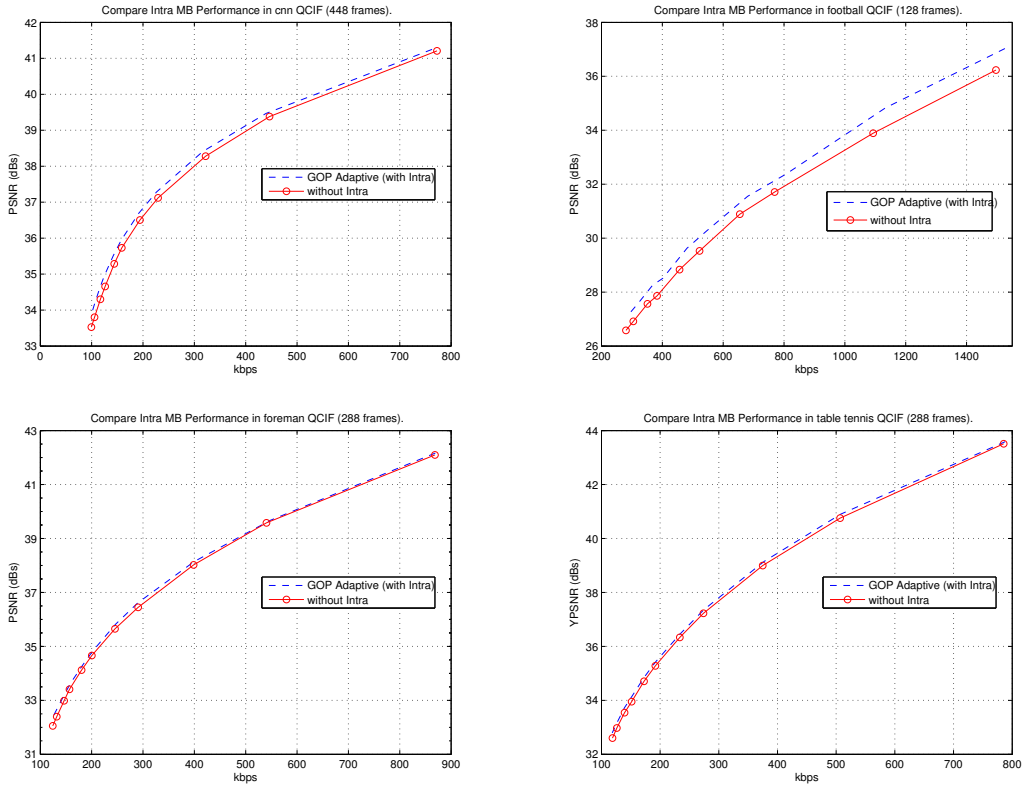


Figure 3.27: Global sequence R-D comparison of the improvement due to GOP Adaptivity.

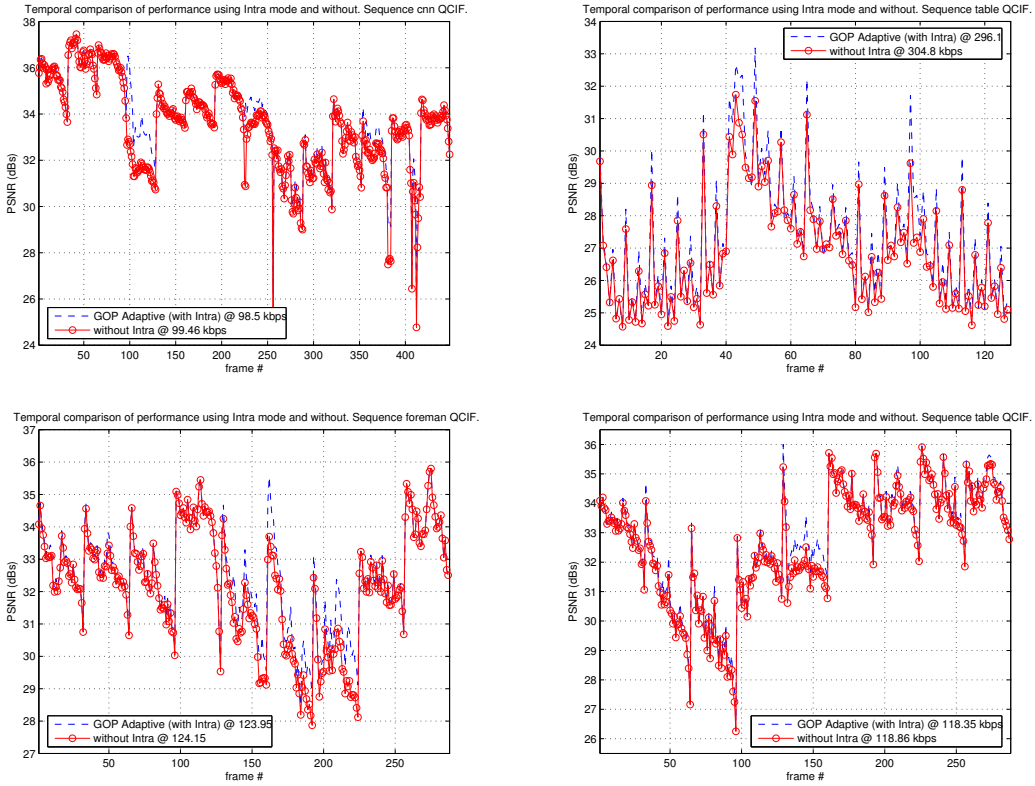


Figure 3.28: Temporal evolution of the PSNR and comparison of the improvement due to GOP Adaptivity.

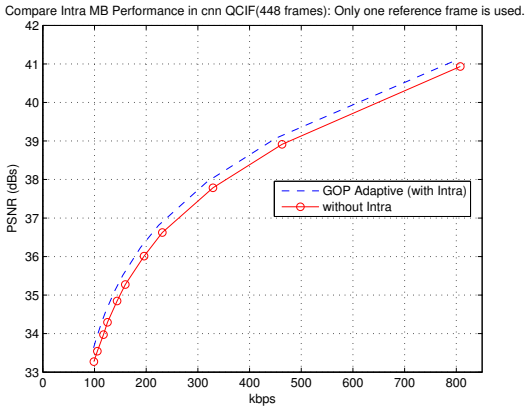


Figure 3.29: Overall sequence R-D comparison of the improvement due to GOP Adaptivity using just one reference frame.

the non Intra-adaptive

In the case where only one reference frame is used for the prediction/update steps, Intra-adaptive R-D improvement is slightly more significant than when two reference frames are allowed. This is illustrated in Fig. 3.29 by the coding R-D performance of the sequence *cnn*.

3.9.4 R-D Performance and Length Adaptation on a Particular GOP

Let us take a GOP where relevant changes appear in the sequence signal and the MC lifting scheme can not efficiently represent them. R-D gain can be as high as 1.0 dB: see the upper left chart in Fig. 3.30 for sequence *cnn*, the upper right chart for *football* and the lower right for *table tennis*.

When only one reference frame is used, Intra-adaptive R-D improvement is slightly more significant, as depicted by Fig. 3.31.

3.9.5 Intra Macro-Blocks and Length Adaptation

Figs. 3.32 and 3.33 show the quantitative usage of Intra Macroblocks to split lifting steps within different contexts. Fig. 3.32 illustrates the proportions of different kinds of prediction modes in the lifting steps. Each bar represents the total of Macroblocks used in each GOP of 32 frames (the fixed number of Intra coded MBs always present in a GOP and commonly used to code the lowest frequency band is not taken into account). In red we observe the percentage of 5/3 wavelet multi-reference prediction modes. In blue appear the usage of Haar wavelet single reference prediction mode. Finally in light green appear the proportion of Intra coded MBs used to locally break particular lifting steps that are not interesting from a R-D point of view. As expected, the use of Intra MBs appears coherent with the temporal scene changes or very fast moving sequence periods.

Fig. 3.33 shows the average frequency of intra MBs at each decomposition level. The index below the column indicates the decomposition level in concordance to the scale of its associated wavelet in the case where non-adaptive lifting steps are used. For dyadic wavelets, the basis function scale evolves according to the level j as 2^{-j} for $j \in \{1, 2, 3, 4, 5\}$. Lifting steps are split when middle or short length wavelet transforms are efficient. Intra MBs are more frequently allocated at low decomposition levels (1, 2 or 3). At high decomposition levels, there is a higher probability that the frame-adaptive scheme finds good reference frames.

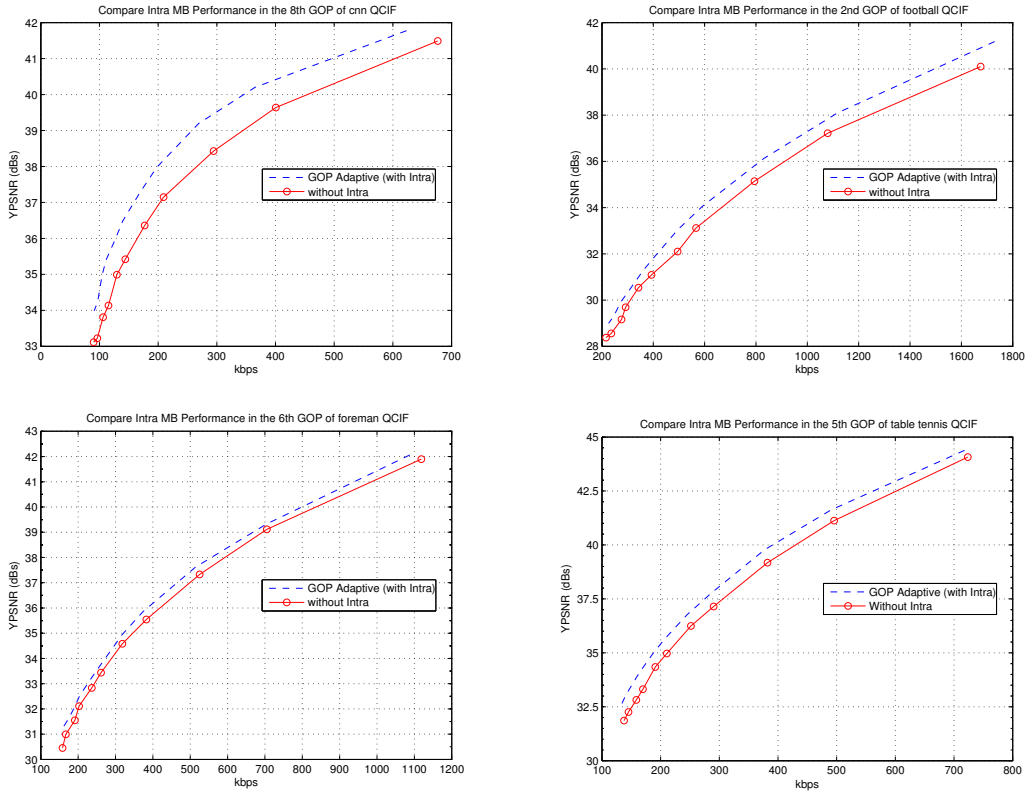


Figure 3.30: Particular GOP R-D comparison of the improvement due to GOP Adaptivity.

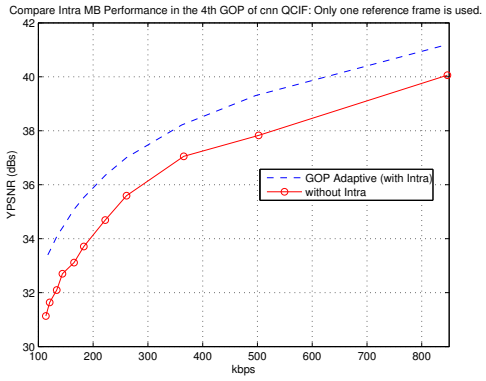


Figure 3.31: 4th GOP R-D comparison of the improvement due to GOP Adaptivity using only one reference frame.

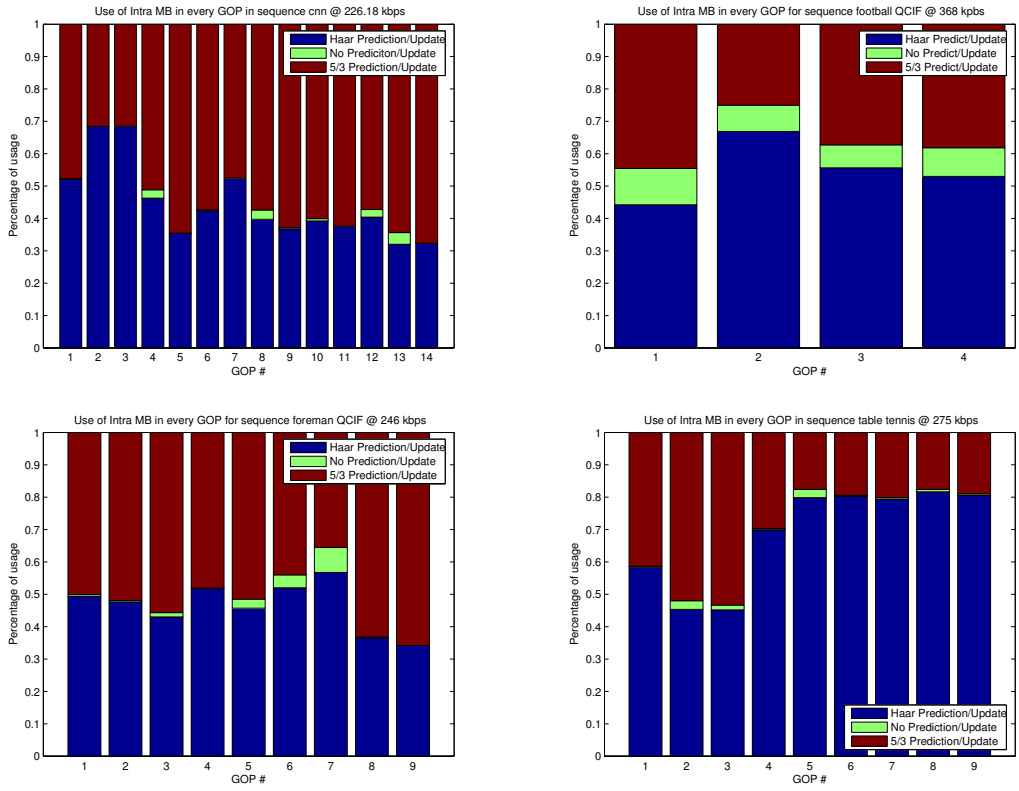


Figure 3.32: Usage of Intra MBs through the different GOPs, GOP length Adaptivity is mainly present in highly moving scenes where MC performs bad and in scene shots.

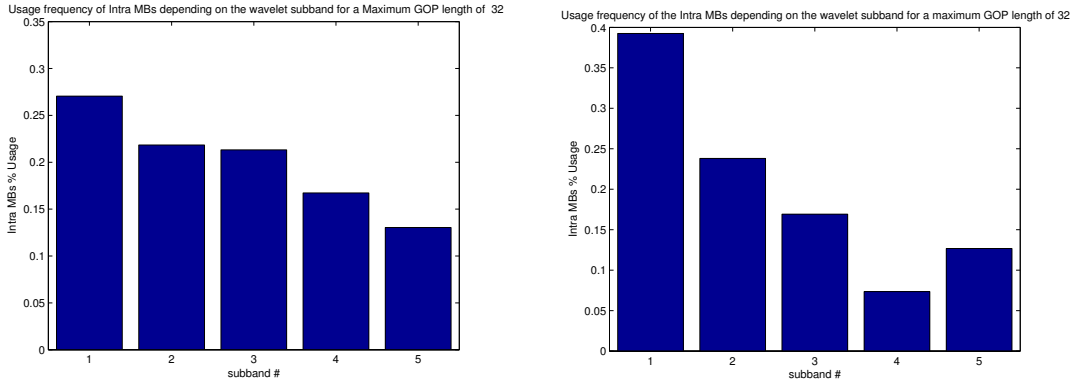


Figure 3.33: Average frequency of usage of Intra MBs depending on the temporal wavelet subband for a maximum GOP length of 32. The lower the number of the wavelet subband, the bigger the scale of details that it represents. The graphic shows the descending trend in the splitting frequency of the temporal lifting scheme. (Left) All four test sequences are considered to generate the statistic. (Right) This statistic does not contain the sequence *football*

However, since *football* is a highly moving sequence, that requires a higher re-injection of the information than other sequences and, thus, it presents a very different kind of motion with respect to other test sequences, we show in Fig. 3.33 the same statistic presented in Fig. 3.33 but without taking into account *football*. This shows that for slower sequences, Intra MBs are concentrated in lower subbands. Hence, longer wavelet transforms are used.

3.9.6 Visual Comparison and Length Adaptation

Finally, a visual comparison between the normally coded sequences with the scheme proposed in [70, 75] and the wavelet decomposition depth adaptive is presented in Fig. 3.34. The most relevant, in addition to the numerical improvements in the previous R-D results, is the visual noise reduction evidenced in the pictures. Noise reduction is present in all four sequences. This noise reduction is due to the fact that the quantization error introduced at big scale wavelet subbands is not spread over all the GOP when the wavelet decomposition is allowed to be split in less deeper decompositions. Indeed, Intra-adaptivity allows for a higher energy compaction in the signal approximation. Hence, signal structures are represented by fewer wavelet coefficients which reduces the introduction of quantization noise in relevant signal components. Moreover, the amount of rate spared thanks to the new lifting mode can be invested in other critical macroblocks. In some cases the noise reduction appears as a relevant increase of the reconstruction detail of objects appearing in the sequence. Notice, for instance, the sharper and clearer appearance of *Yeltsin* in the *cnn* sequence, and the lower blocking effect in the player t-shirt of *football* sequence.

3.10 Conclusions

The use of temporal adaptivity in subband video coding decompositions and motion-compensated temporal transform coding of video signals has been discussed in this chapter.

Although the main major signal division for its processing is the GOP of K pictures, we do not impose a fixed number of temporal decomposition subbands. Our approach is such that GOPs of K pictures are adaptively broken in smaller ones in a spatially local fashion. Like this, the number of wavelet decomposition subbands in the temporal transform is adapted in space and time. Local signal breakpoints are coded independently while wavelet kernels are reserved for those areas of the signal where prediction can be efficiently made using the classic lifting and the MC lifting schemes.

In this chapter, video has been modeled as a 3D signal with several particular characteristics. Based on the fact that natural video signals are sequences of natural images, video can be seen as a 3D piecewise-smooth signal made of piecewise-smooth regions that have certain motions through time. Based on this model, we refer to related theoretical work and discuss local spatio-temporal adaptations of MCTF schemes for video coding. A theoretical R-D analysis has also been done for the 3D transform based video coding case. These have evidenced the need to introduce adaptivity in the signal transform by means of non-linear signal decompositions. Experimental results have validated the formulated assumptions and our theoretical analysis.

Finally, we would like to underline the following remark: Experimental results have also empirically revealed that present implementations of the adaptive wavelet representations based on the lifting scheme present a serious suboptimality due to their level by level (and even MB by MB) optimization strategy. Given the high non-linearity , the optimization problem should be posed from a global combinatorial point of view (mainly for the case of adaptive MCTF, in adaptive 3D Wavelet video coding tree based techniques are possible) in order to obtain optimal results in signal representation and compression. However, the practical feasibility of this is not clear due to the

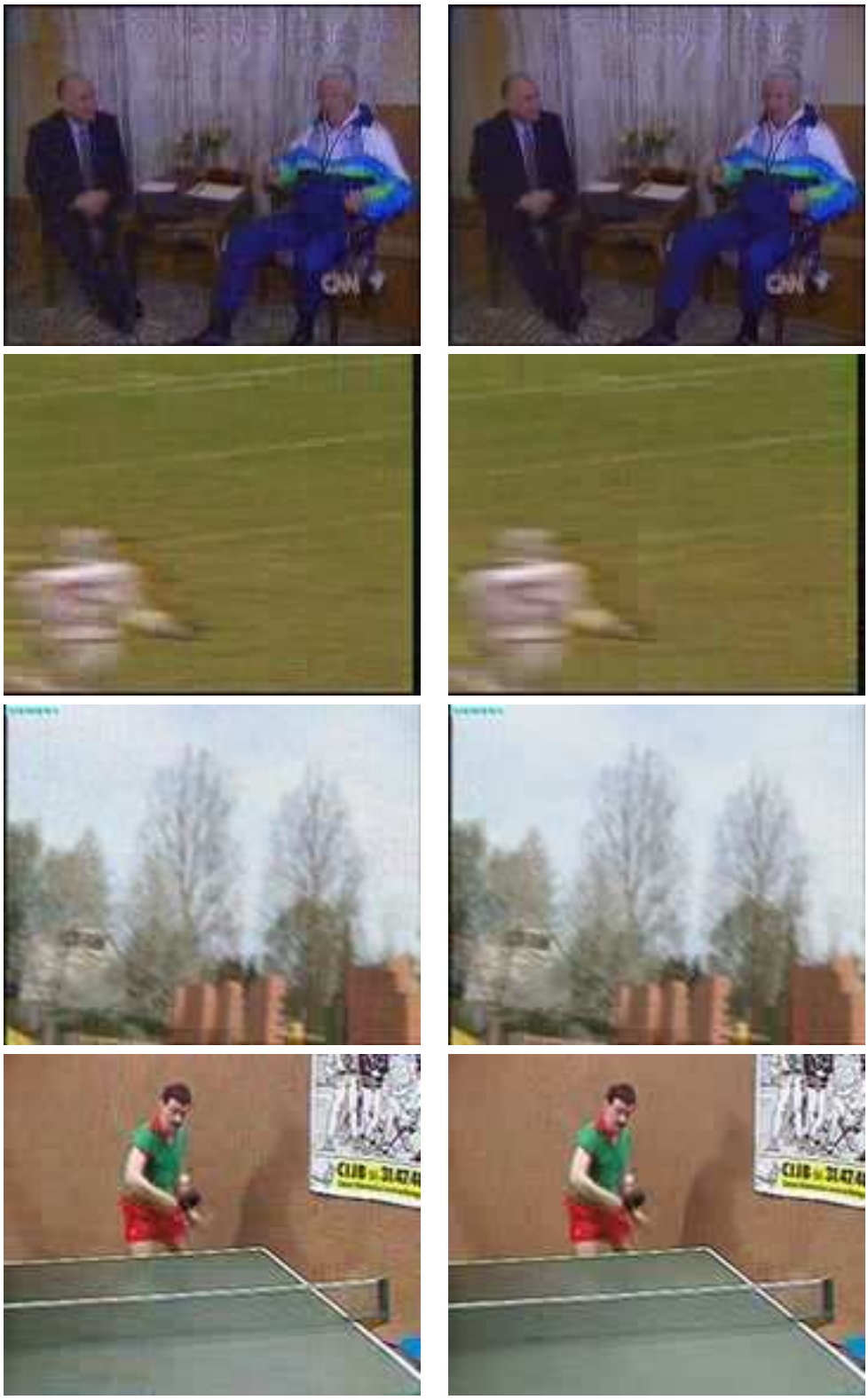


Figure 3.34: Visual quality improvement of selected frames in the test sequences. Left: no GOP Adaptivity. Right: With GOP Adaptivity. Rows from top to bottom: cnn, football, foreman, table tennis. The respective bit rates are the following (for each sequence, the first indicated rate corresponds to the non Intra-adaptive case): cnn 126.21 kbps and 124.637 kbps, football: 1093 kbps and 1007 kbps, foreman: both at 246 kbps, table: 151.38 kbps and 150.92 kbps

large number of dimensions involved in the combinatorial optimization.

In this chapter, we have observed the importance of accurate temporal geometry modeling for efficient video coding. In video signals, spatial geometry is very relevant as well. However, few video representation approaches exploit this fact. Accurate video models should take this into account, such that spatial geometry and its associated temporal motion are jointly represented. This is not an easy task, moreover new unexplored techniques relaying on the use of over-complete bases are probably required. In the next chapters, we pave the way toward sparse and geometry adapted video representations through the study of highly non-linear algorithms with redundant dictionaries, image decompositions on over-complete dictionaries and their application to video representations.

Chapter 4

Sparse Representations and Approximations on Redundant Dictionaries

4.1 Introduction

In many applications, such as compression, denoising or source separation, one often seeks an efficient representation or approximation of a signal f by means of a linear expansion into a possibly overcomplete family of functions:

$$\hat{f} = \sum_{\gamma \in \Gamma} b_\gamma g_\gamma, \tag{4.1}$$

where Γ is a set of functions with cardinality m and $\Gamma \subset \Omega$, where Ω is the set of all basic functions composing the dictionary* $\mathcal{D} = \{g_\gamma : \ \gamma \in \Omega\}$, $b_\gamma \neq 0 \ \forall \gamma \in \Gamma$, and where $\hat{f} = f$ for the case of exact representations. Therefore, $\hat{f} \in span\,(g_\gamma, \ \gamma \in \Gamma)$. In this chapter, f is assumed to be such that $f \in \mathcal{H}$, where $\mathcal{H}$ is a Hilbert Space (in this work, $\mathcal{H}$ is assumed such that $\mathcal{H} \equiv \mathbb{R}^N$).

In this setting, efficiency is often characterized by sparseness of the associated series of coefficients (i.e. the cardinality of Γ). The criterion of sparseness has been studied for a long time and in the last few years has become popular in the signal processing community (see [54, 55, 175, 176], among other). This is because sparseness, concerning the representation or approximation of a given signal, reflects the capacity of efficiently modeling and extracting the main structural components of a given signal f. Efficient signals modeling, which is certainly synonymous of energy compaction in the representation, supplies to many different applications the basics for their working principles. Let us, here, shortly discuss the importance of sparse representations and approximations for some applications:

- Compression: Although not exclusively, efficient representation of signals by sparse approximations contributes to achieve good compression performances (see for example [8, 66, 141]). Using few terms in signal approximation is, up to some degree, related with efficient signal modeling and dimensionality reduction. In signal compression, one searches for good signal

*In this thesis a dictionary is understood as a generic pool of functions (or atoms), containing all available waveforms for representing signals based on the model described in Eq. (4.1).

models which are cheap to represent in terms of bits. Typical models based on the superposition of waveforms (i.e. Eq. (4.1)), require the encoding of the set of coefficients $b_\gamma, \gamma \in \Gamma$ and the set of basis function indexes Γ. Intuitively there is no doubt that the fewer the coefficients, the cheaper to code them. However, a very critical point in signal compression is the coding cost of Γ. This may limit, up to some degree, the possible benefits of sparse signal approximations in compression. The cost of coding Γ, apart from its size, is extremely dependent on the nature and structure of the dictionary in use. Indeed, a very sparse signal model, where the appropriate set of functions Γ is very expensive to code, may be, simply, less efficient in compression than other less sparse approximations, where a simpler dictionary is used. For a given dictionary $\mathcal{D}$, one will be normally interested in having the sparsest representation possible for compression purposes. However, to ensure coding efficiency, special attention must be taken when selecting the dictionary. Such selection must be done in order to maximize sparsity, while keeping a good compromise with the coding cost of Γ.

- Denoising and Restoration: Sparse approximations using appropriate dictionaries are a powerful tool to identify the primitives necessary to represent the main structures underlaying a noisy signal. Efficient energy compaction enables to separate most of the noise from the restored signal (e.g. [96, 97]).

- Source Separation: A challenging problem is the one intending to separate two or more sources from one or several mixtures. Assuming the absence of noise, the use of sparse representations may help finding a representation of the mixtures where sources can be easily identified and separated. If noise is considered, then the problem turns into that of sparse approximations (e.g. see [167, 190]), combining at the same time, source separation and signal restoration.

The reader may have noticed that an explicit difference is made between representations and approximations. This is because they imply different problems and consequently a particular formulation is required for each one of these. In this Chapter different aspects on the use of redundant dictionaries for sparse representations and approximations are reviewed. First, in Sec. 4.2 the definition of sparse approximations and representations is reviewed. Next, the implications of using redundant or non-redundant dictionaries are over-viewed in Sec. 4.3. Since dealing with redundant dictionaries is a difficult task, Sec. 4.4 describes a set of sub-optimal (but necessary) tools and algorithms to deal with those. However, for certain cases, these tools are able to supply optimal solutions. Sections 4.5 and 4.6 describe the situations when optimal solutions can be ensured for a subset of the algorithms described in Sec. 4.4. Finally conclusions are drawn in Sec. 4.7.

4.2 Sparse Representations & Sparse Approximations

The exact representation of a signal f on a dictionary of basis functions $\mathcal{D}$ corresponds to the retrieval of the coefficients vector $\mathbf{b} \in \mathbb{R}^\Omega$ such that:

$$f = D\mathbf{b}, \tag{4.2}$$

where D is the synthesis matrix associated to the dictionary $\mathcal{D}$, i.e. every column of D corresponds to an atom in the dictionary. Hence, the sparsest representation of f on $\mathcal{D}$ corresponds to that concerning the vector $\mathbf{b}$ with the smallest support:

$$\arg\min_{\mathbf{b}} \|\mathbf{b}\|_0 \text{ s.t. } f = D\mathbf{b}. \tag{4.3}$$

Signal approximations are those where a certain error is allowed between the represented signal $\hat{f} = D\mathbf{b}$ and the original signal (f). Hence, the problem of finding the sparsest approximation of f with a maximum of m terms, such that the error norm is minimized, can be stated as:

$$\arg\min_{\mathbf{b}} \|f - D\mathbf{b}\|_2^2 \quad \text{s.t.} \quad \|\mathbf{b}\|_0 \leq m. \tag{4.4}$$

Despite Eq. (4.4), in some applications, one may prefer a slightly different formulation of the approximation problem. Nevertheless the problem remains equivalent and subject to the same limitations. Indeed, one may which to minimize the number of used coefficients given a maximum distortion requirement:

$$\arg\min_{\mathbf{b}} \|\mathbf{b}\|_0 \quad \text{s.t.} \quad \|f - D\mathbf{b}\|_2^2 \leq \eta,$$

where η defines a positive real scalar value.

4.3 Non-Redundant vs Redundant Dictionaries

4.3.1 Non-Redundant Dictionaries

Representations based on non-redundant dictionaries are a particular case of Eq. (4.2). Indeed, non-redundant means that the dictionary is composed by n functions if the space it spans has n dimensions. Let us assume that a non-redundant dictionary, with synthesis matrix A (i.e. $D = A$ in Eq. (4.2)), is in use to represent f. Then,

$$f = A\mathbf{b},$$

where A is a $n \times n$ matrix with all columns linearly independent. This defines a determined system with equal number of equations and variables. Hence,

$$\mathbf{b} = A^{-1}f, \tag{4.5}$$

where for orthonormal bases is such that $A^{-1} = A^T$.

The fact of using a non-redundant dictionary simplifies extremely problems (4.3) and (4.4). Concerning exact sparse representations, the only possible solution is given by Eq. (4.5). If f has a sparse representation on the selected dictionary, then this will be found by the inverse of the dictionary matrix. Another advantage of using non-redundant dictionaries, in particular for the case of orthonormal basis, is the one that concerns the solution of (4.4). Even though it is a non-linearly constraint problem, and even if (4.4) normally requires to solve a combinatorial problem, there exists a closed form to solve it. This converts the combinatorial problem into a set of operations with very few complexity. Indeed, *Donoho and Johnstone* showed in [56, 57] that *Shrinkage* (i.e. coefficients hard thresholding) solves, for the case of orthonormal bases, the problem of Eq. (4.4).

4.3.2 Redundant Dictionaries

When redundant dictionaries are used, the synthesis matrix is no longer a square matrix. The number of basis functions contained in the dictionary outnumbers the dimension of the Hilbert space where f lives. Hence, for a given synthesis matrix B (i.e. $D = B$ in Eq. (4.2)),

$$f = B\mathbf{b} \tag{4.6}$$

where B is a $n \times d$ matrix where $n < d$.

In such a case, **b** has no unique solution. Indeed, (4.6) forms an under-determined system of equations with an infinite number of possible solutions. Hence, the problems stated in (4.3) and (4.4) turn, forcedly, into combinatorial problems, which are NP-Hard.

The reason why it is worth considering redundant dictionaries, even if they give rise to very complex problems to solve, is due to their capacity to supply sparse representations and approximations. Given a certain class of signals, one may define a dictionary of functions such that they have a rich collection of shapes in order to adapt better to the characteristics of the signals to represent/approximate. Such kind of dictionaries are, often, redundant sets of functions able to supply sparser solutions to Eq. (4.3) and (4.4).

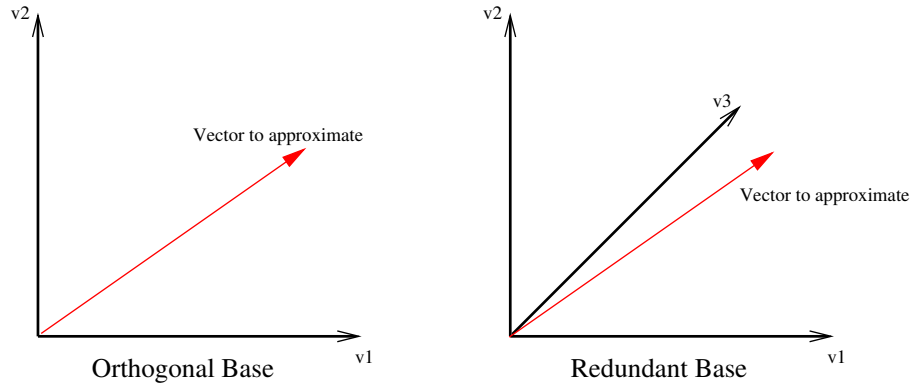


Figure 4.1: Consider the approximation of a vector (in red) by a single vector of a base. In the left example, the closest basis vector is v1. In the right example, the closest vector is v3. In the overcomplete case, the approximation error for any vector is always equal or lower than in the orthogonal case.

Fig. 4.1 depicts a graphical example where one searches to approximate the red vector by means of a single vector from the basis. In the orthogonal case, the best approximation is achieved by projecting the vector onto $v1$. In the redundant case, a much better approximation is achieved by projecting the red vector onto $v3$. In a more general scope, one can prove that for any vector belonging to $\mathbb{R}^2$, the approximation error achieved by a one term approximation is always equal or lower on the redundant base than in the orthogonal base.

4.4 Algorithmic Approaches for Sparse Representations and Approximations on Redundant Dictionaries

Solving problem (4.3) and (4.4) for any signal and any redundant $\mathcal{D}$ has non-polynomial complexity due to the non convexity of the ℓ_0 quasi-norm. Hence, different alternative approaches have been proposed in order to make computationally solvable the retrieval of a solution for **b**. This solution, however, in many cases may not be the sparsest one.

4.4.1 Method of Frames

An approach widely used to pick out a solution, among all those from an under-determined system, is to constrain this to be the one with minimum ℓ_2-norm. This approach is known under the name of Method of Frames (MOF) [38]. Geometrically speaking, MOF selects the solution that is closest to

the origin. It solves exact representations of signals by means of a modified formulation of Eq. (4.3). In this, the non convex ℓ_0 quasi-norm is substituted by a ℓ_2. Hence, the problem to solve is:

$$\arg\min_{\mathbf{b}} \|\mathbf{b}\|_2 \;\; s.t. \;\; f = D\mathbf{b}. \tag{4.7}$$

The great advantage of such a formulation is that it possesses a closed form solution. Indeed, there is a matrix that linearly calculates the minimum-length solution to a system of linear equations. This is the generalized inverse matrix or *Moore-Penrose Pseudoinverse* [25]:

$$\mathbf{b}^+ = D^+ f, \tag{4.8}$$

where $\mathbf{b}^+$ denotes the minimum norm vector of coefficient needed to synthesize f using the synthesis dictionary D and $D^+ = D^T \left(DD^T\right)^{-1}$.

However, MOF presents some problems in what concerns performance. As described in [28], MOF is not sparsity preserving, which means that even if there exist a sparse representation of f, the MOF will, very likely, not recover it. Moreover, MOF is resolution limited due to the operator DD^T. Indeed, this operator determines the size of the sharper features that MOF is able to detect in order to reconstruct signals [28].

4.4.2 Greedy algorithms

Matching Pursuits

Another way of retrieving a signal approximation or representation based on the model of Eq. (4.1) is using the so called Matching Pursuit algorithm. Matching Pursuit (MP) was first introduced by Mallat and Zhang [121, 122] as *a greedy algorithm that decomposes any signal into a linear expansion of waveforms taken from a redundant dictionary.* These waveforms are iteratively chosen to best match signal structures, producing typically a sub-optimal expansion. Vectors are selected one by one from the dictionary, by means of optimizing the signal approximation (in terms of energy) at each step. Even though the expansion is linear (see Eq. (4.1)), MP is a highly non-linear decomposition algorithm.

Let us considering the previous definition of dictionary $\mathcal{D}$. Consider a redundant dictionary $\mathcal{D}$ where atoms belong to $\mathbb{R}^N$, and with N linearly independent vectors. Let $r_k f$ be the residual of an k term approximation of a given signal $f \in \mathbb{R}^N$. A Matching Pursuit is an iterative algorithm that sub-decomposes the residue $r_k f$ by projecting it on a vector of $\mathcal{D}$ that matches $r_k f$ at best.

If we consider $r_0 f = f$, at the first iteration MP will represent the signal as:

$$f = r_0 f = \langle f, g_{\gamma_0} \rangle g_{\gamma_0} + r_1 f \,, \tag{4.9}$$

where $r_1 f$ is the residual vector after approximating $r_0 f$ in the direction of g_{γ_0}. Since $r_1 f$ is orthogonal to g_{γ_0}, the module of f will be:

$$\|r_0 f\|_2^2 = |\langle r_0 f, g_{\gamma_0} \rangle|_2^2 + \|r_1 f\|_2^2 \,. \tag{4.10}$$

As the term that must be minimized is the error $\|r_1 f\|$,

$$\|r_1 f\|_2^2 = \|r_0 f\|^2 - |\langle r_0 f, g_{\gamma_0} \rangle|^2 \,, \tag{4.11}$$

the $g_\gamma \in \mathcal{D}$ to be chosen is the one that maximizes $|\langle r_0 f, g_{\gamma_0} \rangle|$, or, generalizing, $|\langle r_k f, g_{\gamma_k} \rangle|$. This is,

$$|\langle r_k f, g_{\gamma_k} \rangle| = \sup_{\gamma \in \Omega} |\langle r_k f, g_\gamma \rangle| \,. \tag{4.12}$$

From (4.9), one easily sees by induction that the N term decomposition of f is given by:

$$f = \sum_{k=0}^{K-1} \langle r_k f, g_{\gamma_k} \rangle g_{\gamma_k} + r_K f \tag{4.13}$$

and with the same principle we can also deduce from (4.10) that the L^2 norm of the signal f is:

$$\|f\|^2 = \sum_{k=0}^{K-1} |\langle r_k f, g_{\gamma_k} \rangle|^2 + \|r_K f\|^2 \ , \tag{4.14}$$

where $\|r_K f\|$, when dealing with finite dimension signals, converges exponentially to 0 when K tends to infinity and the number of signal dimensions is finite (see [42] for a proof).

Orthogonal Matching Pursuits

Pure Matching Pursuits, with redundant dictionaries, normally needs an infinite number of iterations to capture the whole energy of a signal. However, for any signal belonging to $\mathbb{R}^N$ a set of N linearly independent atoms should be enough to represent it. This motivated a refinement of MP called Orthogonal Matching Pursuit (OMP) [139]. In this, for every newly selected atom by the MP rule, all expansion coefficients are recomputed such that the approximation error becomes orthogonal to all the selected atoms (unlike MP, where only the last selected atom is orthogonal to the residual).

Weak General Matching Pursuits

In some cases it is not computationally possible to find the optimal solution for the atom search of Eq. (4.12), and a suboptimal solution may be computed instead:

$$|\langle r_k f, g_{\gamma_k} \rangle| \geqslant \alpha \sup_{\gamma \in \Omega} |\langle r_k f, g_{\gamma} \rangle| \, , \tag{4.15}$$

where $\alpha \in (0, 1]$ is a sub-optimality factor that depends on the search method. This factor is $\alpha = 1$ when a full search method is used. Sub-optimal search based MP approaches are known under the name of Weak Matching Pursuit (Weak-MP). The reader is referred to [173] for a detailed study.

In [64, 65], a weak form of MP based on Genetic Algorithms [102] is used to handle the extremely big size of the dictionary and the impossibility to fully browse it at every iteration.

4.4.3 Linear Programing: Basis Pursuit

Chen, Donoho and Saunders [28] proposed a new paradigm in an attempt to solve sparse exact representations. In order to avoid the *non-convexity* of Eq. (4.3) and the excessive energy spreading of Eq. (4.7), they proposed to minimize the ℓ_1 norm of the coefficients:

$$\arg\min_{\mathbf{b}} \|\mathbf{b}\|_1 \ \ s.t. \ f = D\mathbf{b}. \tag{4.16}$$

This paradigm shows to offer, in many occasions, a much better behavior than Eq. (4.7) for the retrieval of sparse representations. It is even able to find, in some cases, the same solution as Eq. (4.3). The problem stated in Eq. (4.16) can be solved using polynomial-time linear programming approaches.

4.4.4 Basis Pursuit Denoising: Quadratic Programing

Another instance of problem (4.4) is given by

$$(P_0) \quad \arg\min_{\mathbf{b}} \|f - D\mathbf{b}\|_2^2 + \tau^2 \|\mathbf{b}\|_0 , \tag{4.17}$$

where τ is a positive real value which acts as a threshold parameter to limit the number of non-zero coefficients in $\mathbf{b}$. This problem is sometimes called Subset Selection in statistics. One searches for a sparse approximation of the signal f considering a trade-off between the error and the number of elements that participate to the expansion.

A possible way of overcoming the NP complexity of P_0 is to substitute the ℓ_0 quasi-norm with the convex ℓ_1 norm. This relaxation leads to problem P_1:

$$(P_1) \quad \arg\min_{\mathbf{b}} \frac{1}{2} \|f - D\mathbf{b}\|_2^2 + \gamma \|\mathbf{b}\|_1 , \tag{4.18}$$

where γ is a positive real value used as a threshold to limit the maximum ℓ_1 norm of $\mathbf{b}$. P_1 is the minimization of a convex functional that can be solved by classical Quadratic Programming methods. This relaxation is similar to that leading to the definition of the Basis Pursuit principle for the case of exact signal representation. The fact that this paradigm is called Basis Pursuit Denoising can be explained because it was introduced to adapt BP to the case of noisy data (i.e. to the approximation case) [28]. Note that if $\mathcal{D}$ is orthonormal the solution of P_1 can be found by a *soft shrinkage* of the coefficients [28, 57], while, if $\mathcal{D}$ is a union of orthonormal subdictionaries, the problem can be solved recurring to the Block Coordinate Relaxation method [155], combined with *soft shrinkage* faster than Quadratic Programming.

4.4.5 FOCUSS: A Re-Weighted Minimum Norm Algorithm

Another very interesting approach for solving exact representations is a variation of the MOF approach called *FOCal Underdetermined System Solver* (FOCUSS) [89]. This is a *nonparametric, iterative algorithm for finding localized solutions to undetermined problems with limited data.* The algorithm is iterative, and is composed of two main parts:

1. Retrieval of a low resolution estimate of the sparse signal by means of a simple MOF approach.
2. Pruning process of the first estimation using a generalized Affine Scaling Transformation (AST). That is, an iterated solution is found by scaling the entries with the solution of previous iterations.

A more formal description of the problem solved at every iteration k is the following:

$$\arg\min_{\mathbf{b}} \left\| (W_{a_k} W_{p_k})^+ \mathbf{b} \right\|_2^2 \; s.t. \; f = D\mathbf{b},$$

where W_{a_k} is a diagonal matrix which may contain some *a priori*, W_{p_k} is a diagonal matrix composed by the weights obtained from the solution retrieved in the precedent algorithm iteration and $(\cdot)^+$ denotes the pseudo-inverse.

To solve one iteration of the problem, the procedure may be split in three steps:

$$\begin{array}{lrcl} \text{Step 1:} & W_{p_k} & = & \text{diag}(\mathbf{b}_{k-1}^l) \\ \text{Step 2:} & \mathbf{q}_k & = & (D\, W_{a_k} W_{p_k})^+ f \\ \text{Step 3:} & \mathbf{b}_k & = & W_{a_k} W_{p_k} \mathbf{q}_k, \end{array} \tag{4.19}$$

where l is an user defined parameter to modify the strength of the re-weighting feed-back. For the particular choice of $l = \frac{1}{2}$, FOCUSS provides the solution to the Eq. (4.16), i.e. the same as Basis Pursuit [40, 41].

For the first iteration, W_{p_0} is assumed to be the identity matrix.

If one interprets the weights W_{p_k} like some kind of *a priori* knowledge about the solution, then an interpretation can be: the algorithm computes its own *a priori* information from iteration to iteration.

4.4.6 FOCUSS "Denoising"

Like most of the other algorithms presented in here to solve under-determined systems, FOCUSS has also an extension that tackles the situation of signals approximation. There are two main approaches:

- One is based on the "denoising" version of the MOF approach, i.e. it is based on a *Tikhonov Regularization* approach [89]:

$$\arg\min_{\mathbf{b}} \frac{1}{2} \|f - D\mathbf{b}\|_2^2 + \lambda^2 \left\| (W_{a_k} W_{p_k})^+ \mathbf{b} \right\|_2^2 .$$

- The other FOCUSS *regularized* approach is based on a simple truncation of the singular value decomposition of the re-weighted dictionary (see Step 2 in the algorithm of Sec. 4.4.5) [88, 89]. However this last approach performs the regularization assuming the signal under approximation to be stationary.

As for the exact representation case, FOCUSS "denoising" supplies the solution to the Basis Pursuit Denoising problem (Eq. (4.18)) when $l = \frac{1}{2}$ in the computation of W_{p_k} [40, 41].

4.4.7 Use of Structured Dictionaries: Retrieval of a Best Orthogonal Basis

For certain structured dictionaries, it is possible to develop specific decomposition schemes adapted to the dictionary. Wavelet Packet and Local Cosine packets are well known examples of this kind of dictionaries. Such dictionaries supply a long range of different orthogonal bases which may be adaptively selected depending on some criteria. In [33] is proposed an algorithm to select the Best Orthogonal Basis (BOB) within the collection of different bases. The performance of the algorithm will depend on the signal to approximate. Indeed if the signal has a sparse orthogonal representation BOB may work well. However, if this is not the case, then BOB will completely fail to find an efficient representation.

4.5 Recovery of Exact Sparse Representations Using Redundant Dictionaries

In this section, one finds a summary of recent theoretical results concerning the possibility for *Weak(α)*-MP/OMP and BP [95, 175] to exactly recover a given linear combination of m linearly independent atoms from a redundant dictionary $\mathcal{D} = \{g_j\}_{j\in\Omega}$. That is, if certain conditions on the dictionary are satisfied, then the solutions found by *Weak(α)*-MP/OMP and BP will be also the solutions to the problem of retrieving the sparsest exact representation (Eq. (4.3)).

Γ is defined, here, as the optimal subset of Ω that indexes the m atoms of the sparse representation (4.1) and $\overline{\Gamma}$ as the complement of Γ in Ω. Hence, D_Γ contains only the linearly independent atoms providing the exact sparsest signal representation of f and $\mathcal{D} = \mathcal{D}_\Gamma \cup \mathcal{D}_{\overline{\Gamma}}$. The dictionary matrix D has size $n \times d$, with $d \geq n$, where n is the size of the input signal f and $d = |\Omega|$.

Optimal atoms are not usually known in advance. Therefore, sufficient conditions for exact recovery are based on the internal coherence of the dictionary. A measure of this coherence is given by the cumulative coherence function [175], defined as follows:

$$\mu_1(m, \mathcal{D}) \triangleq \max_{|\Lambda|=m} \max_{i \in \Omega \setminus \Lambda} \sum_{\lambda \in \Lambda} |\langle g_i, g_\lambda \rangle| \,, \tag{4.20}$$

where $\Lambda \subset \Omega$ has cardinality m. Notice that the measure known as coherence of a dictionary (μ) and often used to characterize redundant dictionaries corresponds to the particular case of $\mu = \mu_1(1, \mathcal{D})$. Furthermore $\mu_1(m, \mathcal{D}) \leq m\mu$.

Given a signal $f = \sum_{\gamma \in \Gamma} b_\gamma g_\gamma$, MP/OMP and BP will not necessarily recover the optimal set Γ. The exact recovery of correct atoms will be only ensured if the following *Exact Recovery Condition* [175] (also called *Stability Condition* (SC) [95] for MP) is satisfied:

$$\sup_{i \notin \Gamma} \left\| D_\Gamma^+ g_i \right\|_1 < 1, \tag{4.21}$$

where $(\cdot)^+$ denotes the *Moore-Penrose Pseudoinverse*. In the case of *Weak*-MP [173], the right hand side of (4.21) is simply replaced by α (see [95, 175]). This bound is indicative of the behavior of general weak greedy algorithms and BP with an overcomplete dictionary and a sparse signal. Eq. (4.21) implies that, in order to recover the optimal functions that expand the signal f, these must be different *enough* from any other function of the dictionary not included in D_Γ. As proved in [95, 175, 176], an estimate based, exclusively, on the cumulative coherence holds:

Theorem 4.1 *(Tropp [175],*Gribonval and Vandergheynst [95]*) Let μ_1 be the cumulative coherence function of $\mathcal{D}$ and m is a positive integer such that*

$$\mu_1(m) + \mu_1(m-1) < 1. \tag{4.22}$$

Then, for any index set Γ of size at most m and any $f \in span\,(g_\gamma, \gamma \in \Gamma)$, Eq. (4.21) holds. This is a sufficient condition for Basis Pursuit to recover the optimal representation of a $(\mathcal{D}, m)$-sparse signal f. Moreover, if $\alpha > \mu_1(m)/(1 - \mu_1(m-1))$, then Weak-*MP picks up a correct atom $g \in \Gamma$ at each step.*

It is important to stress that Theorem 4.1 provides a pessimistic bound. There are many cases in which (4.22) is not respected but indeed MP or BP would find the sparsest solution.

4.6 Recovery of General Signals Using Redundant Dictionaries: Sparse Approximations

Akin to the exact representation case, finding a solution to (4.4), when overcomplete dictionaries are used, may be rather difficult or even practically infeasible. Typically used sub-optimal algorithms, like Matching Pursuit algorithms (*Weak*-MP [95]) or ℓ_1-norm Relaxation Algorithms (BPDN [28]), do not necessarily compute the solution of problem (4.4). However, there exist particular situations in which they succeed in recovering the "correct" solution, i.e. the set of atoms giving the sparsest m-term approximation. Very important results have been found for the case where incoherent dictionaries are used.

Prior to reviewing the results introduced above, let us define a series of elements that will be used in the remaining of the section:

- f_m^{opt} is the best m-term approximant of f such that $f_m^{opt} = D\mathbf{c}_{opt}$ where the support of $\mathbf{c}_{opt}$ is smaller than or equal to a positive integer m.
- Given $k \geq 0$, r_k and f_k are the residual and approximant generated by a greedy algorithm at its nth iteration.
- Γ_m is the optimal set of m atoms that generate f_m^{opt}. Often, for the sake of simplicity, this will be referred as Γ.
- The projection of f over the atoms in a set Λ is called $a_\Lambda = DD_\Lambda^+ f$.

4.6.1 Greedy Algorithms: *Weak*-MP

Gribonval and Vandergheynst [95] extended the results of Tropp [175] for the particular case of Orthogonal Matching Pursuit (OMP) to the general *Weak*-MP. Akin to the case of signal representations, the main results consist in the sufficient conditions that guarantee that *Weak*-MP will recover the optimal set of atoms that generate the best m-term approximant f_m^{opt}. A result establishes also an upper bound on the decay of the residual energy in the approximation of a signal that depends on the internal coherence of $\mathcal{D}$. Moreover, a bound is found on how many "correct" iterations can be performed by the greedy algorithm depending on the dictionary and the energy of f_m^{opt}.

Robustness

The sufficient conditions found in [95] that ensure that *Weak*-MP will recover the set of atoms that compose the best m-term approximant are enounced in Theorem 4.2. First of all, it is necessary that the optimal set Γ_m satisfies the Stability Condition [95]. If, in addition, some conditions are satisfied concerning the remaining residual energy at the kth iteration ($\|r_k\|_2^2$) and the optimal residual energy $\|r_m^{opt}\|_2^2$, then an additional atom belonging to Γ_m will be recovered. This condition, called originally the General Recovery Condition in [175], was named, for the case of general *Weak*-MP, the Robustness Condition in [95].

Theorem 4.2 *(Gribonval & Vandergheynst [95]) Let $\{r_k\}_{k\geq 0}$ be a sequence of residuals computed by General MP to approximate some $f \in \mathcal{H}$. For any integer m such that $\mu_1(m-1) + \mu_1(m) \leq 1$, let $f_m^{opt} = \sum_{\gamma\in\Gamma_m} c_\gamma g_\gamma$ be a best m-term approximation to f, and let $N_m = N_m(f)$ be the smallest integer such that*

$$\|r_{N_m}\|_2^2 \leq \left\|r_m^{opt}\right\|_2^2 \cdot \left(1 + \frac{m \cdot (1 - \mu_1(m-1))}{(1 - \mu_1(m-1) - \mu_1(m))^2}\right). \tag{4.23}$$

Then, for $1 \leq k < N_m$, General MP picks up a "correct" atom. If no best m-term approximant exists, the same results are valid provided that $\|r_m^{opt}\|_2 = \|f - f_m^{opt}\|_2$ is replaced with $\|f - f_m^{opt}\|_2 = (1+\eta)\|r_m^{opt}\|_2$ in (4.23), where $\eta \geq 0$ is a sub-optimality factor.

Rate of Convergence

In the following the main result concerning the exponential decay of the error energy bound, as well as the bound on how many "correct" iterations can be performed by the greedy algorithm, is reviewed.

Theorem 4.3 *(Gribonval & Vandergheynst [95]) Let $\{r_k\}_{k\geq 0}$ be a sequence of residuals computed by General MP to approximate some $f \in \mathcal{H}$. For any integer m such that $\mu_1(m-1)+\mu_1(m) \leq 1$, we have that*

$$\|r_k\|_2^2 - \left\|r_m^{opt}\right\|_2^2 \leq \left(1 - \frac{1-\mu_1(m-1)}{m}\right)^{n-l} \left(\|r_l\|_2^2 - \left\|r_m^{opt}\right\|_2^2\right). \tag{4.24}$$

Moreover, $N_1 \leq 1$, and for $m \geq 2$:

- *if $\|r_m^{opt}\|_2^2 \leq 3\|r_1\|_2^2/m$, then*

$$2 \leq N_m < 2 + \frac{m}{1-\mu_1(m-1)} \cdot \ln \frac{3 \cdot \|r_1\|_2^2}{m \cdot \|r_m\|_2^2} \tag{4.25}$$

- *else $N_m \leq 1$.*

4.6.2 Convex Relaxation of the Subset Selection Problem

In [176], the author studies the relation between the subset selection problem (4.17) and its convex relaxation (4.18), and shows that any coefficient vector which minimizes Eq. (4.18) is supported inside the optimal set of indexes if the following condition is satisfied:

Theorem 4.4 *(Correlation Condition, Tropp [176]) Suppose that the maximum inner product between the residual signal and any atom satisfies the condition*

$$\|D^*(f - \mathbf{a}_\Lambda)\|_\infty < \gamma(1 - \sup_{i\notin\Lambda} \|D_\Lambda^+ g_i\|_1).$$

Then any coefficient vector $\mathbf{b}_*$ minimizing the function (4.18) *must satisfy support$(\mathbf{b}_*) \subset \Lambda$.*

In particular, the relationship between the trade-off parameters τ and γ is studied, proving that if the coefficient vector $\mathbf{b}_*$ minimizes the function (4.18) with threshold $\gamma = \tau/(1-\sup_{i\notin\Gamma} \left\|D_\Gamma^+ g_i\right\|_1)$, then the relaxation never selects a non optimal atom and the solution of the convex relaxation is unique.

Theorem 4.5 *(Tropp [176]) Suppose that the coefficient vector $\mathbf{b}_*$ minimizes the function* (4.18) *with threshold $\gamma = \tau/(1-\sup_{i\notin\Gamma} \left\|D_\Gamma^+ g_i\right\|_1)$. Then we have that:*

1. *the relaxation never selects a non optimal atom since support$(\mathbf{b}_*) \subset$ support$(\mathbf{c}_{opt})$.*
2. *The solution of the convex relaxation is unique.*
3. *The following upper bound is valid:*

$$\|\mathbf{c}_{opt} - \mathbf{b}_*\|_\infty \leq \frac{\tau \cdot \left\|\left(D_\Gamma^* D_\Gamma\right)^{-1}\right\|_{\infty,\infty}}{1 - \sup_{i\notin\Gamma} \|D_\Gamma^+ g_i\|_1}. \tag{4.26}$$

4. *The support of $\mathbf{b}_*$ contains every index j for which*

$$|\mathbf{c}_{opt}(j)| > \frac{\tau \cdot \left\|\left(D_\Gamma^* D_\Gamma\right)^{-1}\right\|_{\infty,\infty}}{1 - \sup_{i\notin\Gamma} \|D_\Gamma^+ g_i\|_1}. \tag{4.27}$$

If the dictionary we are working with is orthonormal it follows that

$$\sup_{i\notin\Gamma} \|D_\Gamma^+ g_i\|_1 = 0 \text{ and } \left\|\left(D_\Gamma^* D_\Gamma\right)^{-1}\right\|_{\infty,\infty} = 1,$$

making the previous theorem to become much stronger. In particular we obtain that $\|\mathbf{c}_{opt} - \mathbf{b}_*\|_\infty \leq \tau$ and $|\mathbf{c}_{opt}(j)| > \tau$ [57, 176].

4.7 Conclusions

In this chapter, an introduction to the general framework of signals representations and approximations through the use of redundant dictionaries is presented. The problematic of retrieving sparse solutions to under-determined problems is discussed. The main tools used for this purpose have been described. These tools are just relaxations of the initial sparse problems. Hence, they are not guaranteed to find the optimal solution to sparse representations and approximations. Nevertheless some recent results claim that if incoherent enough dictionaries are used, then the recovery of some sparse solutions can be guaranteed.

In fact, from all these results, one infers that the use of incoherent dictionaries is very important for the good behavior of greedy and ℓ_1-norm relaxation algorithms. However, experience seems to teach us that highly redundant and often coherent dictionaries are more powerful for natural signals approximation. In the next chapter, this problem is solved by carefully studying the relationship between the signal and the dictionary, through the introduction of *a priori* information in the decomposition process.

Chapter 5

Using *a Priori* Models in Sparse Representations and Approximations

5.1 Motivation

In general, the problem of recovering the sparsest signal approximation (or representation) over a redundant dictionary is a NP-hard problem. However, this does not impair the possibility of solving this problem when *particular* classes of dictionaries are used.

As demonstrated in [54, 95, 175, 176], and reviewed in Chapter 4, in order to ensure the good behavior of algorithms like General *Weak*(α) Matching Pursuit (*Weak*-MP), BP and Basis Pursuit Denoising (BPDN), dictionaries need to be incoherent enough. Under this main hypothesis, sufficient conditions have been stated so that these methods are able to recover the atoms from the sparsest m-term expansion of a signal.

However, experience and intuition dictate that good dictionaries for sparse approximations of natural signals can be very redundant and, depending on the kind of signal structures to describe, they may be highly coherent. This is a strong discrepancy between theory and practice.

In this chapter we explore a way of using more coherent dictionaries with *Weak*-MP and Basis Pursuit (BP), while keeping the possibility of recovering the optimal solution. This is done by the use of *a priori* information about the signal to decompose. Examples of this are the Weighted-MP, Weighted-BP and Weighted-BPDN algorithms [48, 49]. We discuss the potentiality of using *a priori* knowledge in the atom selection procedure for sparse representations and approximations in a similar way as also suggested previously in [89]. We do not treat here the issue of how to find a reliable and useful *a priori* knowledge about a signal. This problem strongly depends on the nature of the signal and on the kind of dictionary used. We, nevertheless, give an insight through a realistic example in Section 5.8. The aim of this chapter is the theoretical study of the weighted algorithms in the prospective of achieving sparseness and to compare this with the *state of the art*. Existing results in literature have been obtained for the non-model based case.

This chapter is structured as follows. First, a review of the influence of *a priori* information in sparse representations using MP and BP is discussed in Sections 5.2, 5.3, 5.4, 5.5. Later, the remaining of the Sections are consecrated to the study of the influence of *a priori* models for sparse

approximations. Finally, conclusions are drawn in Sec. 5.9.

5.2 Including *A Priori* Information: Influence on Exact Sparse Representations

In chapter 4 we saw that when using redundant dictionaries there is not a unique signal decomposition. This makes the recovery of the sparsest representation difficult for an algorithm such as *Weak*-MP or BP. However, this can be theoretically ensured when sufficiently incoherent dictionaries are in use. In this section we prove that, if some valuable *a priori* information about the signal to expand is available, the class of dictionaries where BP and *Weak*-MP are ensured to recover the exact optimal solution can be enlarged. The *a priori* knowledge establishes in advance a likelihood for any atom in the dictionary to appear into the representation of a given signal f. This is achieved by suitably weighting the atoms in the dictionary in order to reflect their relevance for the signal f.

Definition 5.1 *A weighting matrix $W = W(f, \mathcal{D})$ is a square diagonal matrix of size $d \times d$. Each of the entries $w_i \in (0, 1]$ from the diagonal corresponds to some measure of the* a priori *likelihood of a particular atom $g_i \in \mathcal{D}$ to be part of the sparsest decomposition of f.*

Weights in matrix W are not arbitrary and are not supposed to be independently and blindly optimized by the algorithm during the subset selection procedure. These values alone are not meant to determine whether an atom shall be included in the selection or not. Weights introduce a fuzzy likelihood according to some auxiliary measure issued from a model that establishes a relationship between data and dictionary (these can even establish some inter-dependence among different subdictionaries). Matrix W values can be obtained by optimization within the subset selection algorithm according to, for example, some parametric signal-dictionary model.

The way $W(f, \mathcal{D})$ should be obtained is particular for each kind of problem and dictionary. Hence, this is not treated in here. Nevertheless, we supply several illustrative examples that will be progressively introduced along this chapter. In the following we will use W_Γ and $W_{\overline{\Gamma}}$ to indicate the diagonal weighting matrices corresponding to D_Γ and $D_{\overline{\Gamma}}$ respectively. It is now possible to define a coherence measure that generalizes Eq. (4.20), where *a priori* information is also taken into account: the *Weighted Cumulative Coherence* function.

Definition 5.2 *The weighted cumulative coherence function of $\mathcal{D}$ is defined as the following data dependent coherence measure:*

$$\mu_1^w(m, \mathcal{D}, f) \triangleq \max_{|\Lambda|=m} \max_{i \in \Omega \setminus \Lambda} \sum_{\lambda \in \Lambda} |\langle g_\lambda, g_i \rangle| \cdot w_\lambda \cdot w_i. \tag{5.1}$$

The weighted cumulative coherence introduces the idea of weighting the correlations among atoms with respect to the *a priori* information we have on f. This new coherence measure considers the fact that all functions from the dictionary do not have the same probability to appear in the signal expansion. Indeed, it is of no use to consider atoms that are not likely to appear in the representation of a given signal, as they would artificially increase the value of μ_1.

5.2.1 Influence of *a Priori* Information on *Weak*-MP: Using Weighted-MP

As seen in chapter 4, General Matching Pursuits [95, 122, 173] iteratively build m-term approximants by selecting at each step the most appropriate term from $\mathcal{D}$ according to a certain rule. Each one of these iterations can be seen as a two step procedure:

1. A selection step where an atom $g_{i_k} \in \mathcal{D}$ is chosen (where $k \geq 0$ indicates the iteration number).
2. A projection step where an approximant $f_m \in span(g_{i_k} \ : \ k \in \{0, ..., m-1\})$ and a residual $r_m = f - f_m$ are generated.

The selection step, at iteration k, can be generally formulated as the maximization of a similarity measure $C(r_k, g_i)$ between the signal to approximate (the residual at the kth iteration: $r_k = f - f_k$) and the dictionary atoms:

$$g_{i_k} = \arg\max_{g_i \in \mathcal{D}} C(r_k, g_i). \tag{5.2}$$

Pure Matching Pursuit uses the modulus of the scalar product as similarity measure, i.e. $C(r_k, g_i) = |\langle r_k, g_i \rangle|$. More generally, Weak-MP allows an additional flexibility factor $\alpha \in (0, 1]$ allowing the selected atom g_{i_k} to be such that $|\langle r_k, g_{i_k} \rangle| \geq \alpha \sup_{i \in \mathcal{D}} |\langle r_k, g_i \rangle|$. The sub-optimality factor α, as demonstrated in [173], does not necessarily prevent the greedy algorithm from converging to a solution (i.e. $\lim_{k \to \infty} \|r_k\|_2^2 = 0$). However, $\alpha < 1$ often affects negatively the speed of convergence of $\|r_k\|_2^2$.

The projection step determines whether Matching Pursuit (MP) or Orthogonal Matching Pursuit (OMP) is in use. The former just guarantees that the atom selected at iteration k is orthogonal to the residual r_k [122]. The latter, constructs the approximant f_k by finding an orthogonal projection of f over the space spanned by all selected atoms until iteration k [139].

The use of the scalar product as similarity measure in *Weak*-MP bears some similarity with searching for the atom g_{i_k} with "Maximum Likelihood" given the residual r_k: the atom g_{i_k} that maximizes the probability $p(g_i|r_k)$ is selected. Thus, $|\langle r_k, g_i \rangle|$ may be intuitively seen as a measure of the conditional probability $p(r_k|g_i)$, and when all g_i are equally probable, maximizing $|\langle r_k, g_i \rangle|$ is equivalent to maximizing $p(g_i|r_k)$. Let us now consider the case where atoms do not have the same *a priori* probability to appear in the optimal set of m atoms (Γ_m). Indeed, we assume that we have at our disposal a *prior* knowledge about the likelihood of each g_i. By means of the Bayes' Rule, when some *a priori* $p(g_i)$ is available, the probability to maximize becomes

$$p(g_i|r_k) = \frac{p(r_k|g_i)\, p(g_i)}{p(r_k)}, \tag{5.3}$$

where the denominator is normally assumed to be constant for any signal r_k. Emulating this, the selection rule of MP can, thus, be modified multiplying the modulus of the scalar product by a weighting factor $w_i \in (0, 1]$, which depends on the atom index i. This is done in order to represent the insertion in the MP selection criteria of some heuristic measure of prior information. Hence, now $C(r_k, g_i)$ in Eq. (5.2) can be considered such that:

$$C(r_k, g_i) = |\langle r_k, g_i \rangle| \cdot w_i. \tag{5.4}$$

We call this family of weighted greedy algorithms Weighted-MP. The Weighted-MP approach does not modify the projection step of the algorithm, allowing to freely select the MP or OMP projection strategy. For the sake of simplicity, Weighted-MP will be used in the remaining of the paper as a general term to refer to both projection approaches. The kind of projection will not be specified unless judged to be relevant. In this work, we assume for simplicity that the *a priori* knowledge ($w_i \ \forall i \in \Omega$) is independent of the iteration of the greedy algorithm* (hence, $\forall k \ p(r_k) = constant$).

The reader may have noticed that the relation between Eq. (5.3) and Eq. (5.4) is not straightforward, and mainly based on intuition. This does not introduce any loss of generality, nor affects

*However, one could decide to update the atom weights at every iteration, leading to take also into account, in some way, $p(r_k)$. This would introduce more flexibility in the formulation of Weighted-MP

the theoretical and practical analysis performed in the following of the chapter. Nevertheless, let us briefly state in here the formal relation between both, (5.3) and (5.4), equations based on some probabilistic assumptions.

Let us model the residue obtained at iteration k if r_{k-1} is approximated with $g_i \ \forall i \in \Omega$ (i.e. $r_k = r_{k-1} - g_i\langle r_{k-1}, g_i\rangle$) as an iid Gaussian set of random variables. Even if this is not the best model for the residual data, it will allow us to derive some formal formulation relating (5.3) and (5.4). If we assume the Gaussian model for r_k, then

$$p(r_k|g_i) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \cdot \exp\left(-\frac{\|r_{k-1} - g_i\langle r_{k-1}, g_i\rangle\|_2^2}{2\sigma_r^2}\right).$$

Considering this, and that $\forall k \ p\left(r_k\right) = constant$, we thus formulate the selection step of our probabilistic greedy algorithm as:

$$\begin{aligned}
g_{i_k} &= \arg\max_{g_i \in \mathcal{D}} \left(p(r_k|g_i)p(g_i)\right) \\
&= \arg\max_{g_i \in \mathcal{D}} \left(\frac{1}{\sqrt{2\pi\sigma_r^2}} \cdot \exp\left(-\frac{\|r_{k-1} - g_i\langle r_{k-1}, g_i\rangle\|_2^2}{2\sigma_r^2}\right) p(g_i)\right) \\
&= \arg\max_{g_i \in \mathcal{D}} \left(\frac{1}{\sqrt{2\pi\sigma_r^2}} \cdot \exp\left(-\frac{\|r_{k-1}\|_2^2 - |\langle r_{k-1}, g_i\rangle|^2}{2\sigma_r^2}\right) p(g_i)\right) \\
&= \arg\max_{g_i \in \mathcal{D}} \left(\frac{1}{\sqrt{2\pi\sigma_r^2}} \cdot \exp\left(\frac{|\langle r_{k-1}, g_i\rangle|^2}{2\sigma_r^2}\right) p(g_i)\right) \\
&= \arg\max_{g_i \in \mathcal{D}} \left(|\langle r_{k-1}, g_i\rangle|^2 + \lambda \log\left(p(g_i)\right)\right), \qquad (5.5)
\end{aligned}$$

where λ is a constant that depends on σ_r^2.

In order to get closer to (5.4) formulation, Eq. (5.5) can be operated such that:

$$g_{i_k} = \arg\max_{g_i \in \mathcal{D}} \left(|\langle r_{k-1}, g_i\rangle| \sqrt{1 + \lambda \frac{\log\left(p(g_i)\right)}{|\langle r_{k-1}, g_i\rangle|^2}}\right). \qquad (5.6)$$

Notice that (5.6) requires that $\lambda \dfrac{\log\left(p(g_i)\right)}{|\langle r_{k-1}, g_i\rangle|^2} \geq -1$. Comparing (5.2), (5.4) and (5.6), one finds that w_i is related to $\sqrt{1 + \lambda \dfrac{\log\left(p(g_i)\right)}{|\langle r_{k-1}, g_i\rangle|^2}}$.

With no doubt, as seen later in this chapter, the challenging point of using Weighted-MP, is the retrieval of practical $w_i \ \forall i$ (or $p\left(g_i\right) \ \forall i$), which are adapted for a particular application.

Other interpretations of Weighted-MP are also possible without changing its analysis and behavior. For example, one may interpret Weighted-MP as a greedy algorithm where the use of non-unit norm atoms within the dictionary is allowed. Unit norm atoms are re-weighted according to some heuristic measure of prior information, which gives some hint about their likelihood to belong to the optimal set Γ. Also, another possible interpretation is the one that assumes the measure $C\left(r_k, g_i\right)$ to be based on an anisotropic norm. This norm has a scaling parameter (w_i) that depends on the measured direction of the signal space (i.e. each $g_i \in \mathbb{R}^N$ determines a line in $\mathbb{R}^N$).

The following theorem establishes the *Exact Recovery Condition* for Weighted-MP/OMP. We can see in it, as well as through all the chapter, how Weighted-MP is able to perform better than Pure MP even if our weighted algorithm is a sort of Weak Greedy Algorithm.

Theorem 5.1 *Given an* a priori *matrix* $W(f,\mathcal{D})$ *and a sub-optimality search factor* $\alpha \in (0,1]$*, then, for any index set* Γ *such that* $f \in span(g_\gamma, \gamma \in \Gamma)$*, Weighted-MP/OMP will recover a "correct" atom at each iteration if*

$$\sup_{g_i \in D_{\overline{\Gamma}}} \left\| \left(D_\Gamma W_\Gamma\right)^+ g_i \cdot w_i \right\|_1 < \alpha. \tag{5.7}$$

The proof issues from introducing the usage of *a priori* knowledge in the method developed in [175]. This may be found in Appendix B.1.

Theorem 5.1 states, as depicted by (5.7), that the use of *a priori* weights will help meeting the sufficient condition that guarantees that a greedy algorithm will recover the elements of the sparsest representation of f. Indeed, as can be observed in (5.7), given a dictionary and an appropriate W_Γ associated to f, the weights that multiply each $g_i \in D_{\overline{\Gamma}}$ may help reducing the supremum in (5.7) compared to Eq. (4.21).

5.2.2 Influence of *a Priori* Information on BP

The BP principle, as shown in [49], can also exploit the usage of *a priori* information by means of a weighting matrix. This can be done using a formulation closely related to that proposed to introduce *a priori* knowledge in re-weighted minimum norm algorithms (*FOCUSS* [89], see Sec. 4.4.5). A variation of BP taking into account the likelihood matrix $W(f,\mathcal{D})$ is given by the Weighted Basis Pursuit principle, introduced by Granai in [91]. This method, previously also suggested in [89], minimizes the ℓ_1 norm of a weighted vector, leaving the constraints unchanged:

$$\arg\min_{\mathbf{b}} \|W^{-1}\mathbf{b}\|_1 \quad \text{s.t.} \quad \mathcal{D}\mathbf{b} = f. \tag{5.8}$$

We recall that the entries of $W(f,\mathcal{D})$ are in $(0,1]$. In this way the atoms with low probability to be selected are penalized by inducing a small weighting factor in W. WBP, just like BP, is nothing but a Linear Programming problem [28, 91].

In the same way as for the case of greedy algorithms, a theorem was derived in [49] following the steps first appeared in [175], that establishes a sufficient condition for WBP to recover an exact sparse superposition of m atoms from $\mathcal{D}$ given $W(f,\mathcal{D})$. This states the following:

Theorem 5.2 *Given a dictionary* $\mathcal{D}$ *and an* a priori *matrix* $W(f,\mathcal{D})$*, Weighted Basis Pursuit recovers the optimal representation of a sparse signal* $f = D_\Gamma \mathbf{b}_{opt}$ *if:*

$$\sup_{g_i \in D_{\overline{\Gamma}}} \left\| \left(D_\Gamma W_\Gamma\right)^+ g_i \cdot w_i \right\|_1 < 1. \tag{5.9}$$

The reader may find the proof in [49].

A single sufficient condition is, thus, available for both WBP and Weighted-MP/OMP, for recovering the "correct" set of atoms involved in the optimal representation of a signal. One can see this condition as a mere proof that results [175] also extend for dictionaries where atoms may have a norm smaller than one (i.e. assuming matrix $D_\Gamma W_\Gamma$ to be the dictionary instead of just D_Γ). However, the value of the present result goes beyond such consideration. This comes from the fact that W_Γ is not a meaningless diagonal matrix with the values smaller or equal than one. It is formally demonstrated that even if such a matrix is introduced, Tropp bounds continue to be valid. Being able to embed in W_Γ additional modeling criteria for the expansion procedure, one may obtain a better signal representation being at the same time, as will be discussed in the following, less constraint by the coherence of the dictionary.

5.3 Exact Recovery Bounds for Weighted Greedy and BP Algorithms

In the case where no weights are used, usually, the optimal atoms are not known in advance and so the *Exact Recovery Condition*, Eq. (4.21), can only be verified *a posteriori*, i.e. once the optimal set of atoms has already been found. Because of that, sufficient recovery conditions based on the internal coherence (μ_1) of the dictionary were proved (see Theorem 4.1 and [95, 174]). In the following, we provide a result that takes into account the fact of using *a priori* information. This result can be compared to that of Theorem 4.1 and underlines the behavior of a greedy and BP algorithm when using an *a priori* model. This shows how, depending on the suitability of the *a priori* knowledge about the signal, a weighted algorithm will outperform the classical non-weighted approach or, to the contrary, in which cases this will be worse. In our case, the exact quantitative information concerning the "reliability" of the *a priori* (this is, as defined later in this section, ϵ_{max}) will, usually, not be available. However, the following results theoretically justify the use of signal adapted algorithms by means of an *a priori* to obtain sparser representations.

5.3.1 Sufficient Condition for Exact Expansions Recovery

Theorem 5.3 *Let $W(f,\mathcal{D})$ be the data dependent weighting matrix and let $\epsilon_{max} \triangleq \sup_{\gamma\in\Gamma} \left|1-w_\gamma^2\right|$. If, for any index set Γ of size at most m, such that $f=\sum_{\gamma\in\Gamma} b_\gamma g_\gamma$, we have*

$$\mu_1^w(m)+\mu_1^w(m-1)<1-\epsilon_{max}, \tag{5.10}$$

then (5.9) holds and WBP recovers the optimal representation of the sparse signal f. Furthermore, if

$$\frac{\mu_1^w(m)}{1-(\mu_1^w(m-1)+\epsilon_{max})}<\alpha \tag{5.11}$$

*is also enforced, then (5.7) holds and Weighted-*Weak*(α) MP will pick up an atom belonging to the optimal set Γ at each step. Moreover, Weighted-*Weak*(α) OMP will exactly recover the sparsest representation of f.*

See Appendix B.2 for a proof.

We will say that the *a priori* is "reliable" when ϵ_{max} is small enough. Hence the following can be stated:

Definition 5.3 *ϵ_{max} is close to zero if "good" atoms (the ones belonging to Γ) are not penalized by the* a priori. *In such a case we state that the* a priori *knowledge is "reliable".*

Since $\mu_1^w(m)\leq\mu_1(m)$, one can intuitively see that a reliable *a priori* knowledge can help a greedy algorithm or BP when the dictionary does not satisfy the hypothesis of Theorem 4.1. This will be possible when the weights corresponding to the atoms in $D_{\overline{\Gamma}}$ are sufficiently small.

Since $\mu_1^w(m)\leq\mu_1(m)$, we claim that considering reliable *a priori* information can help a dictionary unable to satisfy Theorem 4.1 recover the right set of functions. In other words, the use of an *a priori* model within the expansion algorithm allow for using less incoherent dictionaries.

Corollary 5.1 *Given a dictionary $\mathcal{D}$ and the data dependent diagonal matrix $W(f,\mathcal{D})$, where $w_i\in(0,1]$, we can state the following:*

- *For a Weighted MP/OMP with weakness $\alpha = 1$ and WBP a better behavior in the recovery of exact sparse representations is expected with respect to the classical algorithms if:*

$$\mu_1^w(m) + \mu_1^w(m-1) < 1 - \epsilon_{max}$$
$$\textit{and}$$
$$\mu_1(m) + \mu_1(m-1) \geq 1.$$

- *For a Weighted* Weak-*MP a better behavior in the recovery of exact sparse representations is expected with respect to the classical algorithms if:*

$$\frac{\mu_1^w(m)}{1 - (\mu_1^w(m-1) + \epsilon_{max})} < \alpha$$
$$\textit{and}$$
$$\frac{\mu_1(m)}{1 - \mu_1(m-1)} \geq \alpha.$$

Corollary 5.2 *When no* a priori *information is available (i.e. $W(f, \mathcal{D}) = I$), and consequently $\epsilon_{max} = 0$ Theorem 5.3 simplifies to the results found in [95, 174] stated in Theorem 4.1.*

5.3.2 A Toy Example for MP in $\mathbb{R}^3$

Let us consider the following overcomplete dictionary in $\mathbb{R}^3$:

$$D = \begin{pmatrix} 0 & -0.9806 & 0.4472 & -0.5774 \\ 1 & -0.1961 & 0 & 0.5774 \\ 0 & 0 & 0.8944 & -0.5774 \end{pmatrix}.$$

A simple m-sparse signal f is considered with $m = 2$ and defined as:

$$f = 3 \cdot D_0 + 3.059412 \cdot D_1, \tag{5.12}$$

i.e. the optimal set is $\Gamma = \{D_0, D_1\}$. A general graphical representation of $\mathcal{D}$ and f in $\mathbb{R}^3$ can be observed in Fig. 5.1 where the non-orthogonality among vectors can be clearly appreciated.

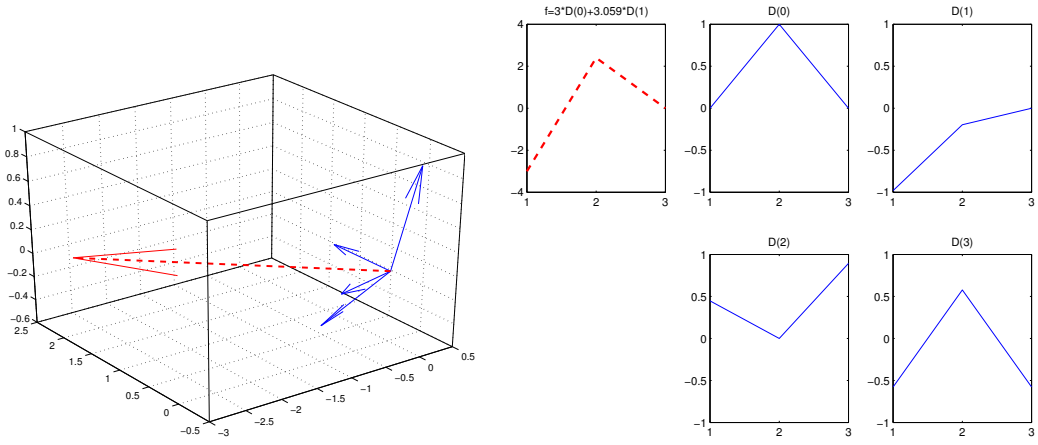


Figure 5.1: Left: 3D representation of the overcomplete dictionary (4 components) and the sparse signal f in $\mathbb{R}^3$. Right: Temporal representation of the signal and dictionary atoms.

According to the coherence measure μ_1, this dictionary has a high coherence, i.e. $\mu_1(1) = 0.7746$. This turns into a complete failure of the sufficient condition (4.22). Indeed, $\mu_1(2) + \mu_1(1) = 2.1265$

which is far above the bound with $\alpha = 1$ required to guarantee the recovery of the optimal set of atoms for any f. As a consequence, MP "derails".

The sequence of atoms selected from the dictionary for pure MP is:

```
MP:
            Step  1:    select = 3    |    Step  6:    select = 1
            Step  2:    select = 2    |    Step  7:    select = 2
            Step  3:    select = 1    |    Step  8:    select = 0
            Step  4:    select = 0    |    Step  9:    select = 1
            Step  5:    select = 2    |    Step 10:    select = 2
```

where the selected 0, 1, 2, 3 are the indexes of D_i.

Let us now consider the possibility that, by some means, it is feasible* to estimate that signal f has around 60% of chances to be embedded in the xy plane. This implies that the scalar products by the vectors D_2 and D_3 can be penalized. Thus, the following weighting matrix can be generated:

$$W(f, \mathcal{D}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0.6 \end{pmatrix}.$$

Notice the assumption that our *oracle*† does not penalize the two vectors implied in the sparsest representation of f ($w_i = 1 \ : \ i = 0, 1$).

Shifting now to the framework of Weighted-MP, the weighted cumulative coherence measure indicates that the effective internal coherence of the dictionary is reduced up to $\mu_1^w(2) = 0.3464$. Moreover the new bound, considering the *a priori*, reads $\mu_1^w(2) + \mu_1^w(1) = 0.9717$, meeting the sufficient requirement to ensure the recovery of the optimal set of vectors Γ. This time, the sequence of atoms selected is quite different and the Weighted-MP algorithm selects only the atoms belonging to the optimal set:

```
Weight-MP:
            Step    1:    select = 1    |    Step  6:    select = 0
            Step    2:    select = 0    |    Step  7:    select = 1
            Step    3:    select = 1    |    Step  8:    select = 0
            Step    4:    select = 0    |    Step  9:    select = 1
            Step    5:    select = 1    |    Step 10:    select = 0
```

Tests on these examples have been performed with the BP and WBP [49] paradigm as well. For this particular case, however, both are able to recover the optimal set of atoms independently of the fact that for BP the sufficient condition of Theorem 4.1 was not fulfilled.

*The present example must be understood as a toy example which only purpose is that of illustrating some of the concepts here explained. Hence, the reader should not worry at this point about how can, in this example, the *a priori* be obtained in practice. More realistic examples are described later in this chapter, where one can *really* establish usable and practical models to extract the *a priori* information.

†The reader may notice that, actually, a simple threshold on the *a priori* $W(f, \mathcal{D})$ would already give the solution to the subset selection problem for a sparse representation. We are obliged to allow this in this particular example due to the extremely low number of dimensions here involved (i.e. the example is in $\mathbb{R}^3$). The example must be considered in a *didactic* sense. Later in this chapter one can see examples on natural signals where feasible and realistic *a priories* are used, and where simple thresholding of the *a priori* matrix does not give the solution to the subset selection problem.

5.4 Rate of Convergence of Weighted-MP/OMP

5.4.1 Theoretical Rate of Convergence

To find a bound on the rate of convergence of Weighted-MP/OMP we follow the path of [95] and [142] where the respective authors look for an equivalent result for the case of *Weak*-MP in the former, and for the particular case of a block based dictionary in the latter. Simply looking at the results found in [95] it is intuitively clear that the knowledge of some *a priori* information should allow for a better bound on the rate of convergence of the representation. Indeed, in the convergence of the *Weak*-MP, the cumulative coherence function appears as a determining factor that drives the speed of exponential decay. Given the fact that $\mu_1^w(m) \leq \mu_1(m)$, we consider that having some *a priori* knowledge contributes to determine a lower bound on the exponentially decaying rate of convergence associated to Weighted-MP/OMP.

Theorem 5.4 *Let $W(f, \mathcal{D})$ be the data dependent weighting matrix that introduces* a priori *knowledge in $\mu_1^w(m)$. Let m be an integer such that:*

$$\frac{\mu_1^w(m)}{1 - (\mu_1^w(m-1) + \epsilon_{max})} < \alpha. \tag{5.13}$$

Then for any subset $\mathcal{D}_\Gamma \subset \mathcal{D}$ with $|\mathcal{D}_\Gamma| \leq m$, and any $f \in span(\mathcal{D}_\Gamma)$, Weighted-MP/OMP picks up only correct atoms at each step and

$$\|r_{k+1}\|^2 \leq \|f\|^2 \left(1 - \alpha^2 \frac{(1 - \mu_1^w(m-1) - \epsilon_{max})}{m}\right)^{n+1}. \tag{5.14}$$

The reader is referred to Appendix B.3 for the proof.

Thus, since $\mu_1^w(m-1) \leq \mu_1(m-1)$ and assuming ϵ_{max} to be small enough, a faster rate of convergence is reached.

5.4.2 A Toy Example for Weighted-MP and MP

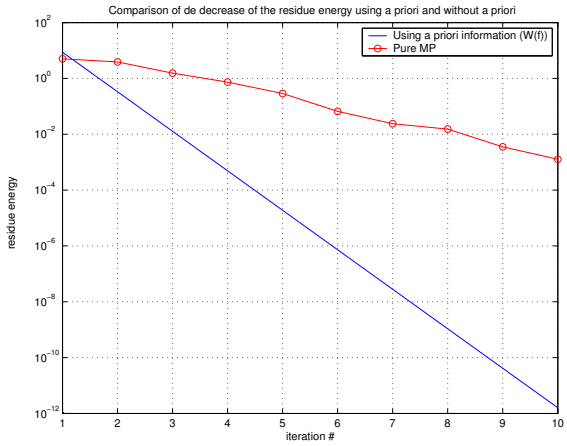


Figure 5.2: Convergence of the approximation error of the example of Fig. 5.1. The respective rates with and without using weights are compared. The use of weights enhances the asymptotic rate of convergence.

To illustrate the theoretical result found in this section, we go back to the toy example presented in sec. 5.3.2 where an overcomplete coherent Dictionary in $\mathbb{R}^3$ is used. As can be expected from Theorem 5.3 and Theorem 5.4 and observed in Fig. 5.2, the rate of convergence of Weighted-MP shows a much faster decay of the error energy than classical MP. Indeed, as guaranteed by the sufficient condition (5.10) and as illustrated in Sec. 5.3.2, the Weighted-MP algorithm gets trapped selecting over and over only vectors from the optimal set Γ. This avoids introducing spurious terms in the signal expansion and allows a faster exponential convergence than in the pure greedy case.

5.5 Examples: Heuristics in a Coherent Dictionary Based on Wavelet Footprints

In this section we provide more realistic examples than those appearing in previous sections. Some experiments on retrieving the sparsest signal representation using a redundant dictionary are shortly presented. Both weighted and classical approaches are used.

Let us explore the representation of piecewise-smooth signals and the use of dictionaries composed by the mixture of an orthonormal wavelet basis and a family of wavelet footprints (see [58]). Wavelet footprints are the functions composed by all wavelet coefficients that a given singularity generates on a orthonormal basis or frame as illustrated in Fig. 5.3.

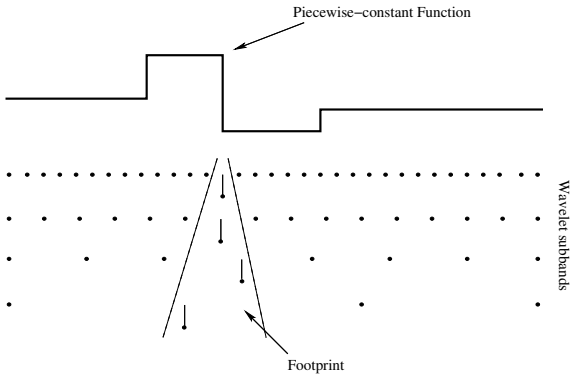


Figure 5.3: Wavelet Footprints description scheme for a piecewise-constant signal [58].

In this example, f is a 1-D signal with 128 samples, which can be sparsely represented by describing the singularities with footprints. We assume that the family of wavelets in use has a sufficiently high number of vanishing moments such that polynomial parts of the signal are efficiently represented by the coefficients of the scaling functions. Moreover the set of discontinuities appearing in the signal are also contained in the dictionary in the form of footprints. For the sake of simplicity, we consider a piecewise constant signal f (see Fig. 5.7). The dictionary is defined by the union of an orthonormal basis defined by the *Symmlet*-4 family of wavelets [121] and the respective family of footprints for all possible translations of the Heaviside function. The later is used to model the piecewise constant discontinuities. The graphical representation of the dictionary matrix can be seen in Fig. 5.4 where the columns are the waveforms that compose the dictionary.

The overcompleteness of the dictionary is evident: the number of atoms is twice the dimension of the signal. In spite of its simplicity, the dictionary presents a very high coherence factor $\mu_1(1) = 0.9606$. It is indeed very difficult for such a dense dictionary to fulfill the bounds of Theorem 4.1. For example, for m=3, $\mu_1(3) + \mu_1(2) = 4.7664$, which is already quite far from the required upper bound. In this example the optimal subset that represents the signal f has size $m = 9$.

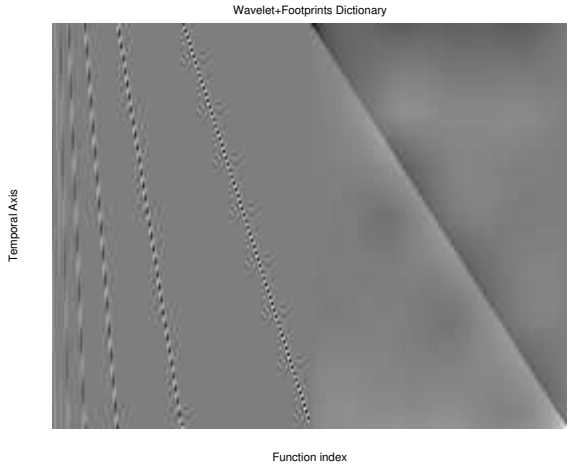


Figure 5.4: Dictionary formed by the Symmlet-4 [121] (left half) and its respective footprints for piecewise constant singularities (right half).

The signal f has been selected such that footprint components are close enough to strongly interact. If they were not overlapping, then any pure greedy algorithm would be able to recover the good representation without problems, given their orthogonality.

The weights of $W(f, D)$ are estimated from the data. This is done following a simple procedure inspired from [58]. This somehow tries to estimate the location of footprints and to penalize those wavelets that overlap with the footprints location. The detailed procedure is depicted in Algorithm 5.1.

Algorithm 5.1: $W(f, D)$ estimation

Require: $\mathcal{D} = \mathcal{D}_{Symmlet} \cup \mathcal{D}_{Footprints}$, define a threshold λ , define a penalty factor β

1: $f_{diff} = D^{+}_{Footprints} \cdot f$ {Footprints location estimation (edge detection)}
2: Threshold f_{diff} by λ putting greater values to 1 or β otherwise.
3: $W^{diag}_{footprints} = f_{diff}$ {Diagonal of the sub-matrix of $W(f, D)$ corresponding to footprints.}
4: Create W^{diag}_{wave} s.t. all wavelets intersecting the found footprints locations equal β, set to 1 otherwise.
5: $W(f, D) = diag\left(\left[W^{diag}_{wave} \quad W^{diag}_{footprints}\right]\right)$;

Algorithm 5.1 is a parametric model that establishes a dependence among signal features, dictionary structure and the interaction between functions from the same dictionary. This conform a simple configurable model that can be tuned by means of parameters λ and β.

In the results presented here, parameters λ and β are set to 0.7 and 0.6 respectively. The resulting vector of weights from the diagonal of $W(f, D)$ is shown in Fig. 5.5. Notice the four spikes in the right part of Fig. 5.5. These point out the index of the footprint functions that are more likely to be components of f. All the spikes in the left part correspond to the wavelet function indexes that interact with the location of the most probable footprints. The choice of parameters may seem a little bit Ad-hoc for this example; in a more general case (as shown later in this chapter), parameters choice should be included in the optimization of the signal decomposition algorithm, using, for example, an Expectation Maximization strategy.

The effect of applying the weights is reflected in the Gram matrix of D and $D \cdot W$ in Fig. 5.6. A reduction on the strength of interference between the dictionary atoms can be observed in the

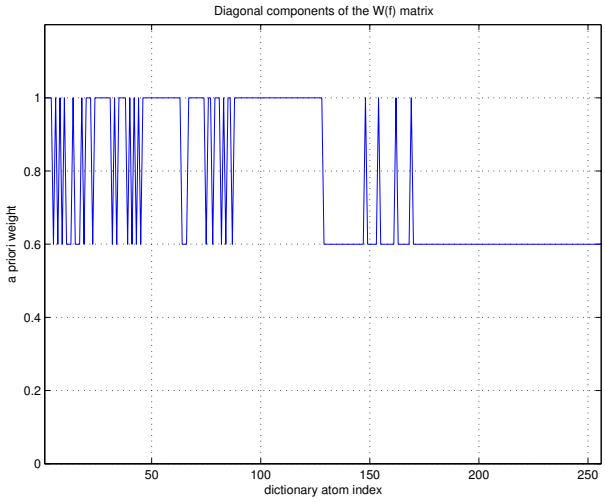


Figure 5.5: Weights involved on introducing the *a priori* information to drive OMP.

Gram matrix of the weighted dictionary.

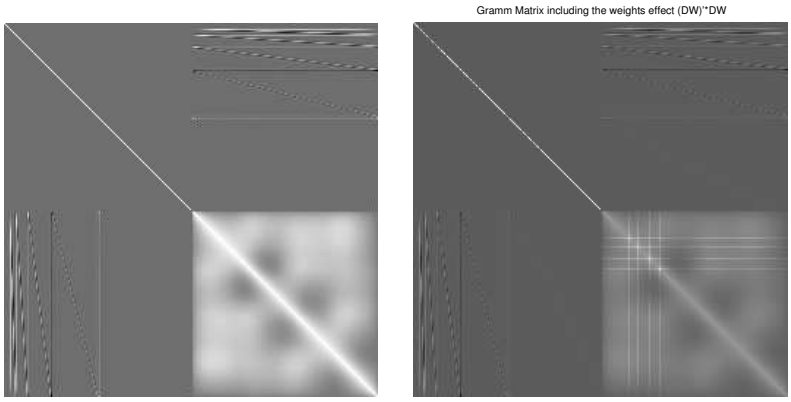


Figure 5.6: Left: Representation of the Gram matrix (i.e. $D^T \cdot D$) of the combined wavelet-footprints dictionary of Fig. 5.4. It clearly depicts the cross products between the different atoms. The upper left side perfectly describes the orthogonality of the Symmlet basis. At the bottom right a sketch of the high coherence among the footprints. Right: Representation of the Gram matrix after applying weights. Notice the reduction of cross-interferences.

Contrary to what the reader could expect now, we are not able to say that given an *a priori* information the sufficient conditions defined previously in this paper are satisfied. Indeed, the signal singularities are so close that their optimal atoms are not incoherent enough to allow the summation $\mu_1^w(m) + \mu_1^w(m-1)$ to be smaller than one. Despite that, we are be able to say that the use of Weighted-OMP (MP and Weighted-MP fail in any case) and Weighted Basis Pursuit helps recovering the optimal representation. This illustrates the intuitive idea that *a priori* information may help the signal representation even if the sufficient conditions of Theorem 5.3 are not satisfied.

The comparative results of representation by means of OMP and Weighted-OMP can be seen in Fig. 5.7. The effect of the *a priori* knowledge to recover the optimal representation is obvious (first picture on the left). The high coherence of the dictionary makes the non-weighted algorithm select

wavelet bases when it should not.

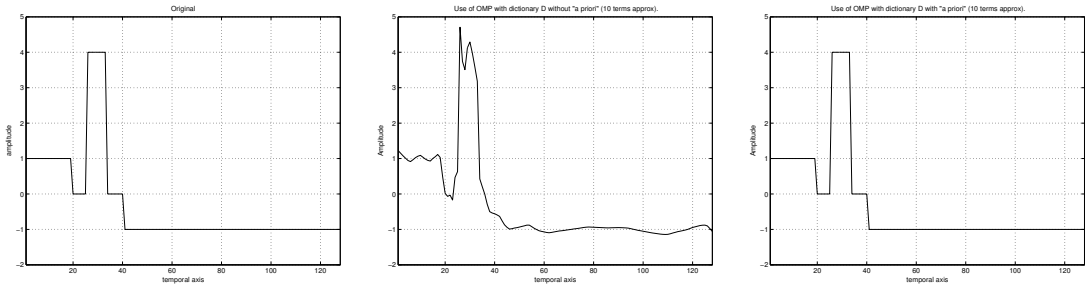


Figure 5.7: Comparison of OMP based approximation with 10 terms using the footprints dictionary (Fig. 5.4). Left: Original signal. Middle: blind OMP approximation. Exact representation is not achieved. Right: (only 9 terms are different from 0) OMP with prior knowledge of the footprints location, in this case exact representation is achieved.

A global view of the impact of using the *a priori* information is presented in Fig. 5.8. Weighting is able to keep OMP on the track for the recovery of the exact-sparse representation unlike classical OMP.

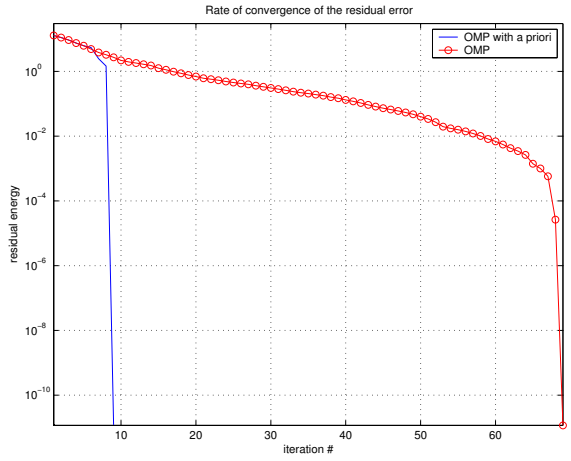


Figure 5.8: Rate of convergence of the error with respect to the iteration number in the experiment of fig. 5.7

5.6 Including *A Priori* Models in Greedy Algorithms for Sparse Approximations

Until here, only exact sparse representations have been considered. However, sparse approximations often may suggest a more relevant interest given their large range of applications: denoising, signal compression, sources separation, etc. Hence, for the rest of the chapter we will explore the effect of using *a priori* knowledge in algorithms for the recovery of the best m-term approximant (f_m^{opt}) of a signal f.

We continue to use the previous Bayesian formulation for greedy algorithms. This allows to formally introduce and analyze the usage of accessory estimators, capable to model some relation between data and dictionary, in the subset selection problem for signal approximations. In this section, sufficient conditions for the recovery of a "correct" atom from the sparsest m-term approximant are established first. After, we study how *a priori* knowledge affects the rate of convergence of greedy algorithms. Finally, an example is presented.

5.6.1 Influence on Sparse Approximations

Theorem 5.5 *Let $\{r_k\} : k \geq 0$, be the set of residuals generated by Weighted-MP/OMP in the approximation of a signal f, and let f_m^{opt} be the best m-term approximant of f over $\Gamma_m \subset \mathcal{D}$. Then, for any positive integer m such that $\mu_1^w(m-1) + \mu_1^w(m) < 1 - \epsilon_{max}$, a suboptimality factor $\eta \geq 0$ associated to the case where the algorithm can not reach the best m-term approximation and*

$$\|r_k\|_2^2 > \left\|f - f_m^{opt}\right\|_2^2 (1+\eta)^2 \left(1 + \frac{m\left(1 - (\mu_1^w(m-1) + \epsilon_{max})\right)\left(w_{\overline{\Gamma}}^{max}\right)^2}{\left(1 - (\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max})\right)^2}\right), \tag{5.15}$$

, where $w_{\overline{\Gamma}}^{max} \triangleq \sup_{\gamma \in \overline{\Gamma}} |w_\gamma|$ and $\epsilon_{max} \triangleq \sup_{\gamma \in \Gamma} \left|1 - w_\gamma^2\right|$ for $\overline{\Gamma} = \Omega \setminus \Gamma$, Weighted-MP/OMP will recover an atom that belongs to the optimal set Γ_m. Moreover, if f_m^{opt} exist and can be reached, then $\eta = 0$.

See Appendix B.4 for a proof.

This means that, if the approximation error at the nth iteration is still bigger than a certain quantity which depends on the optimal error $\|f - f_m^{opt}\|_2^2$, the weighted cumulative coherence and the reliability of the *a priori* information, then another term of the best m-term approximant can be recovered. This is similar to the result of [95], but here the use of *a priori* information results in a smaller bound. A higher number of terms from the best m-term approximant may thus be recovered.

$w_{\overline{\Gamma}}^{max}$ and ϵ_{max} concern the goodness and reliability of the *a priori* information. The reader will notice that these quantities depend on the optimal set of atoms Γ, preventing from establishing a rule to compute them in advance. The role of these magnitudes is to represent the influence of the *a priori* in the results obtained below. Notice that $0 \leq \epsilon_{max} < 1$ and $0 < w_{\overline{\Gamma}}^{max} \leq 1$. ϵ_{max} is close to zero if "good" atoms (the ones belonging to Γ) are not penalized by the *a priori*.

$w_{\overline{\Gamma}}^{max}$ becomes small if all "bad" atoms are strongly penalized by the *a priori* knowledge. If the *a priori* is reliable and $w_{\overline{\Gamma}}^{max}$ is small, then the *a priori* knowledge would be able to directly give the exact solution to the retrieval of the best Γ set. In practice, *a priori* models about the relation between dictionaries and data will be such that $w_{\overline{\Gamma}}^{max} = 1$. The important influence of *prior* knowledge within Weighted-MP is represented by $\mu_1^w(m)$. Indeed, what matters is that *prior* models help handling punctual ambiguities or undesired atom interactions within coherent dictionaries. No need thus that the *prior* model gives any accurate description of Γ.

The general effect of using *a priori* knowledge can be summarized by the following Corollary.

Corollary 5.3 *Let $W(f, \mathcal{D})$ be a reliable* a priori *knowledge obtained from a signal-dictionary model and assume $\alpha = 1$, then for any positive integer m such that $\mu_1(m-1) + \mu_1(m) \geq 1$ but $\mu_1^w(m-1) + \mu_1^w(m) < 1 - \epsilon_{max}$, Weighted-MP/OMP (unlike* Weak*(α)-MP/OMP) will recover the atoms belonging to the best m-term approximant f_m^{opt}. Moreover, for any positive integer m such that $\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max} < \mu_1(m-1) + \mu_1(m) < 1$, Weighted-MP/OMP has a weaker sufficient condition than MP/OMP for the recovery of correct atoms from the best m-term approximant.*

Hence, the correction factor of the right hand side of expression (5.15) *is smaller in the weighted case:*

$$\left(1+\frac{m\left(1-(\mu_1^w(m-1)+\epsilon_{max})\left(w_\Gamma^{max}\right)^2\right)}{\left(1-(\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max})\right)^2}\right)\leq\left(1+\frac{m\left(1-\mu_1(m-1)\right)}{\left(1-(\mu_1(m-1)+\mu_1(m))\right)^2}\right). \quad (5.16)$$

See Appendix B.5 for the proof.

Therefore, Weighted-MP is guaranteed to recover equally good or better approximants than classic MP when reliable *a priori* information is used.

5.6.2 Rate of Convergence of Weighted-MP/OMP

The energy of the series of residuals r_k $(n \geq 0)$ generated by the greedy algorithm progressively converges toward zero as n increases. In the same way, Weighted-MP/OMP with reliable *a priori* information is expected to have a better behavior and a faster convergence rate than the *Weak*-MP for the approximation case. A more accurate measure of the dictionary coherence conditioned to the signal to be analyzed is available: $\mu_1^w(m)$ (where $\mu_1^w(m) \leq \mu_1(m)$). Then a better bound for the rate of convergence can be found for the case of Weighted-MP/OMP. To prove this, we follow the path suggested in [175] for OMP and in [95] for general *Weak*-MP, introducing as before the consideration of the *a priori* information in the formulation. The results formally show how much Weighted-MP/OMP can outperform *Weak*-MP when the *a priori* knowledge is reliable.

Theorem 5.6 *Let $W(f,\mathcal{D})$ be a reliable* a priori *information matrix and $\{r_k\} : k \geq 0$ a sequence of residuals produced by Weighted-MP/OMP, then as long as $\|r_k\|_2^2$ satisfies Eq.* (5.15)*, Weighted-MP/OMP picks up a correct atom and*

$$\left(\|r_k\|_2^2-\left\|r_m^{opt}\right\|_2^2(1+\eta)^2\right)\leq\left(1-\alpha^2\frac{(1-\mu_1^w(m-1)-\epsilon_{max})}{m}\right)^{k-l}\left(\|r_l\|_2^2-\left\|r_m^{opt}\right\|_2^2(1+\eta)^2\right),$$

where $k \geq l$.

See Appendex B.6 for a detailed description of the proof.

Theorem 5.6 implies that the rate of convergence of Weighted-MP, in the same way as Weak-MP, has an upper bound with exponential decay. Moreover, in the case where reliable *a priori* information is used, the bound appears to be lower than in the case where *a priories* are not used. This result suggests that the convergence of suitably weighted greedy algorithms is faster than for the case of pure greedy algorithms. Of course, this is subject to the use of a model that puts in relation both the signal and dictionary. Some fuzzy (i.e. $w_\Gamma^{max} = 1$) and non-penalizing indication about the appropriate atoms may be of great help for the convergence of the algorithm.

Depending on the sufficient conditions specified in Sec. 5.6.1, the recovery of the optimal set Γ will be possible. However, it is not yet clear how long a non-orthogonalized greedy algorithm (Weighted-MP in our case) will last iterating over the optimal set of atoms in the approximation case. For this, in [95] the authors derive a set of bounds intended to give some clue about that for *Weak*-MP. Like the rest of bounds here analyzed, as one may guess, the use of *a priori* models has a positive influence on these. Here, we analyze how *a priories* come to affect the results appeared in [95]. Let us define the number of correct iterations as follows:

Definition 5.4 *Consider a Weighted-MP/OMP algorithm used for the approximation of signals. We define the number of provably correct steps N_m as the smallest positive integer such that*

$$\|r_{N_m}\|_2^2 \leq \left\|f - f_m^{opt}\right\|_2^2 (1+\eta)^2 \left(1 + \frac{m\left(1 - (\mu_1^w(m-1) + \epsilon_{max})\left(w_{\overline{\Gamma}}^{max}\right)^2\right)}{\left(1 - (\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max})\right)^2}\right),$$

which corresponds to the number of atoms belonging to the optimal set that can be recovered given a signal f, a dictionary $\mathcal{D}$ and an a priori *information matrix $W(f, \mathcal{D})$.*

In the case of OMP and Weighted-OMP, N_m will be always smaller or equal to the cardinality of Γ. For *Weak*-MP and Weighted-MP, provided that $\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max} < 1$, the provable number of correct iterations will depend on the final error of the best m-term approximation. In the following theorem, bounds on the quantity N_m are given for Weighted-MP/OMP. To obtain the results we follow [95].

Before stating the theorem, the reader should note that from now on, $w_{\overline{\Gamma}_l}^{max}$ defines the same concept as $w_{\overline{\Gamma}}^{max}$ for an optimal set of atoms Γ of size l, i.e. for Γ_l.

Theorem 5.7 *Let $W(f, \mathcal{D})$ be a reliable* a priori *information and $\{r_k\} : k \geq 0$ a sequence of residuals produced by Weighted-MP/OMP when approximating f. Then, for any integer m such that $\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max} < 1$, we have $N_1 \leq 1$ and for $m \geq 2$:*

- *if* $3\left\|r_1^{opt}\right\|_2^2 \geq m \cdot \left\|r_m^{opt}\right\|_2^2 (1 - \epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max}\right)^2$, *then*

$$2 \leq N_m < 2 + \frac{2 \cdot m}{1 - \epsilon_{max}} \log\left(\frac{3\left\|r_1^{opt}\right\|_2^2}{m \cdot \left\|r_m^{opt}\right\|_2^2 (1 - \epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max}\right)^2}\right). \tag{5.17}$$

- *else* $N_m \leq 1$.

We refer the reader to Appendix B.7 for a detailed description of the proof.

From (5.17) we can draw that the upper bound on the provably correct number of steps N_m is tighter for Weighted-MP if a reliable *a priori* knowledge is used. Indeed, in accordance with Theorem 5.6, which states a tighter residual error convergence bound for Weighted-MP, one can also have a tighter estimate for Weighted-MP about which is the maximum number of good iterations the algorithm might do. If some *a priori* is available, some atom interactions will not influence $\mu_1^w(m-1)$ in Eq. (5.17), unlike in the case of Theorem 7 in [95] where $\mu_1(m-1)$ was used.

Moreover, in a situation where the reliable model used to establish the *a priori* information was discriminative enough, we are sure that there would be additional room for an improvement on the number of correct iterations recovered by the greedy algorithm with respect to [95]. The term $w_{\overline{\Gamma}_m}^{max}$ helps to increase the value of the bound, describing the fact that Weighted-MP can recover a higher number of correct iterations than MP. In addition, compared to the case when no *a priori* information is available [95], the condition for the validity of bound (5.17) is softened in our case. Even though the assumption of good discrimination capabilities of the *a priori* model is somehow unrealistic in practice (i.e. a small value for $w_{\overline{\Gamma}_m}^{max}$ indicates that the model can already discriminate between Γ and $\overline{\Gamma}$), the result of Theorem 5.7, apart of giving a better estimate on the upper bound of N_m (thanks to the use of $\mu_1^w(m-1)$ instead of $\mu_1(m-1)$), it suggests also that using an *a priori* model should have a positive effect on the stability of Weighted-MP. In practice, if the *a priori* is capable to handle some punctual ambiguity that may affect the choice of the appropriate function at

a given MP step, then the benefits for the convergence of the algorithm can be of extreme relevance. This can be the case even if the *a priori* model does not supply a good discrimination between Γ and $\overline{\Gamma}$. We find very interesting practical examples of this fact in Sec. 5.6.3 and Sec. 5.8.

5.6.3 Example: Use of Footprints and Weighted-OMP for Sparse Approximations.

To give an example of approximation using *a priori* information, we consider the case where a piecewise-smooth signal is represented by means of an overcomplete dictionary.

The dictionary is the one used in Sec. 5.5. Such a dictionary does not satisfy at all the sufficient condition required to ensure the recovery of an optimal approximant with more than one term. Moreover, even if the best *a priori* was available, it is also far from satisfying the sufficient condition based on the weighted cumulative coherence. Nevertheless, we consider this example because of two main reasons. The first concerns the fact that sufficient theoretical conditions exposed in the literature are very pessimistic and reflect the worst possible case. The second reason is that, as previously discussed, experience seems to teach us that good dictionaries for efficient approximation of some classes of signals, are likely to be highly coherent. This fact conflicts with the requirement of incoherence for the good behavior of greedy algorithms. Hence, we find this example of special interest to underline the benefits of using *a priori* information and additional signal modeling for non-linear expansions.

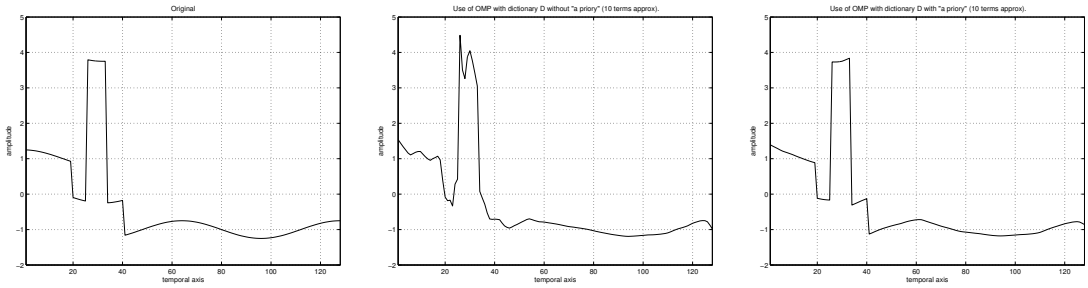


Figure 5.9: Comparison of OMP based approximation with 10 terms using the footprints dictionary (Fig. 5.4). Left: Original signal. Middle: "blind" OMP approximation. Right: OMP with prior knowledge of the footprints location.

We repeat the procedure used in Sec. 5.5 to estimate the *a priori* information based on the dictionary and the input data. We also refer the reader to Sec. 5.8 for a more detailed explanation on the model configuration for the parameter optimization. Fig. 5.9 presents the original signal (left) together with the two approximations obtained in this example: without *a priori* in the middle and with *a priori* on the right. The input signal has a number of polynomial degrees higher than the number of vanishing moments of the *Symmlet*-4. The figures depict clearly the positive effect of the reliable *a priori* information inserted in the Weighted-OMP algorithm. Indeed, with very few components, the algorithm benefits from the *a priori* information estimated from the signal, and gives a much better approximation. A more global view of this behavioral enhancement can be seen in Fig. 5.10 where the rate of convergence of the approximation error is presented. The use of weights is definitively helpful and a considerable reduction of the approximation error is achieved for a small number of terms.

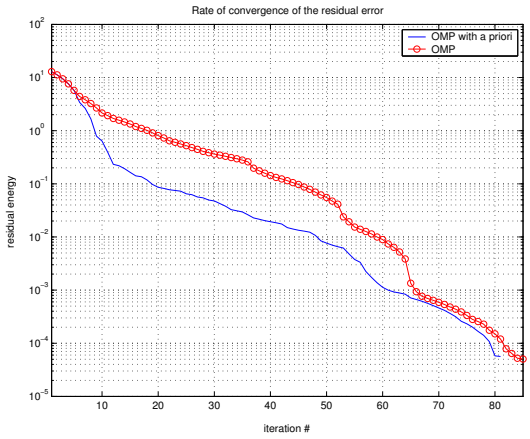


Figure 5.10: Rate of convergence of the error with respect to the iteration number in the experiment of Fig. 5.9

5.7 Approximations with Weighted Basis Pursuit Denoising

Another sub-optimal method to solve the problem in Eq. (4.4) is given by relaxation algorithms, whose recovery capabilities in presence of *a priori* knowledge are shortly reviewed in this section.

A study in [48] investigates the effects of inserting *a priori* knowledge in the convex relaxation of the subset selection problem (see Sec. 4.6), i.e. in the approximation case.

5.7.1 A Bayesian Approach to Weighted Basis Pursuit Denoising

In this subsection the problem of signal approximation is studied from a Bayesian point of view. This leads us to generalize the BPDN principle through the definition of Weighted Basis Pursuit Denoising (WBPDN). First, let us write the model of our data approximation, where $\hat{f}$ is the approximant and r is the residual:

$$f = \hat{f} + r = D\mathbf{b} + r. \tag{5.18}$$

Assuming r to be an iid Gaussian set of variables, the probability that f corresponds to $\hat{f}$, given D and $\mathbf{b}$ is:

$$p(f|D,\mathbf{b}) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \cdot \exp\left(-\frac{\|f - D\mathbf{b}\|_2^2}{2\sigma_r^2}\right),$$

where σ_r^2 is the variance of the residual. In the approximation problem, one aims at maximizing the likelihood $p(\mathbf{b}|f,D)$. Formally, by the Bayes rule, we have

$$p(\mathbf{b}|f,D) = \frac{p(f|D,\mathbf{b}) \cdot p(\mathbf{b})}{p(f,D)},$$

and thus, assuming $p(f,D)$ uniform, it follows that the most probable signal representation is:

$$\mathbf{b}_P = \arg\max_{\mathbf{b}} p(f|D,\mathbf{b}) \cdot p(\mathbf{b}). \tag{5.19}$$

Let us now assume that the coefficients b_i are independent and have a Laplacian distribution with standard deviation σ_i:

$$p(b_i) = \frac{1}{\sqrt{2}\sigma_i} \cdot \exp\left(-\frac{\sqrt{2}|b_i|}{\sigma_i}\right).$$

From (5.19), by computing the logarithm, it follows that

$$\mathbf{b}_P = \arg\max_{\mathbf{b}} \left(\ln(p(f|D,\mathbf{b})) + \sum_i \ln p(b_i) \right) = \arg\min_{\mathbf{b}} \left(\frac{\|f - D\mathbf{b}\|_2^2}{2\sigma_r^2} + \sum_i \frac{\sqrt{2}|b_i|}{\sigma_i} \right).$$

Making the hypothesis that σ_i is constant for every index i, the previous equation means that the most probable $\mathbf{b}$ is the one found by the BPDN algorithm [119]. In fact, this hypothesis does not often correspond to reality. On the contrary, if the variances of the coefficients are not forced to be all the same, it turns out that the most probable signal representation can be found by solving the following problem:

$$(P_1^w) \qquad \min_{\mathbf{b}} \frac{1}{2} \|f - D\mathbf{b}\|_2^2 + \gamma \|W^{-1}\mathbf{b}\|_1, \tag{5.20}$$

where the diagonal matrix with entries in $(0, 1]$ is defined in Section 5.6. One can notice that in Eq. (5.20), the introduction of weights allows to individually model the components of $\mathbf{b}$. This approach is analogous to the one introduced in [49, 91] as well as to the use of *a priories* into *Tikhonov regularization* based methods [89]. From now on, we will refer to P_1^w as Weighted Basis Pursuit Denoising or WBPDN.

The assumption often made about the Gaussianity of the residual is quite restrictive. However, for another particular problem, one could make the hypothesis that this residual has a Laplacian distribution. It is then possible to prove that the most probable signal representation can be found substituting the L^2 measure of the error with the L^1. This leads to the following minimization problem:

$$\min_{\mathbf{b}} \frac{1}{2} \|f - D\mathbf{b}\|_1 + \gamma \|W^{-1}\mathbf{b}\|_1,$$

where $W = I$ if the variances of the probability density functions of b_i are the same for each i. This problem is faced, for example, in [91], where it is solved by Linear Programming techniques.

5.7.2 Relation with the Weighted Cumulative Coherence

As in the case of greedy algorithms, the behavioral bounds obtained by Tropp in [176] for BPDN, may be extended when additional *a priori* models are plugged into the optimization P_1^w problem. This analysis is performed in [48]. Here, the main results are summarized.

Lemma 5.1 *([48]) Given an index subset $\Lambda \subset \Omega$, suppose that the following condition is satisfied:*

$$\|D^T(f - a_\Lambda)\|_\infty < \frac{\gamma}{w_\Gamma^{max}} \cdot (1 - \sup_{i \notin \Lambda} \left\| (D_\Lambda W_\Lambda)^+ g_i \cdot w_i \right\|_1), \tag{5.21}$$

where a_Λ is the optimal approximation of f on Λ. Then, any coefficient vector $\mathbf{b}_*$ that minimizes the cost function of problem P_1^w must have a support contained in Λ.

This proposition, akin to the Correlation Condition Lemma in [176], basically states that, if the atoms of Λ have a small weighted coherence, expressed by the Weighted Recovery Factor, then the support of any vector that solves P_1^w is a subset of Λ. The factors derived from the use of *a priori* information, can rise the value of the right term of (5.21), if they are "reliable" enough, reducing the tightness of the condition. For a given γ, a model that relates signal and dictionary $\mathcal{D}$ will make more robust and performant WBPDN in front of BPDN.

The result of Lemma 5.1 can be further extended in order to establish relations among the algorithmic thresholds γ and τ (see Sec. 4.4.4 to see its meaning), the maximum error that the relaxation algorithm will introduce in the retrieved coefficients and the internal coherence of the

dictionary. In [176], this is performed for the generic case where no *a priori* is used. The result determined that the algorithm required incoherent dictionaries to well operate. The influence of the inclusion of additional models to the optimization process are analyzed in [48] and can be seen in the following Theorem. Two main effects are perceived: First, the requirement of incoherence is softened. Second, the maximal possible error in the retrieved coefficients is reduced.

Theorem 5.8 *([48]) Assume that the real vector* $\mathbf{b}_*$ *solves* P_1^w *with*

$$\gamma = \frac{w_\Gamma^{max} \cdot \tau(1 - \epsilon_{max} - \mu_1^w(m-1))}{1 - \epsilon_{max} - \mu_1^w(m) - \mu_1^w(m-1)}.$$

Then support$(\mathbf{b}_*) \subset \Gamma$ *and*

$$\|\mathbf{b}_* - \mathbf{c}_\Gamma\|_\infty \leq \frac{\tau \cdot \frac{w_\Gamma^{max}}{w_\Gamma^{min}}(1 - \epsilon_{max} - \mu_1^w(m-1))}{(1 - \epsilon_{max} - \mu_1^w(m) - \mu_1^w(m-1))(1 - \mu_1(m-1))}. \tag{5.22}$$

This result is valid in general and illustrates how the distance between the optimal coefficients and the solution found by solving P_1^w can be bounded. In case no *a priori* is given, the bound on the coefficient error is obtained from Eq. (5.22) setting $W = I$. Consequently, $w_\Gamma^{min} = 1, \epsilon_{max} = 0$ and $w_\Gamma^{max} = 1$ (see also [176]):

$$\|\mathbf{b}_* - \mathbf{c}_\Gamma\|_\infty \leq \frac{\tau}{1 - \mu_1(m) - \mu_1(m-1)}, \tag{5.23}$$

where $\mathbf{c}_\Gamma$ is the optimal set of coefficients with support in Γ. Comparing the two bounds, one can observe how the availability of a reliable *a priori* on the signal can help in finding a sparser signal approximation. Let $W(f, \mathcal{D})$ be a reliable a priori knowledge, with $w_\Gamma^{max}/w_\Gamma^{min} \leq 1$. Then for any positive integer m such that $\mu_1^w(m-1) + \mu_1^w(m) + \epsilon_{max} < \mu_1(m-1) + \mu_1(m) < 1$, the error $\|\mathbf{b}_* - \mathbf{c}_\Gamma\|_\infty$ given by the coefficients found by WBPDN is smaller than the one obtained by BPDN.

Hence, the bound stated by Eq. (5.22) is lower than the one in Eq. (5.23), i.e.

$$\frac{\tau \cdot \frac{w_\Gamma^{max}}{w_\Gamma^{min}}(1 - \epsilon_{max} - \mu_1^w(m-1))}{(1 - \epsilon_{max} - \mu_1^w(m) - \mu_1^w(m-1))(1 - \mu_1(m-1))} \leq \frac{\tau}{1 - \mu_1(m) - \mu_1(m-1)}. \tag{5.24}$$

One can prove this result following the procedure of the proof of Corollary 5.3 reported in the Appendix. See [48] for more details.

The reader may notice that if $\frac{w_\Gamma^{max}}{w_\Gamma^{min}} < 1$ the *a priori* information already tells which is the right support of the solution. Indeed, a simple threshold on the weights would find the appropriate set of atoms. This is an unrealistic situation in practice. However, provided that the *a priori* information is reliable, we do not need $\frac{w_\Gamma^{max}}{w_\Gamma^{min}} < 1$ to justify an improvement on the behavior of the algorithm. Suppose that the weights do not penalize the optimal atoms, but only some (not all) of the "wrong" ones: in this case $\frac{w_\Gamma^{max}}{w_\Gamma^{min}} = 1$. In such a situation, given that $\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max} < \mu_1(m-1) + \mu_1(m) < 1$, Eq. (5.24) is still valid. This means that, even if the *a priori* knowledge is imprecise (but reliable), WBPDN can behave significantly better than BPDN.

Note that, once the algorithm has recovered the atom subset, the appropriate amplitudes of the coefficients can be computed by the orthogonal projection of the signal onto the space generated by the selected atoms. Hence, the error introduced by the relaxation algorithm into the coefficients is irrelevant in practice if the good set of atoms is retrieved. This re-projection step is illustrated in Section 5.8.2.

5.8 Examples: Natural Signal Approximation with an *A Priori* Model

In this section we apply the methodology introduced in Sec. 5.6 and Sec. 5.7 to natural signals. We also discuss the problem of extracting reliable *a priori* information on a concrete example. Moreover, we will show how the *a priori* weights can be automatically extracted from the data and optimized in order to maximize the performance of weighted algorithms. We approximate several 1D signals extracted from a variety of columns from "Cameraman" and "Lenna" images, and that can be considered as piecewise-smooth, by using an overcomplete coherent dictionary.

5.8.1 Modeling the Relation Signal-Dictionary

The dictionary, composed by the union of the *Symmlet*-4 orthonormal basis is used to model smooth parts of the signal, and the set of piecewise-constant footprints meant to model discontinuities (see Sec. 5.5 and Fig. 5.4). Since the input signal has 256 samples, D is a matrix of size 256×512. The modeling of the interaction between the signal and the dictionary is performed using the simple approach described in Sec. 5.5. The weighting matrix $W(f, D)$ is generated by means of a pre-estimation of the locations where footprints are likely to be used, assuming that in such locations wavelets have less probability to appear. This discrimination does not penalize locations where a footprint is likely to be placed (thus the weighting factor remains 1). On the contrary, wavelets that overlap the footprint, as well as footprints considered unlikely to be used, get a penalizing factor $\beta \in (0, 1]$. Hence, extracting relevant footprints and wavelets by selecting these corresponding to strong weights would not necessarily yield a good sparse approximation

As one can observe, two parameters configure the model that generates $W(f, D)$: a threshold λ and a penalty weight β. We will shown later that these can be automatically selected by an iterative optimization procedure that minimizes the energy of the approximation error.

5.8.2 Signal Approximation

We resume the general procedure for the signal approximation by these two steps:

1. Estimation of the *a priori* information from the "real world" signal using an *a priori* model.
2. Use of a weighted algorithm (greedy or relaxed) based on the estimated *a priori* knowledge to find the appropriate atoms subset

   Optionally, once these have been selected, their optimal coefficients can be computed again, by means of a simple projection. This is, for the case of BPDN and WBPDN: Let us call $\mathbf{b}_*$ the approximation found by BPDN and $\mathbf{b}_*^w$ the one found by WBPDN. These vectors are thresholded removing the numerically negligible components, and in this way we are able to individuate a sparse support and thus a subset of the dictionary. Let us label the subdictionary found by WBPDN with $\mathcal{D}_*^w$ (composed by the atoms corresponding to the non-zero elements of $\mathbf{b}_*^w$). Once this is given, there are no guarantees that the coefficients that represent f are optimal (see [176] and [48]). These are, thus, recomputed projecting the signal onto $\mathcal{D}_*^w$ and a new approximation of f named $\mathbf{b}_{**}^w$ is found. Exactly the same is done for BPDN, ending up with a subdictionary $\mathcal{D}_*$ and a new approximation $\mathbf{b}_{**}$. Of course, $support(\mathbf{b}_*) = support(\mathbf{b}_{**})$ and $support(\mathbf{b}_*^w) = support(\mathbf{b}_{**}^w)$. Formally the approximants found by BPDN and WBPDN after the projection step are respectively:

$$\begin{aligned} f_{**} &= D_* D_*^+ f = D\mathbf{b}_{**} \text{ and} \\ f_{**}^w &= D_*^w (D_*^w)^+ f = D\mathbf{b}_{**}^w. \end{aligned} \tag{5.25}$$

Furthermore, an iterative version of this two phase algorithm can be considered in order to optimize the parameters that configure the *a priori* model used in the first step: the Expectation Maximization (EM) algorithm. A first approach for the parameters tuning can be a grid search, or a multi-scale grid search. More sophisticated search techniques could also be used. An overview these can be found in [18, 145].

5.8.3 Results

The results obtained from the framework introduced above are illustrated in the following. First, we show the quantitative impact of using Weighted-MP/OMP and WBPDN in terms of the residual error energy. Right after, the use of atoms of the dictionary to represent the main features of the signal is analyzed. Finally, we explore the influence of tuning the two parameters that configure our penalty model.

Approximation Results with OMP

The improvement of Weighted-OMP in the case of sparse approximations is assessed by the rate of convergence of the residual energy, on the right-hand side of Fig. 5.11 and Fig 5.12: the graphs show that after a certain number of iterations, Weighted-OMP selects better atoms than classic *Weak*-OMP. Hence the convergence of the error improves and this yields a gain of up to 2 dBs and 2.5 dBs respectively.

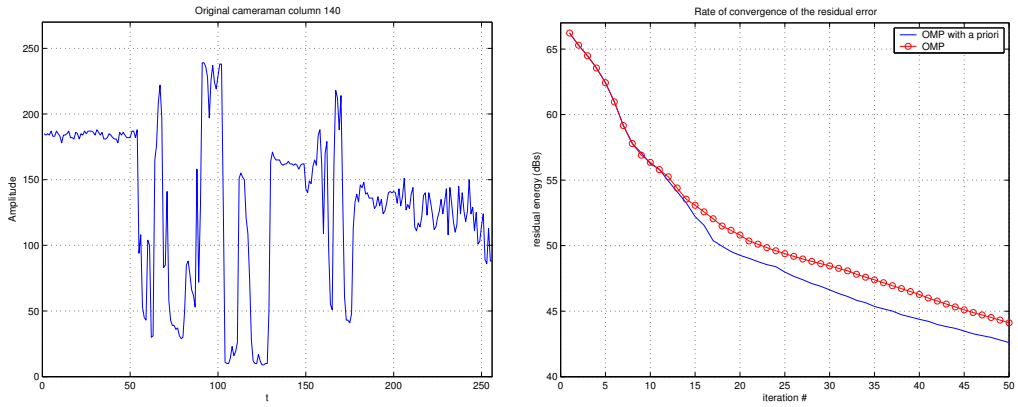


Figure 5.11: Experiment of approximating the 1D signal extracted from the 140th column of "cameraman". Left, 1D signal used in the experience can be seen. Right, the rate of convergence of the residual error. In red can be observed the OMP result. In blue the Weighted-OMP result.

Approximation Results with BPDN

The same signal can be approximated by BPDN and WBPDN. As explained previously, the pursuit algorithm is used only to select a dictionary subset and then the coefficients of the approximation are computed again, by means of a simple projection. Fig. 5.13 shows the decay of the error versus the number of atoms. It is clear how the use of the *a priori* helps the algorithm in finding a better approximation of the signal. The results concerning WBPDN are obtained by adopting a weighting matrix that corresponds to $\lambda = 90$ and $\beta = 0.2$. Notice that these values are not optimal for all the numbers of non-zero coefficients, as can be seen in the area between 34 and 43 selected coefficients

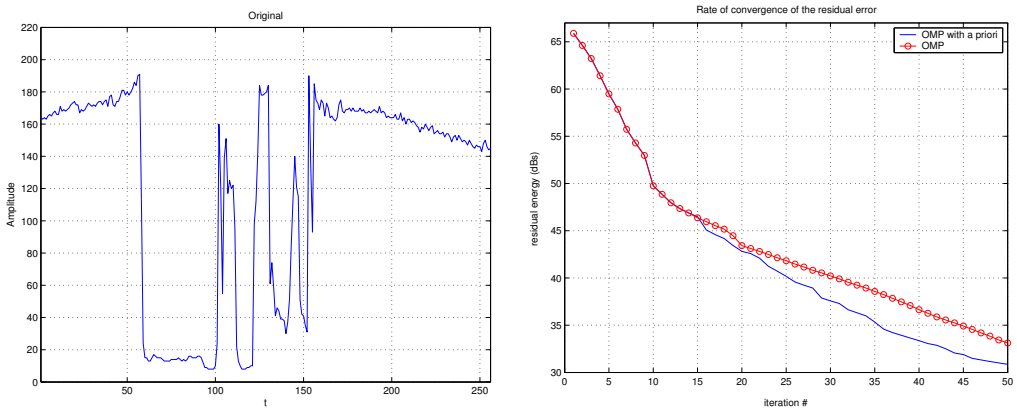


Figure 5.12: Experiment of approximating the 1D signal extracted from the 80th row of the 256x256 "cameraman". Left, 1D signal used in the experience can be seen. Right, the rate of convergence of the residual error. In red can be observed the OMP result. In blue the Weighted-OMP result.

in the graph of Fig. 5.13. Better results can be achieved by tuning appropriately β and λ for any desired m.

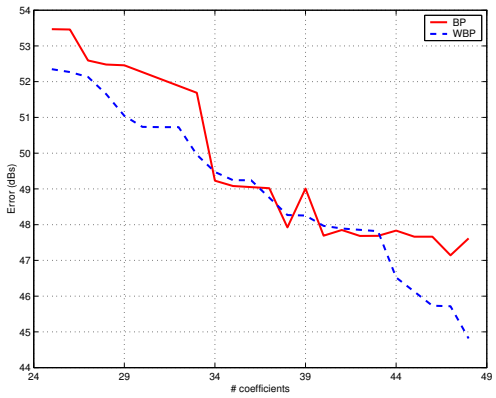


Figure 5.13: Error (in dB) obtained by BPDN and WBPDN [48]. Both results are obtained by using quadratic programming for selecting a dictionary subset and then recomputing the coefficients by re-projecting the signal onto the span of the subdictionary.

Capturing the Piecewise-smooth Component with Footprints Basis

Here, the results intend to underline the importance of selecting the appropriate atom to represent a particular signal feature. In the top row of Fig. 5.14 we can see the resulting approximants after 50 iterations of OMP (left) and Weighted-OMP (right) for the signal corresponding to the 140th column of cameraman. The result obtained by including the *a priori* is 1.51 dBs better than the one obtained by OMP. At this point, it is important to observe the bottom row of Fig. 5.14. These waveforms represent the signal components that are captured exclusively by the footprints atoms and *Symmlet*-4 scaling functions. These signal components should correspond to the piecewise-smooth parts of the signal. However, in the case of OMP (bottom left) the piecewise-smooth component captured by footprints and low-pass functions is far from what one could expect. Intuitively one can understand

that the OMP algorithm is failing in the selection of atoms. On the other hand, the result obtained by Weighted-OMP (bottom right) clearly shows that footprints and *Symmlet*-4 scaling functions are capturing a much more accurate approximant of the piecewise-smooth component of the signal. We can thus argue that a better approximation is achieved by using the *a priori* information, and this leads to a sparser approximation too.

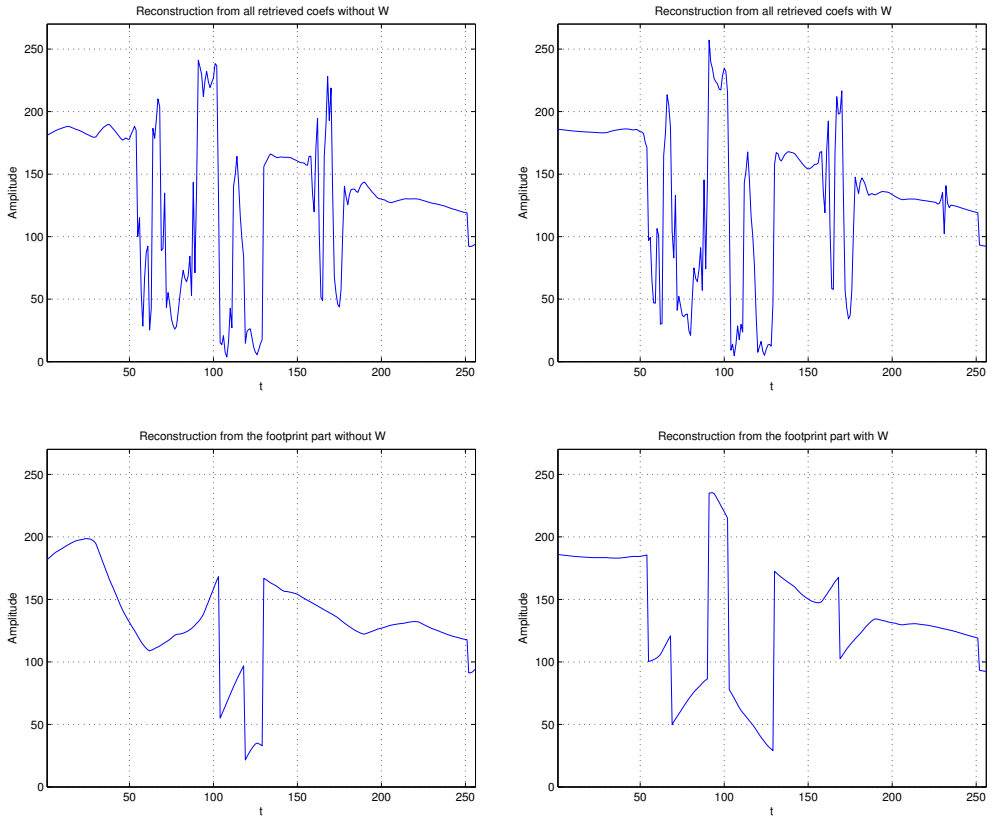


Figure 5.14: Top: Approximation after 50 iterations of OMP with (right) and without (left) *a priori* information. Bottom left: Signal components captured by *Symmlet* scaling functions and Footprints using OMP. Bottom right: Signal components captured by *Symmlet* scaling functions and Footprints using Weighted-OMP.

Parameter Search

Finally, we show the influence of the parameters λ and β in the average quadratic error of the residues obtained by Weighted-OMP, i.e.

$$E\left\{r_k|\lambda',\beta'\right\} = \frac{\sum_{k=0}^{K-1} \|r_k\|^2}{K} \tag{5.26}$$

such that r_k has been obtained fixing $\lambda = \lambda'$ and $\beta = \beta'$.

In Fig. 5.15 and Fig. 5.16, the magnitude of Eq. (5.26) is shown as a function of λ (model threshold) and β (penalty weight). The lower the value of $E\left\{r_k|\lambda',\beta'\right\}$, the higher the probability of the associated model parameters to be the good ones. Hence, it can be easily observed that the

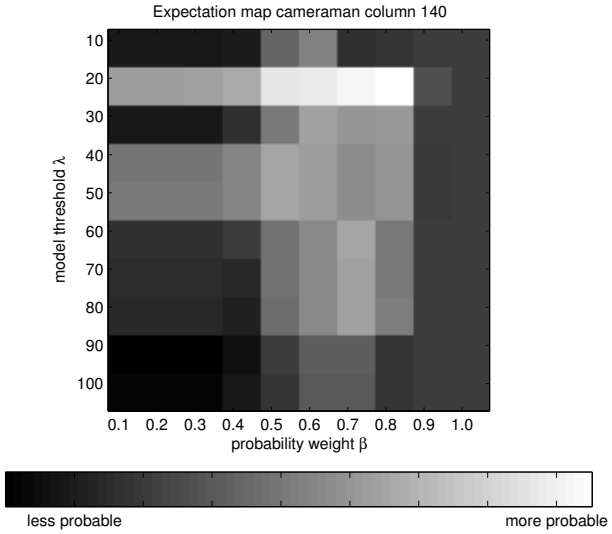


Figure 5.15: Representation of the expectation map depending on the parameters that configure the *a priori* model in the experiment set up in Fig. 5.11. The expectation corresponds to the energy of the residual error.

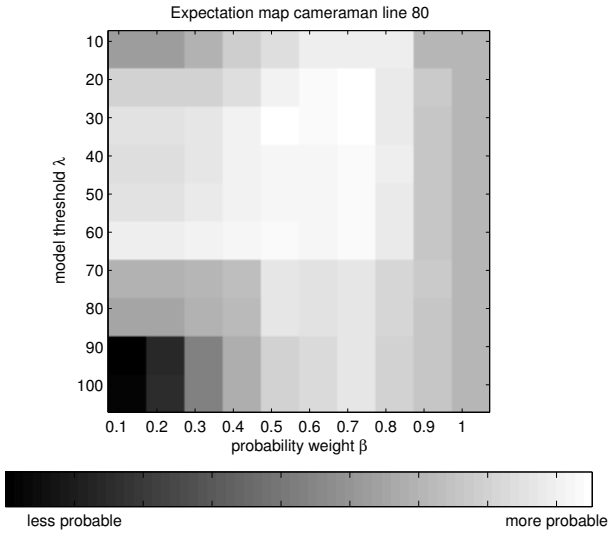


Figure 5.16: Representation of the expectation map depending on the parameters that configure the *a priori* model in the experiment set up in Fig. 5.12. The expectation corresponds tot the energy of the residual error.

optimal configuration of parameters concentrates for each of the figures in a unique global optimum. In this case, the set of optimal parameters that fit the data model can be easily found by some iterative procedure.

Behavior of Weighted-OMP on a Large Set of Piecewise-Smooth Signals

Finally, a larger subset of columns of Cameraman has been selected in order to give a better overview on how our algorithm behaves in average. For this particular experiment, a column out of three has been taken from the picture. Then, the average residual error of all signals, as a function of the greedy iteration is compared for Weighted-OMP and OMP in the left graph of Fig. 5.17. For a fair comparison, an identical analysis is performed on a set of columns from image Lenna. For this particular case, a column out of six has been considered. We give also, in the right graph of Fig. 5.17, a representation of the approximation gains supplied by the weighted algorithm for each of the selected cameraman columns. This shows that depending on the particular structure of the signal, Weighted-MP may supply very significant improvements (up to 4.5 dB better in approximation error). It also reflects the fact that, if the *prior* model is properly defined, one can not get worse results than those of the pure greedy approach, to the contrary results use to be significantly better. Fig. 5.18 depicts the approximation gains of Weighted-MP for each one of the

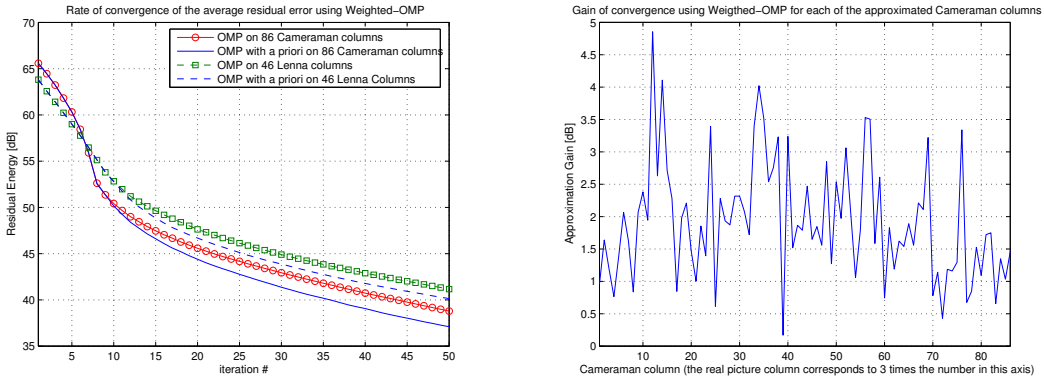


Figure 5.17: left: Average residual error convergence for Weighted-OMP and OMP for 86 columns sampled from image Cameraman and 46 columns sampled from image Lenna. Right: Approximation gain when using Weighted-OMP depending on the column sampled from image Cameraman.

approximated image columns.

As seen in this section, the use of models based on edge estimators can be of great help for good piecewise-smooth signal approximations, using coherent dictionaries and highly non-linear algorithms. One can consider this principle to approximate other signals than those presented in here. For example, in the case of images, the use of edginess measurements may help to better place edge-adapted basis functions and smooth-adapted basis functions, reducing the inconveniences of highly coherent dictionaries. A brief study based on the use of an edginess measure for images and a linear programing decomposition algorithm may be found in [91].

5.9 Conclusions

Sparse representation and approximation require the use of dictionaries capable to catch efficiently the main features and salient structures of signals. Particular applications often focus on a certain

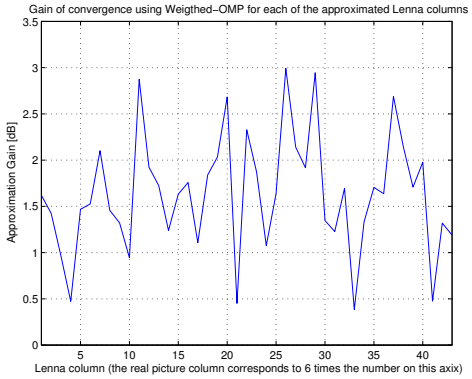


Figure 5.18: Approximation gain when using Weighted-OMP depending on the column sampled from image Cameraman.

class of signals. A wise strategy is, thus, to use dictionaries adapted to this class. Such dictionaries often have high internal coherence, while practical algorithms for the retrieval of sparse approximations, like Weak-MP, BP or BPDN have been proved to work well with incoherent dictionaries. In order to overcome this contradiction, adaptive subset selection algorithms are of key importance to obtain optimal m-term signal approximations.

Weighted variants of Weak-MP, BP and BPDN algorithms, called Weighted-MP/OMP WBP and WBPDN, are introduced. Theoretical results show that these algorithms may supply much better results than classic approaches for highly non-linear signal approximations with coherent dictionaries. In order to guarantee this, sufficiently reliable *a priori* models must be used. Our practical examples show how appropriate *a priories* may be able to characterize the interaction between signal and dictionaries.

A possible direction to explore is to determine some bound on the quantity ϵ_{max} (i.e. the reliability factor) depending on the class of signals to approximate, the selected dictionary and the practical estimators in use (those that generate the *a priori* weights). The knowledge of some bound on ϵ_{max} for certain model, given a class of signals, and a dictionary may be of great help in determining in advance whether a model can be suitable for a particular application. In general, very good feature estimators exist and a lot of experience about them is available in literature. For every particular application, models exploiting particular signal features may be found in order to marry them with most common algorithms used for sparse approximations/representations. In fact, the examples presented in this work may be subject to improvement if more robust estimators were used.

As seen in this chapter, the convergence of highly non-linear algorithms can be modified by the introduction of *a priori* models. This approach will be used in Chapter 7 to help the greedy algorithm in use to extract meaningful 3D spatio-temporal geometric video components.

Chapter 6

Matching Pursuit Geometric Image Approximations

6.1 Motivation

This chapter is the first step toward the study of a practical geometric video representation based on Matching Pursuit expansions. First of all, the problem of representing 2D data has to be tackled. Hence, the present chapter will be fully devoted to this. The work here presented builds upon the seminal study by Vandergheynst and Frossard [178], and subsequent works [64, 65, 78], on the approximation of images using a geometric overcomplete dictionary. In these, an over-complete dictionary of anisotropically refined atoms is used in order to exploit geometry in image representations. Matching pursuits are used to generate signal expansions in a feasible way. However, given the extremely large size of the used dictionary, authors used a suboptimal weak version of MP (see Chapter 4) based on genetic algorithms [43]. As seen later in this chapter, the use of such a suboptimal algorithm significantly reduces approximations efficiency.

Here, a feasible approach for Full Search Matching Pursuit (FSMP) is proposed in the particular case of natural image approximations with anisotropically refined oriented atoms. Thanks to the structure of the dictionary and its spatio-temporal localization, several enhancements allow to speed-up the calculation of the most critical step: the scalar product of the signal with all the functions of the dictionary.

This chapter is structured in the following way: First, the anisotropic refinement dictionary and the methodology used for its construction is revised in Sec. 6.2. Next, we recall in Sec. 6.3 the sub-optimal MP approach based on the genetic algorithm. Once the exposed problem, the proposed feasible Full Search MP approach is described in Sec. 6.4. In this, we expose the benefits of using the FFT to accelerate the scalar products computation and present comparative results with genetic algorithm based MP. In Sec. 6.5, further ways of exploiting the characteristics and structure of the dictionary to reduce computational complexity are presented and discussed. Finally, conclusions are drawn in Sec. 6.6.

6.2 Finding Image Components: Dictionary Design

6.2.1 Image Decomposition

In order to efficiently represent a signal, one needs to know its features and to has make some assumptions. The main assumption often made on images is that they can be represented (or at least approximated) as a finite sum of basis functions:

$$\hat{f} = \sum_{k=0}^{K-1} c_k g_{\gamma_k}, \tag{6.1}$$

where $\hat{f}$ is the image approximation, c_n are the coefficients and g_{γ_n} the selected basis functions.

Clearly, when the number of non-zero terms is much smaller than the discrete signal dimensionality, we are in the case of sparse representations and approximations. Chapter 4 shows that this kind of decompositions are straightforward to obtain when dealing with orthonormal basis or when using a frame method [121]. However, frame methods (see Sec. 4.4.1) do not provide sparse enough signal approximations. In addition, computing the dual basis is not always straightforward. Thus redundant dictionaries require an algorithm that gives a sparse decomposition: an interesting solution may be brought by greedy algorithms.

The sparse image model only implies that the image can be represented or approximated by a finite sum of basis functions. It makes no assumption on the nature of the basis functions. In order to have efficient representations, as discussed in the introductory Chapter, basis functions have to be adapted to the specific features of signals. In the case of images, this corresponds to adapt to contours, smooth regions and textures. Smooth areas are often better represented with non-zero mean functions. These functions have a very important low frequency component. If polynomial approximations are desired, scaling functions of wavelet basis can be envisaged [180]. The definition of texture is somehow trickier. Several authors have worked on the subject, defining texture in several ways, going from just considering it as dense bundles of edgy features to seeing it as conglomerates of multi-scale oscillating patterns [129, 147, 179]. Following the discussion of the introductory chapter, contours are often the most relevant and meaningful feature in natural images. Hence, they merit the use of appropriate and adapted functions to efficiently describe them. Contours are assumed to be 1D continuous smooth functions [51, 53, 62, 63, 115]. Taking this model and trying to approximate edges with the smallest number of piecewise linear segments, one understands the need of a dictionary of functions with high geometric meaning.

Some geometric, edge adapted dictionaries, may also handle with some success texture. Edge dictionaries present, in some cases, an oscillatory component. However, the same can not be said for smooth regions and low frequency components. Hence, before expanding the signal onto the geometric dictionary, the low-pass frequency part is subtracted from the signal and represented by a downscaled version (see Fig. 6.1). Low-pass representations are intended to capture the main components of smooth parts. As depicted in Fig. 6.1, local average in images is represented in our case by a limited number of coefficients that depend on the downsampling rate (this is given by the number of dyadic downsampling stages in the diagram). For simplicity, a cubic spline low pas filter has been taken here to generate the multi-scale pyramid of low-pass image versions.

6.2.2 Geometric Dictionary Generation

Images modeled by (6.1) are puzzles with $c_n g_{\gamma_n}$ terms as pieces. As we are interested in an efficient representation of edges, we need signal pieces $c_n g_{\gamma_n}$ with an elongated form. These pieces need,

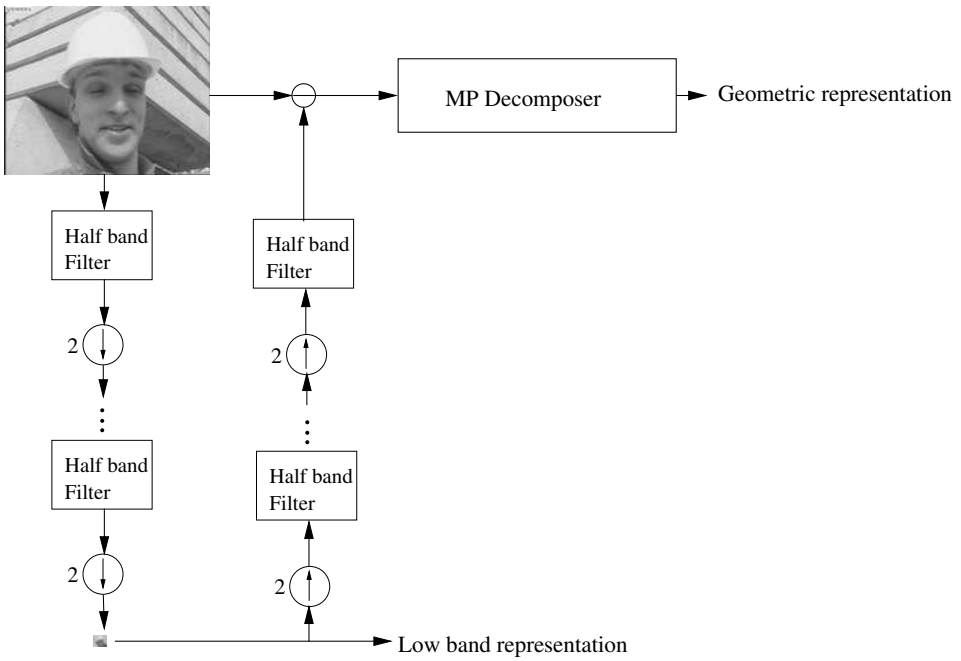


Figure 6.1: Decomposition diagram to obtain Eq. (6.1) representation. The summation terms are divided in two main kinds: low frequency components (some few coefficients, the number depends on the downsampling rate), and the remaining signal components represented by means of the geometric dictionary.

moreover, to be adapted to the great variety of geometric properties that elongated structures may have in an image.

In this section, we first review the generic formalism for the group based design of dictionaries. Then, the geometric dictionary used in the remaining of this work is described.

Group Theoretical Design of Dictionaries

The main goal of group based design of dictionaries is to create complete dictionaries by applying basic operations on a simple signal template g. The interest arises from the set of advantages that such a formulation provides. The more relevant are the following:

- It is a systematic way to generate a large family of functions with an extensive variety of geometric properties,
- if g is a continuous analytic function, then (6.1) implicitly generates a continuous model of f (we assume that f is discrete) which may of use for applications where one desires to perform interpolating operations,
- group operations used to generate the dictionary can, then, be used to easily apply group transformations to the data,
- as shown in [66, 78], several of these properties can be exploited for scalability and transcoding operations in image and video coding applications.

In the following, the unitary operator of a group is defined. This is then used in the result that states the completeness of a *group designed* dictionary.

Definition 6.1 *([16, 78]) Let G be a group and $\mathcal{H}$ be a Hilbert space. A unitary representation $\mathcal{U}$ of G in $\mathcal{H}$ is a homorphism between G and the set of unitary operators on $\mathcal{H}$ with composition as the group law:*

$$\begin{aligned} \mathcal{U} : G &\mapsto U(\mathcal{H}) \\ \gamma &\mapsto \mathcal{U}(\gamma), \end{aligned}$$

The reader is referred to [16] for details and properties of groups and group laws.

Based on that, the following result states the necessary property for a dictionary $\mathcal{D}$, generated by the group based method, to expand any signal $f \in \mathcal{H}$.

Lemma 6.1 *([16, 78]) Let G be a locally compact group and $\mathcal{U}$ a unitary representation of G in a Hilbert space $\mathcal{H}$. Let $\mathcal{U}(\gamma)g$ denote the action of the unitary operator $\mathcal{U}(\gamma)$ on the function g. The unitary representation is said to be* irreducible *if it does not admit any non trivial invariant subspaces. That is if $\mathcal{S} \subset \mathcal{H}$ is such that*

$$\mathcal{U}(\gamma)g \in \mathcal{S}, \quad \forall g \in \mathcal{S} \text{ and } \forall \gamma \in G,$$

then either we have $\mathcal{S} = \mathcal{H}$ or $\mathcal{S} = \{0\}$.

This can also be stated such that, for any $g \in \mathcal{H}$, all transformations of g under G build a set of functions which is generator of $\mathcal{H}$:

$$\mathcal{D} = \{\mathcal{U}(\gamma)g | \forall \gamma \in G\}, \quad span(\mathcal{D}) = \mathcal{H}.$$

As stated in [78], this kind of construction, induces to covariance of the dictionary with respect to group action. This is, indeed, defining $g_\gamma = \mathcal{U}(\gamma)g$:

$$\mathcal{U}(\gamma_0)g_\gamma = \mathcal{U}(\gamma_0 \circ \gamma)g,$$

where $\circ$ indicates the group law.

Anisotropic Refined Dictionary: A Semi-Structured Dictionary

The geometric dictionary that concerns this work is generated by applying a set of diverse transformations to a mother function g. Our dictionary is spanned by a family of unitary operators U_α:

$$\mathcal{D} = \{U(\gamma)g, \gamma \in A\}, \tag{6.2}$$

for a given set of geometry transformations A. g defines the structural and morphological properties of the dictionary functions, while A captures most of its geometrical properties. In the following, we detail the generation of A in the particular case of [178] to design an Anisotropic defined (AR) dictionary for image approximations.

Definition 6.2 *([78, 178]) Let the index set A contain translations, rotations and two dilations, one for each principal direction:*

$$A = \left\{\mathbf{b} \in \mathbb{R}^2, \theta \in SO(2), (a_1, a_2) \in \mathbb{R}_*^+ \times \mathbb{R}_*^+\right\}, \tag{6.3}$$

where $SO(2)$ the notes the group composed of dilations and rotations. The action of A on an atom g is defined by:

$$U_\alpha g = \mathcal{U}(\mathbf{b}, \theta) D(a_1, a_2) g, \tag{6.4}$$

where $\mathcal{U}$ is a representation of the Euclidean group:

$$\mathcal{U}(\mathbf{b},\theta)\, g(\mathbf{x}) = g\left(r_{-\theta}(\mathbf{x}-\mathbf{b})\right), \tag{6.5}$$

and D acts as a dilation operator:

$$D(a_1,a_2)\, g(\mathbf{x}) = \frac{1}{\sqrt{a_1 a_2}} g\left(\frac{x}{a_1},\frac{y}{a_2}\right), \tag{6.6}$$

where $\mathbf{x}=(x,y)$ is the vector of plane coordinates.

When $a_1 = a_2 = a$ then $D(a,a)$ is nothing but the similitude group of the plane. Avoiding rotations, then it corresponds to a group studied in [17]. Clearly, A has not group structure in here. It presents, however, covariance with respect to the Euclidean group. The dictionary of Definition 6.2 is linked to Eq. 6.2 in the following way: each γ in Eq. 6.2 corresponds to a particular setting of $\{\mathbf{b}, a_1, a_2, \theta\}$.

The dictionary used in the scope of this chapter is generated from the following mother function:

$$g_{AR}(u,v) = C(4u^2-2)\exp\left(-(u^2+v^2)\right), \tag{6.7}$$

where C is a normalizing constant and (u,v) are, in this case, the plane coordinates. Eq. (6.7) is a 2D separable kernel made of the tensorial product of two main structures. One (the y axis) is formed by a Gaussian function that confers to it the capacity to represent smooth structures in this direction. The perpendicular direction, is formed by a *Mexican Hat* (or *Marr*) wavelet [121, 126]. This intends to represent big variations within the signal. The wavelet direction of (6.7) will respond to features with properties similar to lines and edges. A 3D visualization of an AR atom can be seen in Fig. 6.2.

Mexican Hat wavelets have just two vanishing moments [13] (thus, a limited capacity to ignore polynomial surfaces), and a very localized spatial and Fourier extension (see Fig. 6.3). In effect, Eq. (6.7) has an optimal spatio-temporal energy localization. It has a low oscillatory behavior which will translate (as it will be seen in examples) in a very low ringing effect in the approximation of images. In addition, this kind of function is very similar to the response of V1 cells of visual cortex [136], which makes coding artifacts smooth enough in order to be not too annoying for the Human observer.

Eq. (6.7) atoms have also been chosen because they can be easily anisotropically scaled, rotated and translated, in order to properly adapt to edge representation. In fact, edgy features are found in images within a large range of lengths, orientations and scales. All these transformations, represented in Definition 6.2 by the indexes set A, apply in the following way to $g_{AR}(u,v)$:

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \frac{1}{sx} & 0 \\ 0 & \frac{1}{sy} \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x-b_x \\ y-b_y \end{bmatrix}. \tag{6.8}$$

In (6.8), the first matrix by the left corresponds to anisotropic scaling of the function. Next comes the rotation matrix and finally, at the right, a translation operation can be seen. Notice that (x,y) denotes the image coordinates. During the generation of the dictionary, only one constraint, a part from those issuing from the image size, needs to be taken into account. In order to avoid ill-formed functions inappropriate to edges, the following relation must always be respected: $sy \geq sx$.

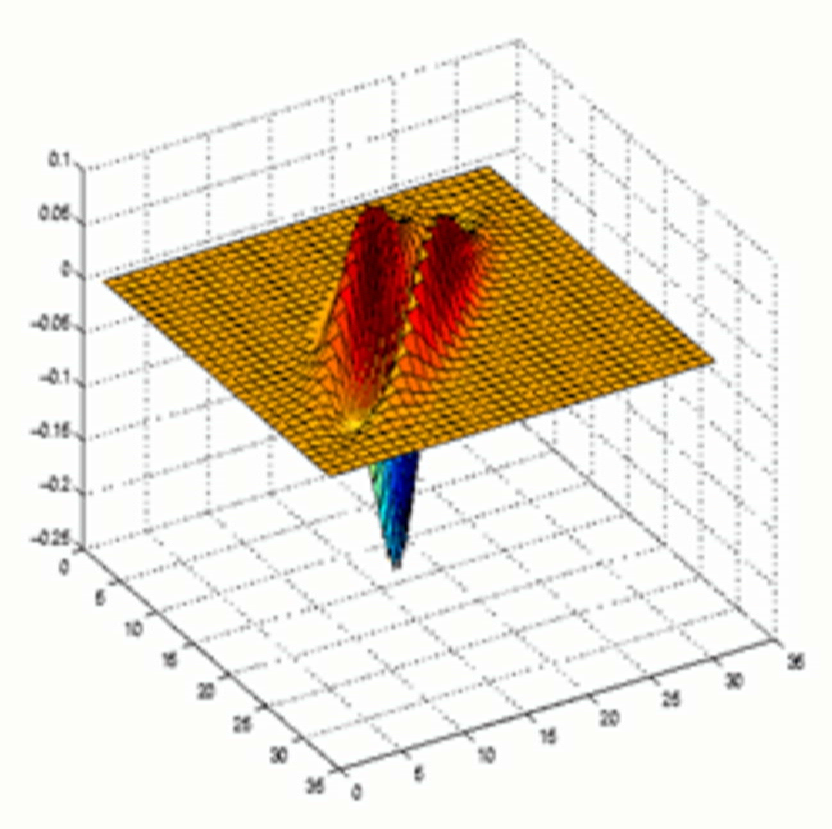
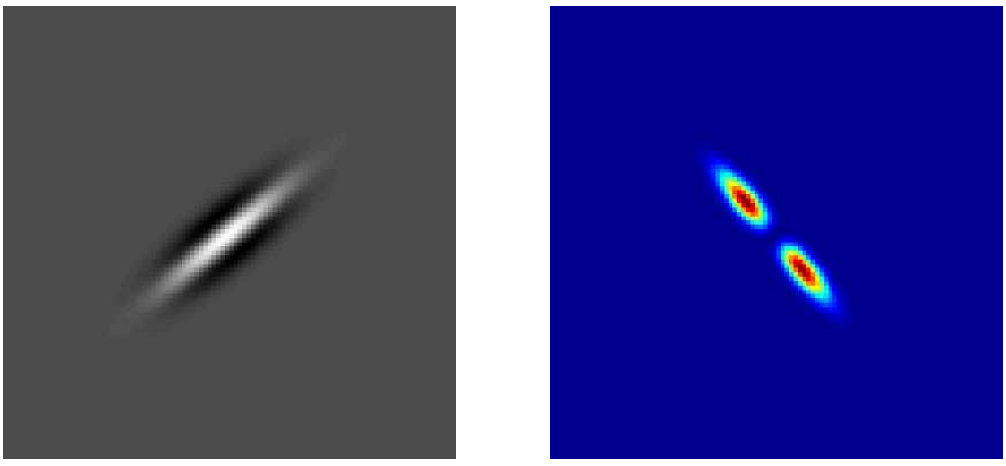


Figure 6.2: 3D view of an anisotropically refined atom in the spatial domain based on Eq. 6.7.



(a) Spatial domain **(b)** Frequency domain

Figure 6.3: An example of one atom of the dictionary in the spatial domain and in the frequency domain.

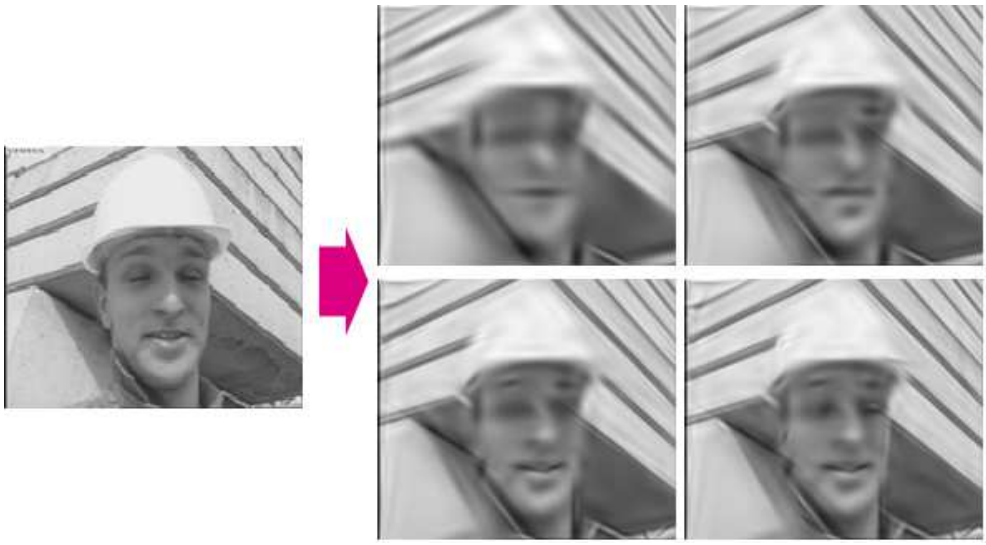


Figure 6.4: Approximation of the Foreman image. left: original picture. right: set of 4 approximations, the approximations with 50, 100, 150 and 200 AR atoms are presented in raster order.

6.3 Genetic Algorithm based MP: a Weak Matching Pursuit Implementation

The straightforward computation of all scalar products required by full search Matching Pursuit is of very high complexity. In order to reduce the computation time, one can take a suboptimal solution. Due to the non-linearity and non-convexity of the problem, standard minimization algorithms, such as steepest descent, do not perform well, because of the presence of many local minimas. One way of implementing a quicker computation of a global solution is through the use of Genetic Algorithms [43, 64, 65]. The use of this optimization algorithm for every MP iteration leads to the so called Weak Matching Pursuit (see Sec. 4.4.2). Indeed, instead of choosing the optimal solution at every iteration, a sub-optimal one is selected.

Genetic Algorithms may be appropriate for MP because they can converge whatever the minima of the problem is. The disadvantage of this kind of algorithms is that the convergence degree is merely statistical. One can never know how close to the optimal solution we are. However, greedy algorithms have been demonstrated to converge even though the solution chosen at every iteration is not the optimal [173]. Hence, the use of a Genetic Algorithm will not cause the algorithm to diverge, but it will slow down the convergence.

The use of a Genetic Algorithm (GA) in the framework of MP for image approximations was proposed in [65, 78]. The GA takes the atom parameters (translation, scaling and rotation) as *genes*. Each group of five parameters defining an atom is an *individual*. The group of N_{ind} atoms being evaluated, or the group of N_{ind} individuals alive in a certain moment is the *population*. The individuals of the population will have descendants through *crossover*, *mutations* and *survival of the fittest*, which will form the new *generation* of the population. Normally the algorithm will go on until a certain number of generations N_g is achieved or until the minimal error has been reached. A diagram of how this algorithm works is depicted in Figure 6.5.

The computational complexity of the GA directly depends on the number of individuals in the population and on the number of allowed generations. The number of scalar products to compute in order to obtain an MP coefficient through the GA will be, at most, $N_{ind} \cdot N_g - N_g + 1$ (from

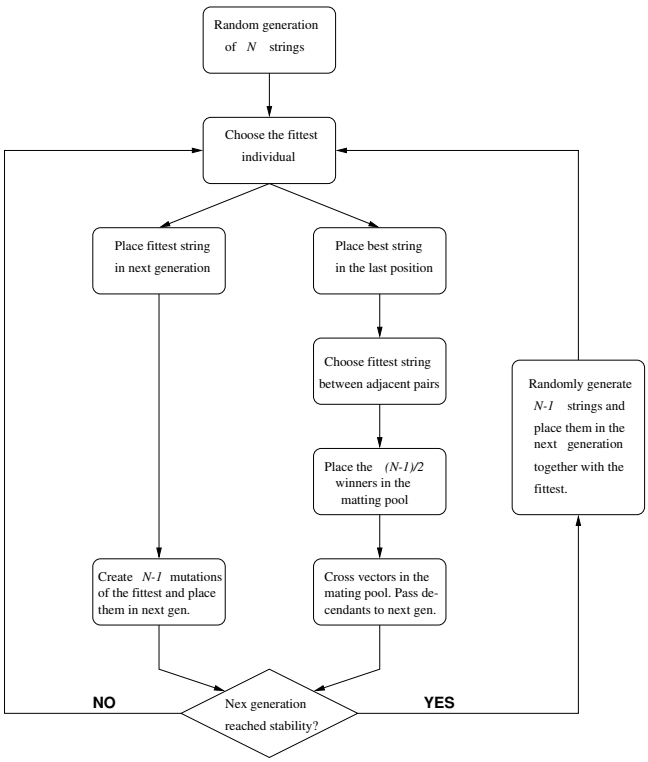


Figure 6.5: Genetic Algorithm block diagram.

one generation to the other some individuals can be kept without change due to the randomness of mutations and crossovers). Indeed, the computation of the winner is done only once, because then it passes from one generation to the other. If no constraint is given on the dictionary (so, dictionary atoms are not necessarily bounded, and its size is taken equal to the image size), performing a scalar product will imply $N = Size_x \cdot Size_y$ products and sums (with $Size_x, Size_y$ being the image size). The final computational complexity of the GA is given by:

$$O\left(N \cdot N_{ind} \cdot N_g\right). \tag{6.9}$$

When compared to the "Brute Force" Full Search MP implementation (see section 6.4.1), this complexity is lower, but still high, specially if taken into account that the GA introduces an approximation error which is difficult to quantize due to the randomness of the method.

6.4 Full Search MP

Sub-optimal algorithms can advantageously reduce computational complexity. However, as seen in Sec. 6.3, a price has to be paid for this. In such a case, approximation quality is degraded by a certain factor due to the heuristic nature of the algorithm. This can negatively affect the performance of the application that uses the generated signal expansion. In fact, errors committed during the functions selection state will require further projections in order to be corrected. Errors turn into a decrease of the coefficients decay and, consequently, a reduction of the approximation sparsity.

For applications where sparsity and accuracy play a critical role, there is a determinate interest in the availability of feasible Full Search MP approaches. In this way, the best possible MP expan-

sion may be achieved avoiding heuristic or uncontrolled imprecisions during the function selection procedure.

6.4.1 "Brute Force" Full Search MP

According to the MP algorithm principle (see Sec. 4.4.2), at every iteration, it is necessary to browse through all functions that form the given dictionary in order to perform all scalar products between dictionary atoms and the signal. Hence, the complexity of a full search MP iteration directly depends on the size of the dictionary times the complexity of a single scalar product. In general, and without any constraint on the dictionary in use, the scalar product can be considered to involve all pixels of the image (non-separable atoms with no compact support). Thus, considering both contributions,

$$O\left(N_F \cdot N\right) \tag{6.10}$$

operations are needed to compute all scalar products, where $N = Size_x \cdot Size_y$ and N_F is the number of dictionary functions.

The choice of atoms in MP expansions aims at finding the best spatio-frequential representation at the best location within the image to be represented. Considering the fact that a precise localization of the "spatio-frequential" feature may be of use, all possible displacements for a certain shape of atom must be taken into account (see the group based design of dictionaries in Sec. 6.2.2). Hence, most functions of a certain dictionary are just displaced versions of a model sharing the same frequential properties. Thus,

$$N_F = M_F \cdot Size_x \cdot Size_y = M_F \cdot N, \tag{6.11}$$

where M_F is the number of model functions and the N the total number of allowed displacements within an image.

It follows from Eq. (6.11) that the complexity of a full search MP with a dictionary, such as the previously described, is given by

$$O\left(M_F \cdot N^2\right) . \tag{6.12}$$

Of course, in this complexity estimation, the cost of generating all functions of the dictionary is not taken into account. We assume that functions are generated once and stored into memory.

Given the dictionary example of Sec. 6.2, we see that the dictionary generation is also an important point to take into account concerning the complexity of the whole expansion algorithm. Actually, the fact of having an analytical generating expression for the dictionary functions may lead to recompute the concerned atom at every step. This, fruit of some pragmatism from a programing point of view, may bring a very important overhead to computations. The cost of computing a general atom is, *a priori*, related with the size of the image. Hence assuming that the number of operations (Δ) needed to compute a pixel is $\Delta << N$ turns to be $O\left(N\right)$. Anyway, this bound may not be representative of the real complexity. Depending on the function to be computed at every pixel (see Eq. (6.7)), it may be required to consider a cost of $O\left(N \cdot \Delta\right)$.

In order to reduce the impact of the complexity (Δ), look-up tables may be defined with a sufficiently dense set of pre-computed values of the critical function (e.g. "exp()" in Eq. (6.7)). The corresponding value can be then approximated by its nearest neighbor in the look-up table. This approach can be very interesting in cases where all dictionary atoms are generated using the same analytical expression or a similar one (e.g. group theoretical design dictionaries).

A more exhaustive solution consists to store all dictionary functions such that no extra calculation is needed (apart from scalar products and memory transfers). This is equivalent to dispose of a numerically defined dictionary. Here, the complexity problem of computing all atoms, becomes a

memory problem, in terms of size and bandwidth. In a general framework with no constraints on the dictionary, it would require $O(N_F \cdot N)$ memory words. This can be absolutely impractical for a very dense dictionary.

Coming back to our example dictionary (Eq. (6.12)), $O(N)$ functions per each of the M_F frequency models represents a $O\left(M_F \cdot N^2\right)$ in terms of memory requirements, which is absolutely impracticable. On the other hand, the dictionary is translation invariant and only a single numerical centered version of each of the M_F models is necessary to be stored. A simple convolution would solve all scalar products for each one of the frequency models.

6.4.2 Spatial Invariance in Scalar Product Computations and Boundary Renormalization

Considering the operations performed in one iteration of the full search MP:

$$|\langle r_k f, g_{\gamma_n}\rangle| = \sup_{\gamma\in\Gamma} |\langle r_k f, g_{\gamma}\rangle| \,, \tag{6.13}$$

and the translation invariance, then

$$|\langle r_k f, g_{\gamma}\rangle| = \left|\langle r_k f, g_{\gamma_{M_F}^{dx,dy}}\rangle\right| = \sum_{\tau}\sum_{\lambda} r_k f(\tau,\lambda) g_{\gamma_{M_F}}(\tau - dx, \lambda - dy), \tag{6.14}$$

where $\gamma_{M_F} \in \Gamma_{M_F}$ and Γ_{M_F} is the sub-set of different frequency models in Γ. This means all different functions Γ_{M_F} s. t. $\gamma_{M_F}^{dx,dy}\ \forall M_F$ and $dx = \frac{N}{2}, dy = \frac{N}{2}$.

A common problem in image processing is the finite nature of signals. Images are defined over a finite domain. If this is not taken into account, then sparsity in representations may decrease.

A typical example is given by image representations through the classical wavelet transform, which introduces many high valued coefficients if no specific periodization, mirroring or boundary wavelet are applied [121].

An image $I(x,y)$ defined in a domain s.t. $x, y \in [0, Size_{x,y})$ can be considered as the windowed version of a infinite 2-D signal I_{inf}:

$$I(x,y) = \Pi\left(\frac{x - Size_x/2}{Size_x}, \frac{y - Size_y/2}{Size_y}\right) I_{inf}\,(x,y)\,, \tag{6.15}$$

where $\Pi\,(x,y)$ is a squared 2-D window defined for $x, y \in [0,1)$.

In our case, the use of periodic extensions of the image would definitely have a negative impact in the number of functions needed to represent the boundary, as artificial structures would be introduced. Mirroring would be a better choice and the energy spreading on the coefficients would be significantly lessen. Anyway, this approach introduces also artificial structures such as bending of oriented features (e.g. edges) on the image boundary.

A possible solution is to introduce into the dictionary additional boundary adapted functions. The proposed supplementary functions are the windowed version of the anisotropically refined functions described in this chapter. They need to be reweighted with a normalization factor such that they continue to have unitary norm. In this way, coefficient expansions continue to have the same properties defined for MP signal decomposition on unitary dictionaries. The contribution of this new class of elementary functions is the ability to catch regular structures of the signal that suddenly terminate at image boundaries.

The scalar product generation can thus be reformulated as:

$$|\langle r_k I, g_\gamma \rangle| = \left| \langle r_k I, g_{\gamma_{M_F}^{dx,dy}} \rangle \right| =$$

$$\sum_\tau \sum_\lambda r_k I(\tau,\lambda) \frac{g_{\gamma_{M_F}}(\tau - dx, \lambda - dy)}{\left\| g_{\gamma_{M_F}}(\tau - dx, \lambda - dy) \cdot \Pi\left(\frac{\tau - dx - Size_x/2}{Size_x}, \frac{\lambda - dy - Size_y/2}{Size_y} \right) \right\|_2}. \tag{6.16}$$

$$\Pi\left(\frac{\tau - dx - Size_x/2}{Size_x}, \frac{\lambda - dy - Size_y/2}{Size_y} \right) =$$

$$\sum_\tau \sum_\lambda r_k I(\tau,\lambda) \cdot \tilde{g}_{\gamma_{M_F}}(\tau - dx, \lambda - dy),$$

where $\tilde{g}_\gamma$ represents the new dictionary of functions that is not space invariant due to a weighting factor that depends only on the spatial location.

Thus, we consider to separate the calculation of the scalar product into two steeps:

- First, the common part of all functions of the same kind (same non bounded spatio-frequential properties but not same position) are used in the convolution.
- Afterward, each obtained coefficient is weighted by the normalization factor of its projecting function that takes into account its intersection with boundaries.

Note that the computation of normalization factors increases the complexity. In a general dictionary, this complexity depends on the total number of functions of the dictionary and the size of the signal. Considering the general situation where all functions of a dictionary may intersect with boundaries, independently of its position within the image, and Eq. (6.11), it turns out that N_F normalization factors are needed and up to

$$O\left(M_F \cdot N^2\right) \tag{6.17}$$

operations can be necessary to compute them. These normalization factors are used at every iteration of the MP algorithm. Furthermore, they do not depend on the signal. These features allow computing them once at the beginning of the expansion process and storing them their posterior use in subsequent MP iterations. Eq. (6.11) trivially implies

$$O\left(M_F \cdot N\right) \tag{6.18}$$

memory words, which in some cases and depending on the dictionary size (and image size) is feasible.

6.4.3 FFT Based Full Search MP: From Scalar Products to Spatial Convolution

In Sec. 6.4.1, the computation of the whole set of scalar products was presented as a very computationally complex task. Therefore, a fast convolution algorithm must be used in order to make possible a full search strategy with large dictionaries. This can be done thanks to the Convolution Theorem [20], which states the way of using the FFT for this purpose [22, 133]. In this way, Eq. (6.12) becomes:

$$O\left(M_F \cdot N \log\left(N\right)\right). \tag{6.19}$$

Furthermore, functions can be directly stored into memory in their transformed version such that only one inverse FFT per Fourier template is needed. In the case of a dictionary of non-complex

functions, the amount of memory needed for storing Fourier domain versions is not bigger than those in the spatial domain. However, the spatial zero padding has to be taken into account in the transformed version of the function. Furthermore, the normalizing mask has to be stored as well, which increases even more the memory needs. The asymptotic memory requirements is finally of:

$$O\left(M_F \cdot N\right) \tag{6.20}$$

memory words.

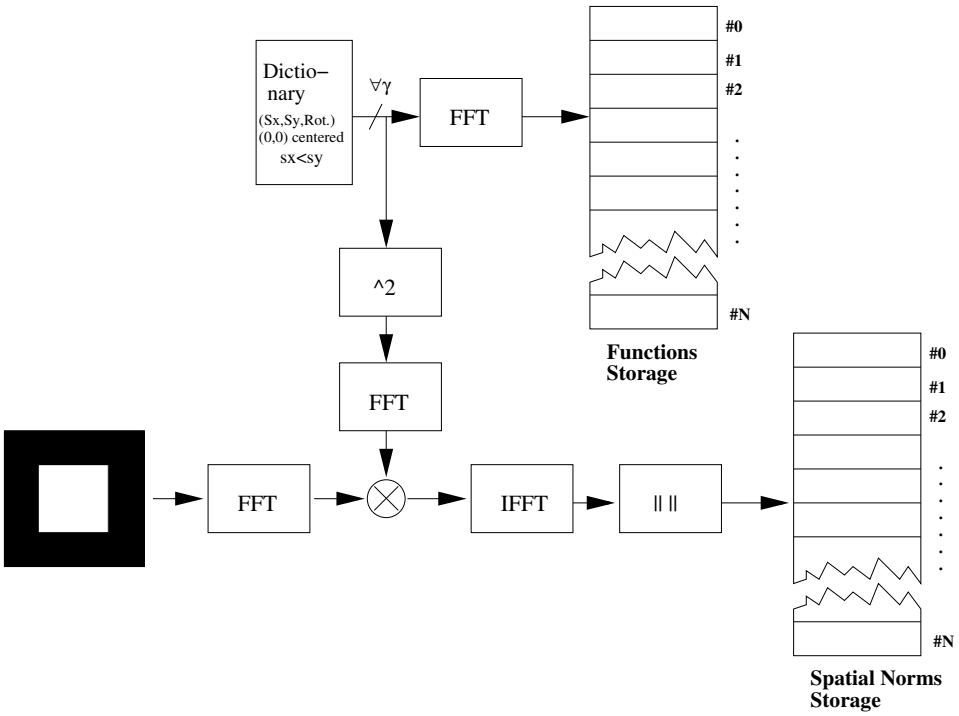


Figure 6.6: Description of the setting up procedure generating the look up tables to be used in Fig. 6.7.

In Fig. 6.6 appears the scheme used for the generation and storage of dictionary atoms and weighting masks. Weighting masks are generated by convolving a centered squared atom by a image size pulse function of unitary amplitude.

Fig. 6.7 illustrates the general procedure used to generate all scalar products from the stored Fourier atoms and spatial weighting masks. Convolutions are performed on the Fourier domain. Once done, it only remains the supremum search in order to find the appropriate result.

As we will show in the following sections, the Matching Pursuit approach can benefit from very interesting improvements in terms of computational cost and memory requirements when special care is taken on the structure of the dictionary. Before that, let us examine some results of the full search algorithm and compare them to those obtained by the GA based approach.

6.4.4 Results: Full Search vs Genetic Algorithm Search

In this section we compare the performance of the FSMP algorithm versus the Weak MP based on the GA. In table 6.1, we state the relation of computation times required by each algorithm and the approximated quality achieved. Experiments have been performed on a Dual Xeon 2.2 GHz

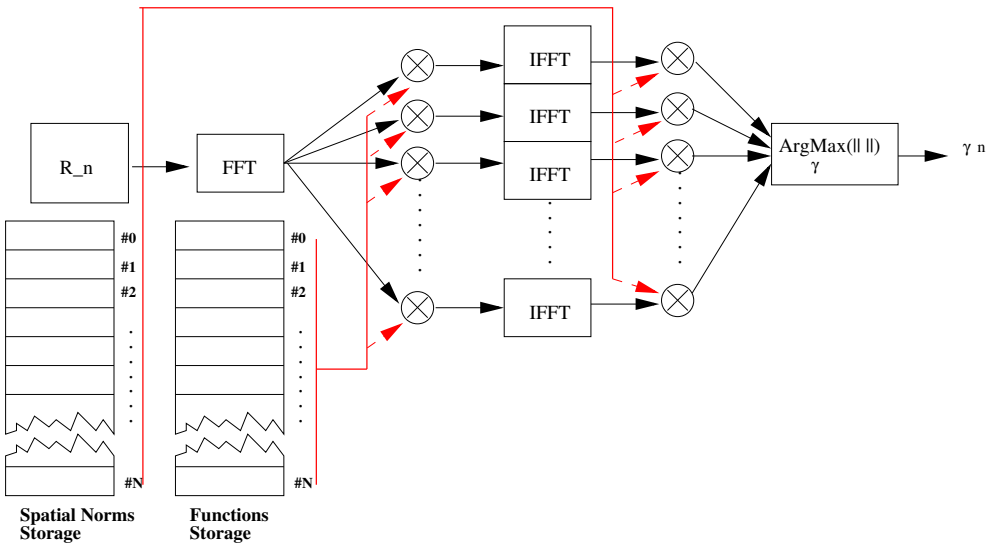


Figure 6.7: Schematic description of the full search algorithm for one iteration of MP. Look up tables are used to hold the DFT of the dictionary functions and the normalizing masks.

with 2GB of RAM. We note the improvement of 3.8 dBs achieved by our FSMP in images with an important piecewise-smooth component like *Lenna* and *Cameraman* with a significantly smaller computational time. In the case of *Baboon*, the improvement is only 1.6 dB due to the higher complexity of the signal (much more textured). In the 3 cases, approximations have been generated by recovering the first 300 terms of their MP expansions.

Image	Algorithm	Computation Time	PSNR
Lenna 128x128	GAMP	156' 26.909"	26.1506 dB
	Full Search MP	126' 29.792"	29.9076 dB
Cameraman 128x128	GAMP	157'3.249"	25.0485 dB
	Full Search MP	128'13.957"	28.8270 dB
Baboon 128x128	GAMP	160'19.048"	25.1374 dB
	Full Search MP	128'9.170"	26.7932 dB

Table 6.1: Comparison of the computational time and the image quality obtained with the Genetic Algorithm, using 39 individuals and 75 generations, and the Full Search algorithm.

Figures 6.8, 6.9, 6.10 show a visual comparison of both algorithms. In the FS case, the exhaustive search at each iteration reduces the *geometric* noise introduced by the GA. The GA, is not able to converge precisely at every iteration in order to select the best atom. This error appears under the form of lines and ridges. They are very noticeable when they correspond to relevant salient geometric structures like edges.

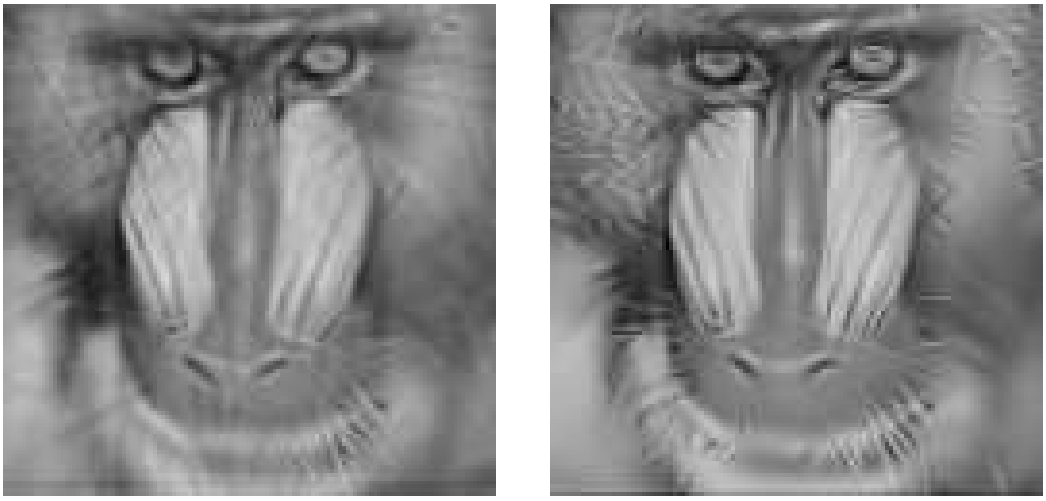
Finally, to give an overview of the convergence through MP iterations, Fig. 6.11 presents a comparison of both, FS and GA based algorithms. The Lenna image has been used for experiments. As it can be seen, the FS algorithm outperforms through all MP iterations the weak greedy algorithm based on the GA.

(a) Genetic Algorithm **(b)** Full Search Algorithm

Figure 6.8: Visual comparison of Lenna 128x128 decomposed with MP with 300 coefficients (a) with the Genetic Algorithm and (b) with the Full Search MP.

(a) Genetic Algorithm **(b)** Full Search Algorithm

Figure 6.9: Visual comparison of Cameraman 128x128 decomposed with MP with 300 coefficients (a) with the Genetic Algorithm and (b) with the Full Search MP.



(a) Genetic Algorithm **(b)** Full Search Algorithm

Figure 6.10: Visual comparison of Baboon 128x128 decomposed with MP with 300 coefficients (a) with the Genetic Algorithm and (b) with the Full Search MP.

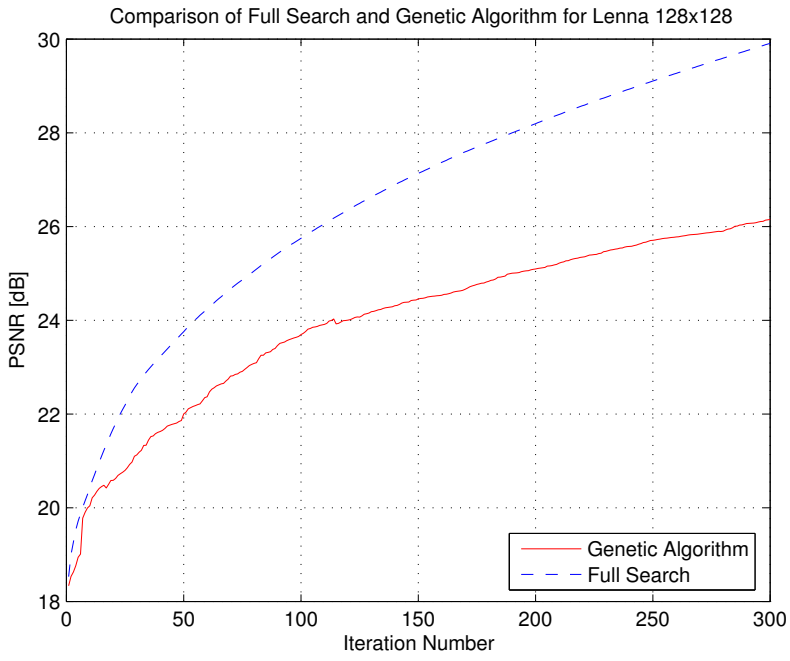


Figure 6.11: Evolution of the PSNR with the iteration number for the Genetic and the Full Search Matching Pursuit algorithms. The image used in this case is the gray-scale 128x128 Lenna.

6.5 Exploiting the Dictionary Features

In this section, we explain how computational and memory complexity can be further reduced. Several results illustrate gains and limits of methods used to reduce the complexity.

6.5.1 Taking Profit of Spatio-Temporal Energy Localization: Compact Support and Atoms Approximation

In terms of computational complexity, it is very interesting to deal with compact support basis functions. We can define two types of compactness: compactness in frequency and compactness in space. Natural image decompositions often use dictionaries where functions have localized support in space, e.g. the basis functions used for classic dyadic wavelet decompositions [121]. Concerning the dictionary used in this work, even if functions do not have strictly compact support, their energy is mostly compacted in a very localized fashion. Moreover, the fact that atoms envelop is Gaussian shaped implies that the energy is localized in space and frequency domains. As seen in the following, this allows for some approximations of the dictionary which significantly reduce the memory usage and decrease the number of multiplications needed at each full search MP iteration, without introducing any distortion nor affecting the rate of convergence of the MP algorithm.

A first advantage of spatially localized dictionaries is the possibility to use the so called M-fold MP strategy [82]. This consist in recovering M atoms at each MP iteration. Indeed, distant atoms in space that have a localized support are often incoherent among them. Some heuristics can be based on this and contribute to speed-up by a factor M.

Thanks to the incoherence among spatially distant atoms, further savings are possible. Indeed, for a given MP iteration, once an atom is selected, all scalar products of atoms that are highly incoherent with the selected one do not need to be recomputed. This allows a way of parallelizing MP. Scalar products of different influence zones may be computed in parallel. In some cases, this strategy, can also be coupled with the M-fold approach, leading to an even better general performance.

Our full search solution is based on the massive use of the FFT in order to reduce the complexity scalar products computation of all translations of a given atom. For this purpose, all centered Fourier atom versions are stored into memory such that they need to be computed only once. The fact that an atom and the normalizing modulus masks (as described in Sec. 6.4) may have few significant non-zero samples in an image makes possible to optimize the storage by efficiently indicating where they are. Moreover, the run-length description of their positions can also be used to reduce the number of multiplications involved in the algorithm.

Memory Compression

Fig. 6.12 illustrates the kind of information contained in maps that represent the frequency domain form of atoms (see the first row), and normalizing factor maps (see the third row). It can be seen, in the second and fourth row of Fig. 6.12, respectively the areas of each map where relevant information is cumulated for the Fourier domain atoms and normalization masks. Concerning the Fourier domain version of atoms in this examples, one sees within the second column, in white, the support where the modulus of the Fourier transform is greater than a given threshold (10^{-3} in this example). In the fourth column, areas in white are the spatial locations where the renormalization factor differs from 1 by more than a certain threshold - 10^{-3} in the presented example - (i.e. factors smaller than 1 are used to re-normalize scalar products of atoms that significantly cross image boundaries). In both cases, only original values corresponding to the white *footprints* are required to be stored, the

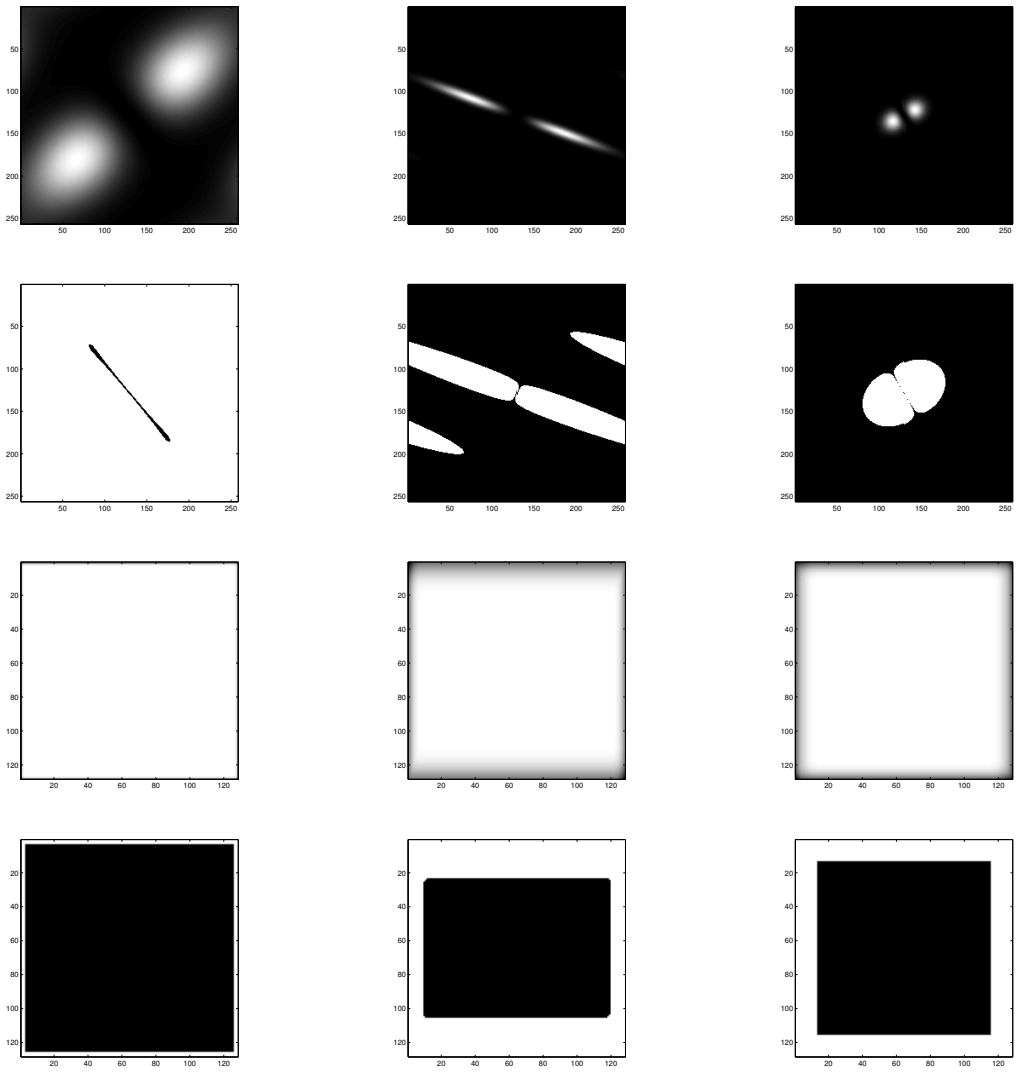


Figure 6.12: First row: frequency modulus of three selected dictionary atoms. Second row: Respective support of the significant values that need to be stored in memory (values with modulus greater than 0.001). Third row: Spatial renormalization maps that correspond to the atoms of the first row. Fourth row: Binary mask that determines where values different from 1.0 with a significant difference (greater than 0.001) are located.

rest can be approximated by zero for the Fourier atoms case and by one for the re-normalizing mask case.

What makes feasible the compression in both cases is that zeros and ones cumulate in consecutive memory positions. Hence, nothing easier than using the basic run-length coding [156] to store into memory more efficiently default values (zero or one) for every kind of map. Of course, better techniques using quad-trees or prune-join trees (like in image compression techniques [163]) could be used, however this is out of the scope of this work.

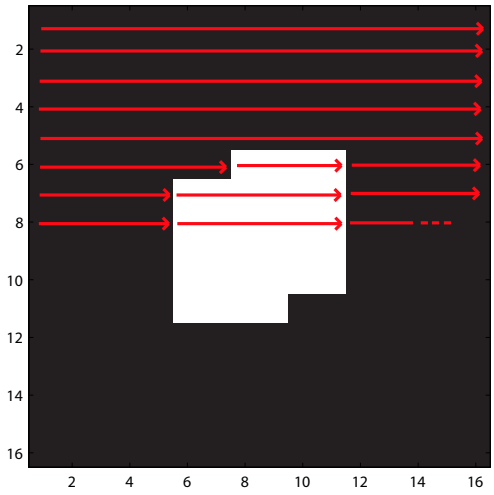


Figure 6.13: Every Fourier domain atom is stored in a run-length fashion. All values considered insignificant are set to zero. All consecutive zeros (considering a raster scanning of images) are efficiently stored using a single integer. Significant values are stored one by one together with an integer number that specifies how many significant values are consecutively aligned.

Fig. 6.13 depicts the raster scanning used for storing a simple example. In a top-down fashion, the signal is scanned line by line. Two lists of values are kept: one contains the list of elements in every run of significant pixels and the other contains the length of each run of values (significant and insignificant).

Examples

In the following, the effect of approximating the stored Fourier domain atoms and normalizing maps is analyzed. We compare in 6.14 and 6.15 the way memory compression affects the quality of the approximation of Lenna (gray-scale 128x128 pixels) when using the FS algorithm. The degree of memory compression is varied by means of changing the threshold used to decide which are the significant values to keep an use. Experiments are done by ranging the threshold from zero up to 100.

As it can be appreciated in the figures, the dictionary can be approximated such that up to more than 75% of memory is saved without any distortion increase. Fig. 6.15 visually shows, that a very high threshold can be set before important distortions are perceived.

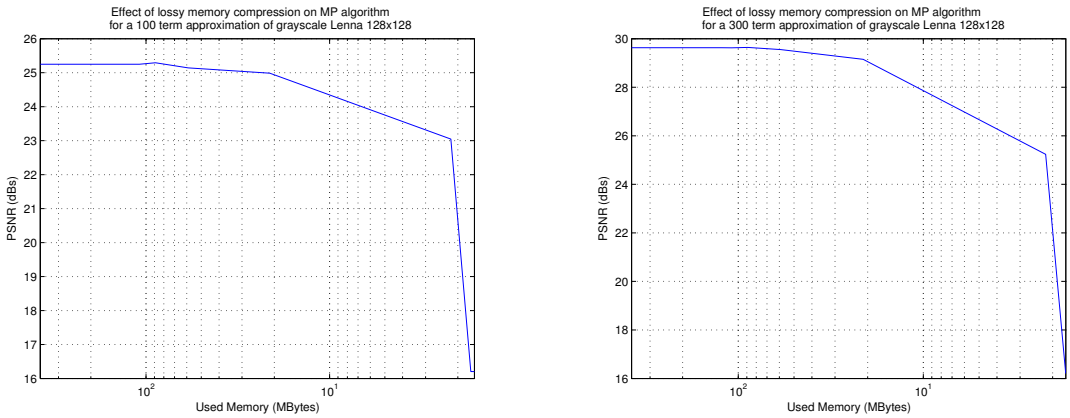


Figure 6.14: Approximation PSNR vs memory used of the FFT based full search algorithm for Lenna 128x128. Up to 75% in memory savings can be achieved without loss in approximation convergence. In the left the approximation is done with 100 terms. In the right, 300 terms are used.

Figure 6.15: Visual comparison of the different approximation of Lenna 128x128 for the amounts of memory of: (from left to right and then up/down) 377 MB (threshold 0), 122 MB (threshold 10^{-4}), 22 MB (threshold 1) and 2.2 MB (threshold 10).

6.5.2 Steerability of Atoms and Complexity Benefits

Steerability is an interesting tool when dealing with orientable functions. This is based on the principle that certain classes of kernels, can be tuned according to a set of geometric parameters. These can be decomposed in a linear combination of few functions sampled from that parameter space. For example, for the kind of atoms used in this work, if both axes have the same scaling factor ($1/sx = 1/sx = a$ in Eq. (6.8)), then it is possible to generate all possible orientations of these from just 3 basic functions with different orientations [77]. It is common to use orientations of 0, $\frac{\pi}{3}$ and $\frac{2\pi}{3}$ (see Fig. 6.16). However, the concept of steerability can be extended to any other affine transformation a part from that of the rotation. In some cases, the steered functions are only approximations of ideal ones. An extensive review of steerability can be found in [77, 99, 100, 123, 143]. In terms of computational advantages, the use of steerability would significantly contribute to

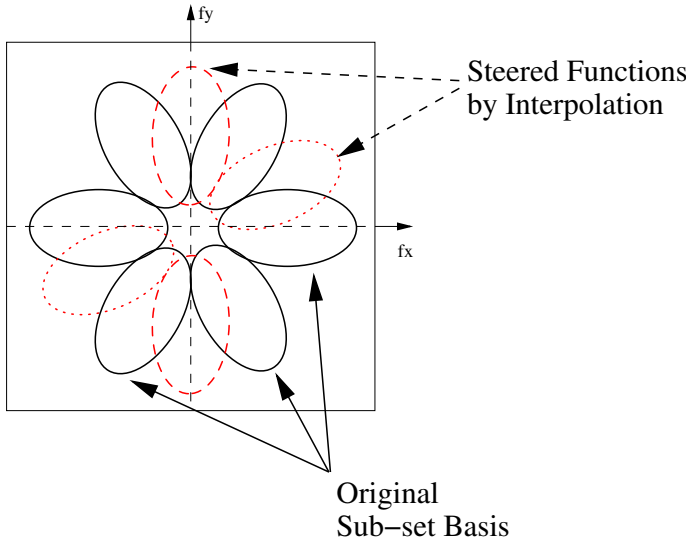


Figure 6.16: In some cases, a whole set of functions from the dictionary can be generated from the linear combination of a subset of these. This is the case, for example, for isotropically scaled Gaussian second derivatives [77]. Only three real filters (drawn in black) are needed to generate all remaining orientations.

reduce the number of necessary inverse Fourier transforms. To give an example, let us take again the isotropically scaled second derivative of a Gaussian. Consider we desire to compute the scalar products of all translated versions of N_θ different orientations of our steerable kernel. For that purpose, if we store into memory a Fourier domain version of the function for each one of the N_θ orientations, as described in Sec. 6.4, a total number of:

$$O(N) \cdot N_\theta + O(N \log N) \cdot N_\theta \tag{6.21}$$

operations will be necessary (where N is the size of the data). In (6.21), the first term corresponds to the product in the transformed domain for filtering, and the second indicates the complexity associated to all inverse Fourier transforms. If steerability is used instead, and all needed orientations are obtained by linear combination of a basic set composed by three functions, the number of necessary operations turns into:

$$O(N) \cdot 3 + O(N \log N) \cdot 3 + O(N) \cdot 5 \cdot (N_\theta - 3), \tag{6.22}$$

where the first two terms correspond to the complexity required by the Fourier domain filtering and the last term corresponds to the steerability based computation of the remaining orientations. The factor 5 in the last term of (6.22) is due to the three products and two additions of the linear combination used to steer filters. Fig. 6.17 illustrates the potential savings of using steerability in the computation of scalar products for the simple example of isotropically scaled anisotropic second derivatives of a Gaussian. Even for small N, the use of steerability is computationally advantageous. Moreover, the bigger the N the more resources are spared. According to the particular example given by (6.21) and (6.22) for $N_\theta = 36$, when $N \to \infty$ the spare factor converges to 12. Moreover, in terms of memory usage, the amount of needed storage falls down of a factor $\frac{N_\theta}{3}$. This not only reduces the memory consumption but also the time required to transfer data from memory to the processor.

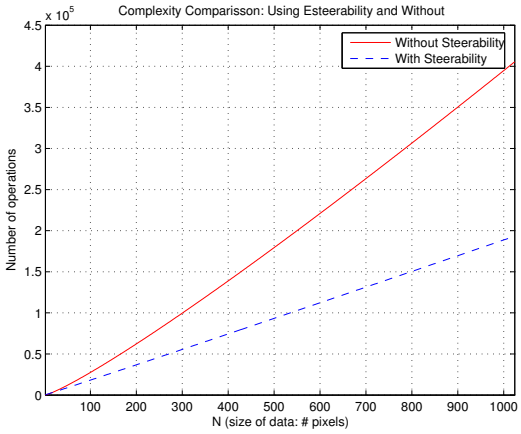


Figure 6.17: Comparison of the growth of complexity with and without using steerability as a function of N. In this graph, we have assumed $N_\theta = 36$.

Anisotropic scalings can be approximated by simple linear combinations of isotropically scaled atoms that are aligned according to their longitudinal axis. In fact, linear combinations of few scalar products obtained with the angular steerability of isotropically scaled atoms, should allow the recovery of approximate values for scalar products corresponding to anisotropically scaled versions of (6.7) [77, 99, 100, 123, 143].

Although advantages of introducing the steerability principle in the analysis step of each MP iteration are stated here, its practical implementation and integration is beyond the scope of this work and will be left for future work.

6.6 Conclusions

In this chapter, a strategy has been presented to allow the implementation of full search matching pursuit for very dense dictionaries with spatially invariant atoms. This appears to be of capital importance for the efficient approximation of images, if compared to suboptimal approaches like genetic algorithm based ones. Indeed, the technique presented here has allowed the investigation and implementation of a very interesting flexible low bit rate image coding scheme [66, 67, 79].

The use of the FFT allows to quickly calculate all scalar products of the displaced versions of a template thanks to the *Convolution Theorem*. Moreover, we have seen how some dictionaries allow to introduce relevant enhancements in order to reduce complexity and memory usage. Spatio-

frequential energy localization of dictionary functions is of key importance for that. Furthermore, for the particular dictionary used in this work, the use of steerability techniques seem very appropriate to further reduce complexity.

Very fast algorithms can be developed for particular dictionaries. Structured dictionaries and those that can profit from the steerability properties may contribute to usable, real-time, signal decomposition techniques based on redundant dense dictionaries. More detailed research is left as future work.

Chapter 7

A Geometric Video Representation Using Redundant Dictionaries

7.1 Motivation: Sequence Modeling

Natural image sequences are composed of successive projected snapshots of 3D objects. Considering these objects to be smooth and their trajectories to be smooth functions of time, one usually assumes that image sequences are well modeled by smooth transformations of a reference frame [183]. Of course this assumption has intrinsic limitations: natural sequences display a wide variety of transient behaviors such as occlusions, appearance and disappearance... A schematic illustration is depicted in Fig. 7.1.

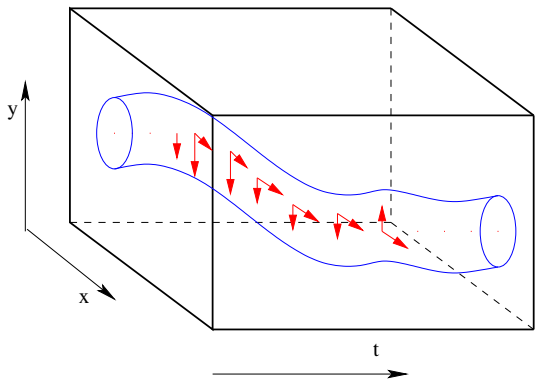


Figure 7.1: Schematic smooth evolution of an object through time.

This basic smoothness assumption is also "visually" justified in Fig. 7.2 where we display a section (a line here) of the Foreman sequence. Time is on the horizontal axis and the whole image looks very smooth, despite the shaky nature of the sequence.

Local geometric transformations are tightly linked with the motion model and the nature of the real 3D scene. When the support of moving regions is sufficiently small, a simple translational model can successfully represent motion. This is the key ingredient of most block matching techniques in motion compensation: the anchor frame is chopped into small primitives which are assumed to just

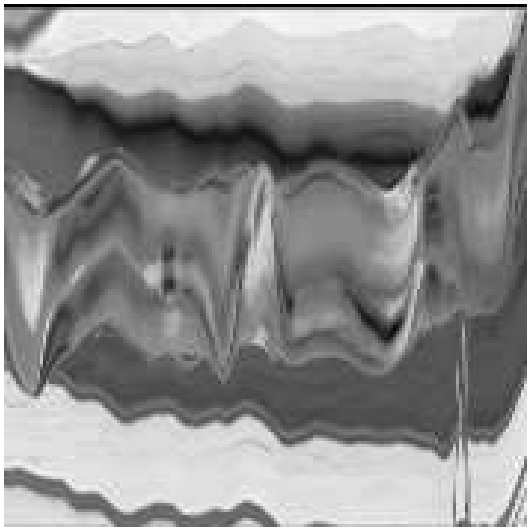


Figure 7.2: Temporal evolution of a pixels row (the 77th from QCIF version) in foreman (from frame 0 until frame 176).

translate in time. The whole model is then represented by a translation vector field and a reference image. Complex and more accurate models have also been considered, for example affine models [24] (see also Chapter 2). These allow for local expansions or contractions and are usually represented by mesh deformations [10]. More precisely the motion model is locally represented by a linear geometric transformation:

$$\mathbf{u} = \mathbf{A}\,(\mathbf{x} - \mathbf{d})\,. \tag{7.1}$$

The 2×2 squared matrix $\mathbf{A}$ and vector $\mathbf{d}$ implement translation, rotation, shearing and scaling operators:

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \frac{1}{sx} & 0 \\ 0 & \frac{1}{sy} \end{bmatrix} \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \\ \mathbf{d} &= \begin{bmatrix} d_x \\ d_y \end{bmatrix}, \end{aligned} \tag{7.2}$$

where sx and sy are scale parameters, m is a shearing factor to change the angle among x and y axis ($m = 0$ if axis are kept perpendicular) and θ parametrizes geometric changes due to rotations.

Most video representation paradigms separate motion information from image structures. In this chapter, we would like to jointly represent image geometrical structures *and* their evolution through time.

Similarly than for the case of images, geometric structures are very relevant in video signals. In the framework of video, 2D geometric features, in many cases, follow temporal geometric transformations. A sparse spatio-temporal representation of video would be based on the superposition of 3D primitives that capture spatial geometry and temporal evolution.

In Sec 1.1.3, several geometric image representation approaches are recalled. A very interesting approach, due to its flexibility, seems to be the use of large over-complete libraries of basis functions, able to represent salient 2D geometric features of the signal. Hence, a possibility is to port such an approach to video in order to extract the 3D features that compose the signal.

The construction of a dictionary of 3D atoms, representing spatial geometry and temporal evolution, is practically unfeasible. The size of the problem and the high number of freedom degrees, would make the search of the most appropriate atoms an intractable task. Moreover, temporal geometric transformations are often so complex that a geometric dictionary able to represent them would be extremely coherent. The adverse effect of such a coherent dictionary to highly non-linear algorithms, like MP, can be expected to be even worse than for the 2D case. It is necessary to adopt strategies that take some *a priori*, about the signal, into account (see Chapter 5). In the following, basic assumptions done for the recovery of 3D spatio-temporal geometric features are exposed.

Given a set of images belonging to a sequence, the changes suffered from frame I_t to I_{t+1} are modeled as the application of an operator F_t on the image I_t such that

$$\begin{aligned} I_{t+1} &= F_t\left(I_t\right), \\ I_{t+2} &= F_{t+1}\left(I_{t+1}\right) = F_{t+1}\left(F_t\left(I_t\right)\right), \\ I_{t+3} &= ... \end{aligned} \tag{7.3}$$

where the subindex t indicates time.

From the image model of Eq. (6.1) and (7.3), $\hat{I}_{t+1}$ is modeled as a transformation of the geometric representation of $\hat{I}_t$ (where $\hat{I}_t$ stands for an approximation of I_t):

$$\hat{I}_{t+1} = F_t\left(\sum_{\gamma\in\Gamma} c_\gamma^t \cdot g_\gamma^t\right). \tag{7.4}$$

A relation needs to be established between F_t and the transformation of each one of the 2D components involved in the frame approximation. This is why we make the hypothesis that F_t is composed by the set of F_t^γ that independently transform each one of the frame expansion terms, i.e.

$$\hat{I}_{t+1} = \sum_{\gamma\in\Gamma} F_t^\gamma\left(c_\gamma^t \cdot g_\gamma^t\right). \tag{7.5}$$

No global simple joint transformation model can be established for the transformation of all geometrical primitives. Hence, local transformation models need to be considered to represent complex motion transformations from frame to frame. Since the basic elements of frames approximations are g_γ^t atoms, local motion transformation models are applied to these.

In the following, F_t is sometimes referred to as a *deformation*. The action of each F_t^γ in (7.3) corresponds to a geometrical operation on g_γ and to a change of its coefficient c_γ^t. Intuitively, this mechanism intends to implement a local change of scale, position and orientation of each primitive (see Fig. 7.4(a) and 7.4(b)). The sequence of deformations $F_t^\gamma : t \in [T_1, T_2]$ and the 2D atom g_γ^t form a 3D primitive that represents how local scene geometry flows through time.

In this chapter, a representation of video sequence based on the superposition of 3D primitives, taking into account spatial geometry and temporal motion, is presented. Given the initial MP decomposition of a frame at time t (as can be found in Chapter 6), 3D primitives are implemented as the series of geometrical deformations of the 2D atoms used to represent the initial frame. These can be seen as: the evolution of each one of the 2D atoms found in the initial frame is tracked through time. To guarantee smoothness in trajectories, Markov Random Fields are used to impose regularity. This approach toward geometric video approximations is studied from a *sparse approximations* with greedy algorithms point of view. The problem is modeled according to the theoretical aspects discussed in Chapter 5.

The present chapter is structured as follows: Sec. 7.2 formally states the optimization problem for the forward prediction of frames in order to extract the deformation operators. Next, the use of greedy algorithms, to solve the previously proposed problem, is discussed in Sec. 7.3 and Sec. 7.4.

An *a priori* based greedy algorithm is introduced in Sec. 7.5, Sec. 7.6 and Sec. 7.8 to reduce the instable behavior of MP, due to the very high coherence of the dictionary used. Experimental results are presented in Sec. 7.7.

Considering the investigated representation, a coding scheme is proposed in Sec. 7.9. This section discusses as well the obtained coding results. Another application is considered for the investigated video representation. 3D video primitives may be used as video features for multi-modal audio-visual sequence analysis. A brief introduction to the framework together with some results can be found in Sec. 7.10. Finally, conclusions are drawn in Sec. 7.11.

7.2 Video Approximation: Tracking 2D Image Features Through Time

In this chapter we study how to approximate F_t^γ operators in order to recover spatio-temporal geometric video components. The F_t^γ estimation is done such that the set of functions g_γ^t and g_γ^{t+1}, at time t and $t+1$ respectively, belong to the dictionary $\mathcal{D}$:

$$\forall \gamma,\ \forall t \quad g_\gamma^t \overset{F_t^\gamma}{\longmapsto} g_{\gamma'}^{t+1} \text{ s.t. } g_\gamma^t, g_{\gamma'}^{t+1} \in \mathcal{D}. \tag{7.6}$$

This is imposed for several reasons. The fact that g_γ^t and $g_{\gamma'}^{t+1}$ belong to the same dictionary $\mathcal{D}$ allows the use of the same fast atom search used for image approximations (see Chapter 6 and [61]) to solve F_t^γ. In the case where the parametric description of a sequence is coded, a quantization on the evolution of geometric parameters γ is required. For a better approximation performance, the quantization operation of atoms parameters has to be embedded within the decomposition loop of the greedy algorithm [78]. Quantization is performed such that 3D features slices (i.e. the piece of 3D feature contained in a video frame), belong to the 2D dictionary in use. In effect, if one desires to reconstruct a particular frame from the spatio-temporal geometric video representation, this may be done using always the same dictionary $\mathcal{D}$.

The set of all possible transformations F_t^γ is an approximation of the affine model of local transformations defined for sequences. This approximation intends to supply a trade off between adaptation flexibility and dictionary complexity, i.e. it does not include the model of shearing and is limited by the granularity of the dictionary parameters.

In order to recover a set of 3D primitives to generate sparse approximants for sequences, the transformation of geometrical features must be estimated. According to (7.3), in this work, this is formulated from a frame to frame point of view. Fig. 7.3 schematically describes the approximation of operator F_t in Eq. (7.3). A more practical example can be seen in Fig. 7.4(a) where the approximation of a simple synthetic object by means of a single atom is performed. The first and third row of pictures show the original sequence and the second and fourth rows provide the reconstruction of the approximation. Fig. 7.4(b) shows the parametric representation of the sequence. We see the temporal evolution of the coefficient c_γ^t, and of the other parameters.

The problem that we are facing may be seen as a constrained optimization for signal approximation. The recovery of F_t^γ mappings at each frame is formulated as follows:

$$\min_{F_t} \left\| I_{t+1} - \sum_{\gamma \in \Gamma} F_t^\gamma \left[c_\gamma^t \cdot g_\gamma^t \right] \right\|_2 \quad \text{subject to} \quad \text{Cost}\left(F_t\right) \leq \xi, \tag{7.7}$$

where "Cost" is a constraint depending on the application. This optimization turns out to be very complex and even NP-hard depending on the formulation and the *Cost* measure. However, and depending on the scale of the problem, it may sometimes be possible to find a global optimum.

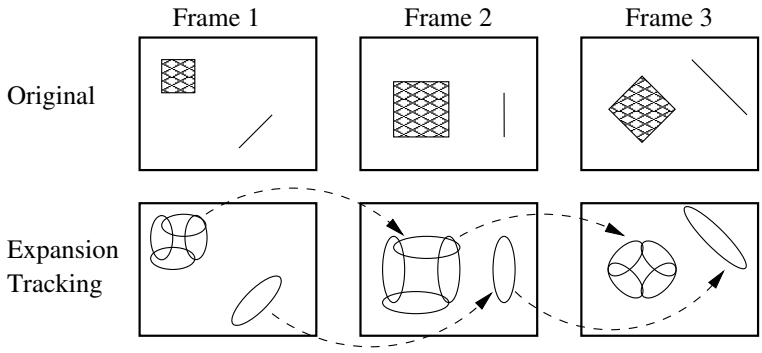
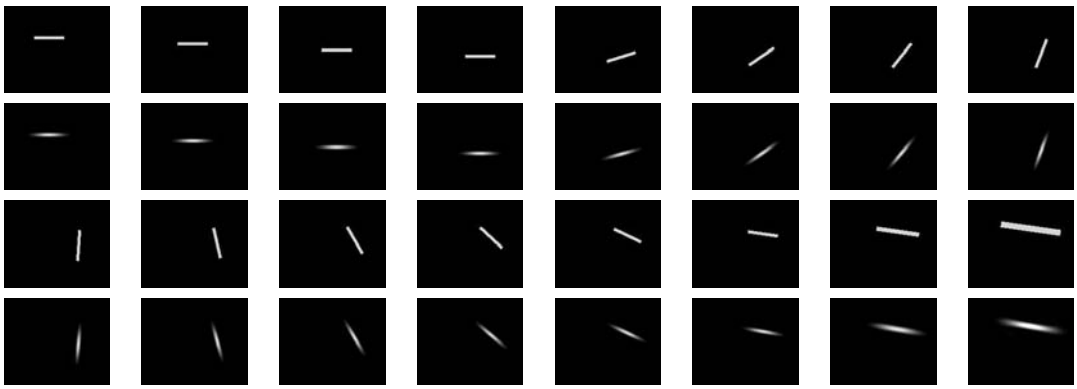
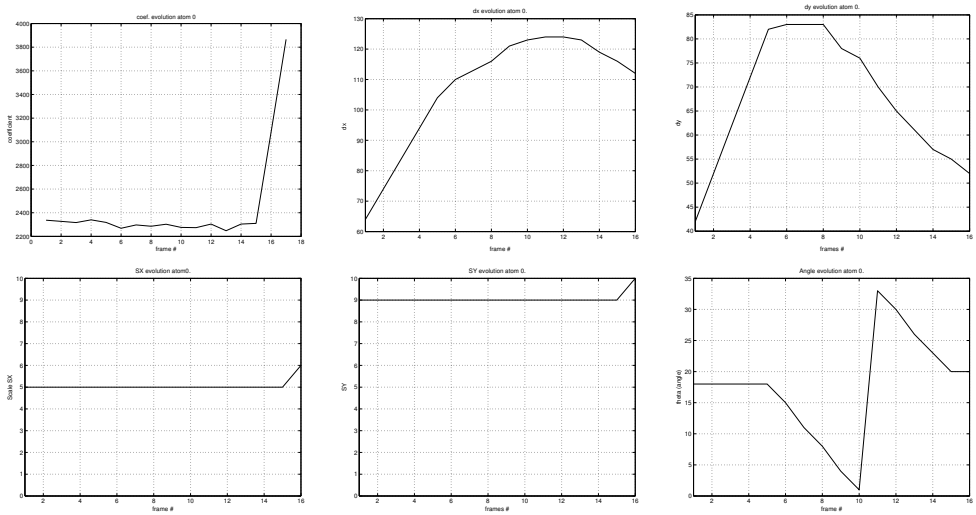


Figure 7.3: Successive schematic updates of basis functions in a sequence of frames. In the second row, ellipses represent schematically the possible positioning of some AR 2D atoms (see Chapter 6).



(a) Synthetic sequence approximated by 1 atom: First and third row show the original sequence made by a simple moving object. Second and fourth row depict the different slices that form a 3D geometric atom.



(b) Parameter evolution of the approximated object; from left to right and from up down, we find: coefficient, x position, y position, x (short axis) scale, y (long axis) scale, angle.

Figure 7.4: Approximation of a synthetic scene by means of a 3D atom.

Example 7.1 *Consider a simplified context where only the selection of functions is involved (thus, coefficients $c \subset \{0,1\}$) and where the costs for every function can be predefined in a fixed manner in the form of a vector ($\mathbf{w}$ where $\forall i\ w_i \geq 0$) where each w_i determines the selection cost of the ith function and the cost constraint can be rewritten as $\mathbf{w}^T\mathbf{c} = Cost$. Thus, Eq. (7.7) could be redefined as:*

$$\min_{c_\gamma{}^{t+1} : \gamma \in \Gamma} \left\| I_{t+1} - \sum_{\gamma \in \Gamma} G_\gamma^{t+1} c_\gamma{}^{t+1} \right\|_2 \quad \textit{subject to} \quad \mathbf{w}^T\mathbf{c} \leq \xi, \tag{7.8}$$

where G_γ^{t+1} is the set of functions in which a given g_γ^t may potentially transform and every selection vector $c_\gamma{}^{t+1}$ a boolean vector such that $\forall\gamma\ 0 \leq \|c_\gamma{}^{t+1}\|_1 \leq 1$. To be clearer, $\mathbf{c}$ is the concatenation of all c_γ. Looking at Eq. (7.8) the problem resembles a lot to the formulations of the retrieval of the best m-term approximation, except for the additional linear constraint on the costs ($\mathbf{w}$), which modifies slightly the constraint norm. More generally, problems of the kind of (7.8) are instances of the Knapsack *problem [127].*

7.3 Tracking Frame Deformations: Using a Greedy Algorithm

In order to obtain a parametric representation in terms of the evolution of geometrical components, a greedy approach is considered for the progressive approximation of every video frame. This approach, very close to that used to obtain sparse approximations of still images, consists on approximating every primitive transformation F_t^γ in a successive manner. However, some further considerations are needed given the assumed motion model and the resulting extremely highly coherent dictionary of functions. Due to dictionary coherence, and the fact that more than one atom is normally necessary to represent signal structure, direct MP full search in a frame at $t+1$ does not have any guarantee to recover the corresponding deformed atoms from frame t. Greedy algorithms are sub-optimal and myopic: they are limited by the resolution and coherence of the dictionary (see Chapter 4). In the case of motion, a simple image deformation and dictionary parameters granularity may induce MP to select the wrong primitive transformation. Moreover, this can contribute to propagate the primitive selection error to posterior MP iterations.

The main aspects to consider when defining the greedy algorithm to retrieve the temporal dimension of 3D geometric video primitives are: limited computational complexity, ensure locality of the algorithm and *smoothness* of motion parameters.

Imposing additional constraints to the selection rule of MP [122] can sometimes be modeled by weakening its greedy nature. As detailed in Sec. 4.4.2, weak greedy algorithms recover, generally, worse approximations. However, as underlined in Chapter 5, if a greedy algorithm is modified in a proper way (e.g. by introducing reliable *a priori* information) their performance may be significantly improved.

7.3.1 Greedy Local Search

The assumption of smoothness on the primitives deformation from frame to frame in our sequence model imposes that the transformation F_γ^t, of a given atom g_γ^t, can not result in any function $g_\gamma \in \mathcal{D}$. As in the case of Block Matching (BM) [183] some constraint on the search space can be set. Solutions beyond the search space are considered to be very improbable. Furthermore, the functional to be optimized is non-convex (7.7). This may yield a slightly better match away from the appropriate place, breaking consequently the structure of the approximation. A local greedy

heuristic is defined by means of a sub-dictionary $\mathcal{D}' \subset \mathcal{D}$ associated to every g_γ^t. The search for the transformation will be performed in $\mathcal{D}'$ solely. We can consider a range of variations $\Delta\gamma$, i.e. in position ($\pm\Delta b_x,\ \pm\Delta b_y$), scale ($\pm\Delta sx,\ \pm\Delta sy$) and angle ($\pm\Delta\theta$):

$$\mathcal{D}'_\gamma = \{g_{\gamma'} : \gamma' \in [\gamma - \Delta\gamma, \gamma + \Delta\gamma]\}\,. \tag{7.9}$$

One may consider the use of quasi-Newton methods, combined with other techniques such as line search or trust-region globalization techniques [34], due to the availability of an analytic expression (Eq. (6.7)) to compute gradients and the non-convex, non-linear form of the problem. However, the complex topology of the objective function makes it likely to fall in local minima.

A local full search based on the computation of all the matching positions will avoid local minima at a reasonable cost if properly implemented [61]. As in the case of static images, the use of a precomputed Fourier version of the different atoms generated from Eq. (6.7), allows the computation of the projection for all atoms translation with a single FFT (see Chapter 6).

7.3.2 Use of Motion Model Constraints: Multi-Objective Optimization

The use of a very redundant dictionary (Chap. 4 and Chap. 6) improves signal modeling but at the expense of a weaker discrimination between atoms. In Chapter 5, we stated that some additional information is needed in order to select a good candidate. As suggested in [120], in a similar approach for motion estimation, the inclusion of *a priori* information in the selection functional may help achieving estimates of frame to frame primitive transformations that are more respectful with the sequence model (see Sec. 7.1). A possible approach can be to impose a regularity constraint among neighboring primitives. Some interdependence is assumed among primitives belonging to the same structure (Sec. 7.5). Besides that, additional motion estimates performed with classic estimation techniques (e.g. region matching techniques) can also be considered [183]. In a coding perspective, however, estimating regularity by means of coding rate could be more appropriate (Sec. 7.8). Representation (and consequently 3D spatio-temporal primitives) is then conditioned by the defined coding scheme.

7.4 Which are the Limits of Using MP?

The use of a greedy strategy implies that only one of the F_t^γ is optimized at every iteration, without taking into account the possible interdependence this might have with the other $F_t^{\gamma'} : \gamma \neq \gamma'$ operators. If F_t^γ are independent $\forall\gamma \in \Gamma$ (i.e. $g_\gamma^{t+1} \perp g_{\gamma'}^{t+1} \quad \forall\gamma \neq \gamma' \quad \gamma, \gamma' \in \Gamma$), each one of them can be optimized independently, leading the algorithm to work perfectly. With no doubt, as discussed previously, given the non-orthogonality of our dictionary, and the fact that those that will be selected will probably be non-orthogonal among them, it is not clear whether an MP like algorithm will succeed in giving a good solution to Eq. (7.7). The present problem is analogous to the recovery of the best m-term sparse approximation problem [95, 174, 176]. A relation between the structure of the dictionary, the signal nature and the algorithm used can be established. Good MP behavior is constrained to the incoherence of dictionaries and enhanced *a priori* models [48, 49].

7.4.1 The Block Structure of the Problem and its Relation with Dictionary Coherence

The dictionary used for predicting of video frames is composed by blocks of candidate functions (as described in Sec. 7.3.1). Each block has been generated by all admissible transformations of a given

primitive from the previous frame. In the approximation of future frames, according to the assumed motion model, only one of these elements will be taken into account, i.e. just an atom from every block. In this situation and in a way similar to [142], some constraints exist that can be exploited in the definition of an upper bound for the good behavior of MP.

Theorem 7.1 *Let $\mu_{1_B}(m)$ be the inter-block coherence defined in [142] (see Appendix C.1) for a block dictionary $\mathcal{D} = \bigcup_l \mathcal{D}_{B_l}$ and let μ_{D_B} be the biggest possible inner product among two different functions into a block. If the signal f is such that*

$$f \in span\left(g_{\gamma_{B_l}} : l \in [0, m-1],\ g_{\gamma_{B_l}} \in \mathcal{D}_{B_l}\right), \tag{7.10}$$

(i.e., f belongs to the space generated by a set of m atoms each of them belonging to a different dictionary block) and

$$\frac{\mu_{D_B} + \mu_{1_B}(m-1)}{1 - \mu_{1_B}(m-1)} < \alpha, \tag{7.11}$$

the Weak(α) algorithm will recover the set of correct atoms that compose f (see Appendix C.1 for a proof).

Thus, for $\alpha = 1$, we need $\mu_{D_B} + 2\mu_{1_B}(m-1) < 1$. This result implies that in order to recover the "correct" atoms, μ_{D_B} and $\mu_{1_B}(m-1)$ can not be *big* at the same time. A very redundant dictionary (μ_{D_B} close to 1) will need very incoherent blocks in order to ensure the right selection of functions. In the other way round, if blocks are coherent enough, μ_{D_B} needs to be sufficiently small to ensure the good recovery of optimal signal components.

Example 7.2 *Motion estimation by means of space domain correlation like block matching methods [116, 183] can be seen as a particular case of this problem. Consider a correlation based approach where all matching candidates for a given image block are normalized and have zero mean. The anchor frame is divided in non-overlapping blocks. Each of these blocks has to be approximated by the most similar block selected from a set of blocks in the reference frame. This set of blocks correspond to all the possible blocks that belong to a neighboring area (see Fig. 7.5). They would correspond to one of the dictionary blocks described above. Furthermore, since anchor frame blocks do not overlap, dictionary blocks are orthogonal. Thus, as long as there are no identical pixmap pieces into a given dictionary block (i.e., $\mu_{D_B} < 1$), the recovery of the optimal anchor frame expansion is ensured by Theorem 7.1.*

7.5 Using Regularity Constraints: A Bayesian Approach of the Problem

The main factors that may have an influence in recovering "correct" F_t^γ operators have been discussed in the previous section. They directly influence the capacity of appropriately modeling the video sequence and can be summarized as:

- Matching Pursuit does not work as one would like with coherent dictionaries.
- $\mathcal{D}$ is made of a finite set of atoms, which implies that only a limited set of positions, scales and rotations are available.
- Motion is assumed to be uniform over the support of an atom.

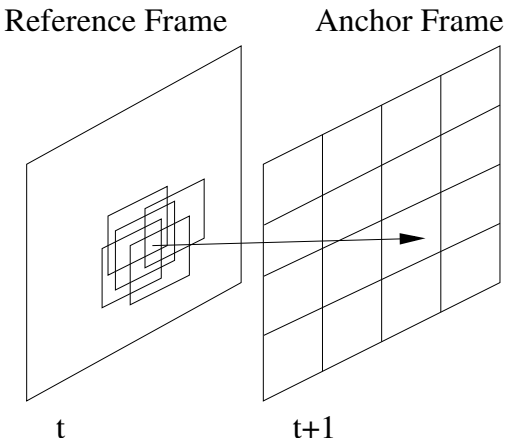


Figure 7.5: BM using a fixed block size anchor frame. Each set of candidate blocks to match into a block of the anchor frame are, according to Sec. 7.4.1, an orthogonal dictionary block.

The inclusion of an *a priori* model in the greedy selection criteria is thus necessary to reduce instability on the recovery of 3D primitives. A first solution is to perform a local search on a reduced subspace. However more complex models and *a priori* informations can be taken into account. In this section we show how Bayesian modeling can be used to tackle that problem.

7.5.1 Probability Model to Optimize

We reformulate the greedy selection criteria from the probabilistic point of view presented in Chapter 5. As explained there, taking the strongest scalar product consists in selecting the most probable atom. However, much more involved models can be considered. A Bayesian modeling of the problem can be performed if some *a priori* information or knowledge about the parametric sequence description is available. We make the assumption that neighboring 3D atoms present similar temporal deformations. This regularity can be inserted in the optimization problem by means of the *Cost* term in (Eq. (7.7)). In the greedy formulation, a Bayesian functional that maximizes the Maximum a Posteriori (MAP) probability will integrate the regularity assumption. We consider a Markov Random Field (MRF) framework to define probabilistic relations among atoms.

Thus, for every MP iteration we optimize:

$$\begin{aligned} \Delta\gamma_n &= \arg\max_{\Delta\gamma_n} \left\{ p\left(\Delta\gamma_n, \Delta c_n \mid \mathcal{R}_n^{t+1} f, g_{\gamma_n}^t\right) \right\} \\ &= \arg\max_{\Delta\gamma_n} \left\{ p\left(\mathcal{R}_n^{t+1} f, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right) \cdot p\left(\Delta\gamma_n, \Delta c_n\right) \right\}, \end{aligned} \tag{7.12}$$

such that $\Delta\gamma_n$ represents the parameter differences between $\gamma_n^{t+1} \in \Gamma$ and $\gamma_n^t \in \Gamma$, and $\mathcal{R}_n^{t+1}$ is the nth iteration frame residual at time $t+1$. In Eq. (7.12) the most probable transformation is taken given a residual at $t+1$ and the corresponding g_γ at time t for a given greedy step n. By the Bayes' rule, this is equivalent to maximizing the probability of matching a given $g_{\gamma_n}^t$ with the residual $\mathcal{R}_n^{t+1}$ conditioned to the probability of the transformation $\Delta\gamma_n$ and the temporal change on the projection coefficient Δc_n. The matching probability $p\left(\mathcal{R}_n^{t+1} f, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right)$ can be defined as a function of an estimated residual error energy $\left\| \hat{\mathcal{R}}_{n+1}^{t+1} f \right\|^2$ for the retrieval of function g_{γ_n} at iteration n. Atoms are assumed to deform under consistent motion transformation. Thus, no change in the coefficient

will be considered (except for scale changes) in the estimation of the most probable motion:

$$\hat{\mathcal{R}}_{n+1}^{t+1} f = \mathcal{R}_n f^{t+1} - \overline{\langle \mathcal{R}_n^t f, g_{\gamma_n}^t \rangle} g_{\gamma_n}^{t+1}, \tag{7.13}$$

where $\overline{\langle \mathcal{R}_n^t f, g_{\gamma_n}^t \rangle}$ is normalized according to a possible re-scaling of $g_{\gamma_n}^{t+1}$ with respect to $g_{\gamma_n}^t$, i.e $\overline{\langle \mathcal{R}_n^t f, g_{\gamma_n}^t \rangle} = \langle \mathcal{R}_n^t f, g_{\gamma_n}^t \rangle \sqrt{\Delta sx \Delta sy}$. At time t, $\mathcal{R}_{n+1}^t f \perp g_{\gamma_n}^t$, in order to minimize the energy of the projection error. In the same way, after motion transformation, $g_{\gamma_n}^{t+1}$ should be such that $\left\| \hat{\mathcal{R}}_{n+1}^{t+1} \right\|^2$ is also minimized.

The probability measure assumes the Gaussianity (by the central limit theorem [138]) and independence of error samples $\mathcal{R}_{n+1}^t f(x,y)$ (although this is not often the case for this class of signals). Based on previous approaches of the block matching and MRF fields [9, 149], we consider:

$$p\left(\mathcal{R}_n^{t+1} f, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right) = \frac{1}{Z} \prod_{x,y} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{|\hat{\mathcal{R}}_{n+1}^{t+1} f(x,y)|^2}{2\sigma^2} \right) \tag{7.14}$$

where Z is a normalizing constant and $\sigma^2 \approx E\left[\left| \hat{\mathcal{R}}_{n+1}^{t+1} f(x,y) \right|^2 \right]$. Note that $\hat{\mathcal{R}}_{n+1}^{t+1} f$ is considered to have zero mean. In fact, prior to any operation, a low pass approximation is removed from every frame (see Sec. 7.6). Introducing the evaluation of σ^2 in Eq. (7.14) we obtain the conditioned optimization criteria:

$$p\left(\mathcal{R}_n^{t+1} f, g_{\gamma_n}^t \mid \Delta\gamma_n, \Delta c_n\right) \approx \frac{C_1}{\sqrt{\left\| \hat{\mathcal{R}}_{n+1}^{t+1} \right\|^2}}, \tag{7.15}$$

where C_1 is a constant.

The probability $p\left(\Delta\gamma_n, \Delta c_n\right)$ imposes the model that drives the transformation F_t^γ of the g_γ^t and the associated coefficient. It is thus defined as the conditioned probabilities of the $\Delta\gamma$ and Δc_n in the framework of MRFs. At every iteration, MP will try to select a new atom that maintains regularity with all previously selected primitives in the neighborhood. Earlier atoms are trusted to generate the MRF for the future appearing atoms. This unbalanced criteria derives from the fact that first atoms of the MP decomposition capture more energy, thus they tend to represent much more significant (i.e. reliable) features from the signal. The inheritance of atoms deformations, through MP iterations, may rather suggest to the reader a Markov Chain (or tree) structure of the problem (instead of a MRF structure). In any case, and without loss of generality, one can assume that at a given stage of the algorithm a MRF is available as a product of this Markov Chain.

When no first reliable estimate of $p\left(\Delta\gamma_n\right)$ is available, an initial tentative needs to be performed trying to match the whole region where the atom is supported. Atoms interaction, MP sub-optimality and simplicity of atoms waveform may reduce the reliability of the estimation provided by a single atom matching. This will be explained later in Sec. 7.6.

We can formulate $p\left(\Delta\gamma_n, \Delta c_n\right)$ as:

$$p\left(\Delta\gamma_n, \Delta c_n\right) = p\left(\Delta c_n \mid \Delta\mathbf{d}_n, \Delta\mathbf{s}_n, \Delta\theta_n\right) \cdot p\left(\Delta\mathbf{d}_n, \Delta\mathbf{s}_n, \Delta\theta_n\right), \tag{7.16}$$

where Δc_n (temporal variation of the nth atom scalar product with the residual) depends on the choice of the new γ parameters. Considering $\Delta\mathbf{d}$, $\Delta\mathbf{s}$, $\Delta\theta$ independent, Eq. (7.16) turns into:

$$p\left(\Delta\gamma_n, \Delta c_n\right) = p\left(\Delta c_n \mid \Delta\mathbf{d}_n, \Delta\mathbf{s}_n, \Delta\theta_n\right) \cdot p\left(\Delta\mathbf{d}_n\right) \cdot p\left(\Delta\mathbf{s}_n\right) \cdot p\left(\Delta\theta_n\right). \tag{7.17}$$

Each of the probability functions has the form of a MRF. That is, they may be modeled by a Gibbs distribution [114]:

$$p(x) = \frac{1}{Z_x} \exp\left(-\frac{E_x\left(x\right)}{T_x} \right), \tag{7.18}$$

where $E_x(x)$ is an energy function that characterizes the MRF and how neighboring variables are related, while T_x stands for its variance.

From Eqs. (7.12), (7.15), (7.17) and (7.18) the functional to be optimized can be expressed as:

$$\begin{aligned}\Delta\gamma_n = \quad & \arg\min_{\Delta\gamma_n}\left\{\frac{1}{2}\log\left(\left\|\hat{\mathcal{R}}_{n+1}^{t+1}\right\|^2\right) + \lambda_{\Delta c_n}E_{\Delta c_n}(\Delta c_n) + \right. \\ & \left.\lambda_{\Delta\mathbf{d}_n}E_{\Delta\mathbf{d}_n}(\Delta\mathbf{d}_n) + \lambda_{\Delta\mathbf{s}_n}E_{\Delta\mathbf{s}_n}(\Delta\mathbf{s}_n) + \lambda_{\Delta\theta_n}E_{\Delta\theta_n}(\Delta\theta_n)\right\}\end{aligned} \tag{7.19}$$

where $\Delta\gamma_n = \{\Delta\mathbf{d}_n, \Delta\mathbf{s}_n, \Delta\theta_n\}$ and each λ_x is a function of the statistics parameter T_x in Eq. (7.18).

7.5.2 Regularity Models

The general regularized expression to solve at every greedy iteration (7.19), requires the definition and modeling of each regularizing term E_x. In the following, the definitions of the Gibbs distributions arising in the MAP estimation are described together with the parametric modeling of the MRF.

Coefficient Model

Temporal variations of coefficients Δc_n should be small in ideal tracking of a primitive. In any case, coefficients may not change sign. Changes to coefficients should be driven mainly by the change of scale of the approximating function.

To induce its temporal regularity, a normalized quadratic distance between the coefficients at time t and $t+1$ is considered for $E_{\Delta c_n}(\Delta c_n)$:

$$E_{\Delta c_n}(\Delta c_n) = \left(\frac{c_n^{t+1} - c_n^t \cdot \sqrt{\Delta sx \Delta sy}}{c_n^t \cdot \sqrt{\Delta sx \Delta sy}}\right)^2, \tag{7.20}$$

where previous c_n^t are re-normalized with respect to the scale transformation. One can observe that Eq. (7.20) is normalized in order to be independent of n.

Geometric Models

Displacement, change of scale and rotation constraints, are measured as the euclidean distance between the value under test and the most likely (ML) transformation estimated from previous MP iterations at every image location. Hence they can be represented as:

$$\begin{aligned}E_{\Delta\mathbf{d}_n} &= \left(d_x^n - \hat{d_x}^n\right)^2 + \left(d_y^n - \hat{d_y}^n\right)^2 \\ E_{\Delta\mathbf{s}_n} &= \left(s_x^n - \hat{s_x}^n\right)^2 + \left(s_y^n - \hat{s_y}^n\right)^2 \\ E_{\Delta\theta_n} &= \left(\theta^n - \hat{\theta}^n\right)^2,\end{aligned} \tag{7.21}$$

where $\hat{d}$, $\hat{s}$ and $\hat{\theta}$ correspond to the ML estimates (see Sec. 7.5.4 for details about their calculation). The use of motion information from the first appearing atoms to regularize the selection criteria of new ones can be seen as a way to propagate the motion information from more reliable atoms to less reliable ones.

7.5.3 Setting the Motion Model

Eqs. (7.20) and (7.21) define the potential among variables of the functional to optimize. However the model lays on the values assigned to the λ_x of Eq. (7.19). These values are unknown a priori and depend on the data to be analyzed since they represent the statistics that characterize the random variables Δc_n, $\Delta \mathbf{d}_n$, $\Delta \mathbf{s}_n$, $\Delta\theta_n$. In this work they have been considered to be constant for the whole sequence. Their value, as defined by Eqs. (7.18)-(7.21) is proportional to the standard deviation of the variables implied in the energy functionals described before. Hence, for a general sequence, their value needs to be trained. However, λ_x values will just be valid in average for the real transformation given the heterogeneous nature of motion in a general sequence. A detailed analysis of the proper adaption of a statistical model is out of the scope of this work. We focus on the understanding of the use of greedy approaches and parametric over-complete dictionaries for the approximation of image sequences.

7.5.4 Motion and Probability Fields Estimation

The transformation estimates are computed from all the atoms that interact in a certain region. In the example presented in this work (Eq. (6.7)) atoms have a localized support in space. Even though it is not strictly finite (see Fig. 7.7), amplitude decay is fast enough such that atoms located sufficiently far away can be considered as not interacting. Furthermore, the decay of the Gaussian envelop of (6.7) can be considered as well as an indicator that the strength of constraints (7.20) and (7.21) has to increase the closer an atom is from another, i.e. it is logical to consider that two such atoms must have a more coherent motion.

λ_x Modeling

From Eq. (6.7), the atom envelop is a bi-variate Gaussian with the same dimensions (sx, sy) as the atom itself:

$$\begin{aligned} p_\gamma(u,v) &= K \exp\left(-\left(u^2+v^2\right)\right) \quad \text{s.t.} \\ u &= \frac{\cos\theta\,(x-dx)+\sin\theta\,(y-dy)}{sx} \\ v &= \frac{-\sin\theta\,(x-dx)+\cos\theta\,(y-dy)}{sy}, \end{aligned} \tag{7.22}$$

where K is a constant. This model is assumed to represent the influence law of the transformation of a given atom in a neighborhood. Thus, $\forall x$, λ_x depend on the spatial location and are proportional to $p_\gamma(u,v)$. That is, the variance of the probabilities described in Sec. 7.5.1, depends on the spatial location and decreases as a function of scale and the distance with respect to the center of an atom. Indeed, the model defined by lambdas in Section 7.5.3 must have a local influence related to the structure of the signal. λ_x values model tightness in atom interactions. The formulation of every λ_x is that of a constant (tuned in order to fit the deformation model) multiplied by the bivariate model of (7.22):

$$\lambda_x(x,y) = C_{\lambda_x} \cdot p_\gamma(u,v).$$

As one can observe, the value of λ_x depends on the area of influence of each atom g_γ. In an area where more than one atom overlap, for ever particular spatial location, the value of the atom having the highest $p_\gamma(u,v)$ will be considered.

Motion Parameter Estimates

The motion parameter estimates $\hat{d_x}, \hat{d_y}, \hat{s_x}, \hat{s_y}, \hat{\theta}$ of Eq. (7.21) are estimated from the preceding $n-1$ atoms of a frame expansion. They are the maximum likelihood estimates according to the energy probability associated to each atom.

In fact, considering that a given frame energy can be represented as the sum of the square of the coefficients in a MP expansion:

$$\|I_{t+1}\|^2 = \sum_{n=0}^{\infty} |c_n|^2 , \tag{7.23}$$

we may approximate the probability associated with the nth atom as a fraction of $\|I\|^2$

$$p(\gamma_n) = \frac{|c_n|^2}{\|I\|^2}. \tag{7.24}$$

The conditioned probability that a given AR atom contributes to spatial location (x, y) can be modeled through Eq. (7.21). Thus,

$$p(x, y \mid \gamma_n) = \frac{K}{\sqrt{sx \cdot sy}} \exp\big(-\big(u(x,y)^2 + v(x,y)^2\big)\big). \tag{7.25}$$

Hence, the motion parameters induced by atom $g^t_{\gamma_n}$ at point (x, y) have probability:

$$p(\gamma_n \mid x, y) = \frac{p(x, y \mid \gamma_n)p(\gamma_n)}{\sum_n p(x, y \mid \gamma_n)p(\gamma_n)}. \tag{7.26}$$

For localized atoms in space, we can see that the summation in the above equation will only integrate those atoms close to position (x, y) (i.e. due to their amplitude decay -Eq. (7.22)-). Giving as example the case of the most likely displacement, or translational motion $E\{\mathbf{d} \mid x, y\}$ at a given (x, y), we formulate it as the average of all the transformations induced by all the atoms at a given point:

$$\hat{\mathbf{d}} = E\{\mathbf{d} \mid x, y\} = \sum_n \hat{\mathbf{d}}(x, y)_n \cdot p(\gamma_n \mid x, y). \tag{7.27}$$

The same applies to the remaining geometrical parameters $\hat{\mathbf{s}}, \hat{\theta}$.

In Eqs. (7.23) - (7.26) the whole set of terms for the expansion of I are considered. However in a practical and realistic situation, only a truncated version of the representation can be considered. Indeed, to estimate (predict) the motion that the nth atom will follow, only the precedent $(n-1)$ available atoms are considered for the statistical measurements and calculations.

7.6 Implementation Issues

7.6.1 Signal Representation

3D spatio-temporal features are extracted in a frame by frame fashion. At each particular frame, the strategy employed in Chapter 6 to approximate images is adopted. This means that the dictionary used to represent 2D geometry is composed by the AR set of functions of section 6.2.2 and low frequencies are condensed by means of a Laplacian pyramid in a highly downsampled subband (Fig. 6.1). Geometric changes from frame to frame of these 2D AR atoms form the third dimension of spatio-temporal atoms.

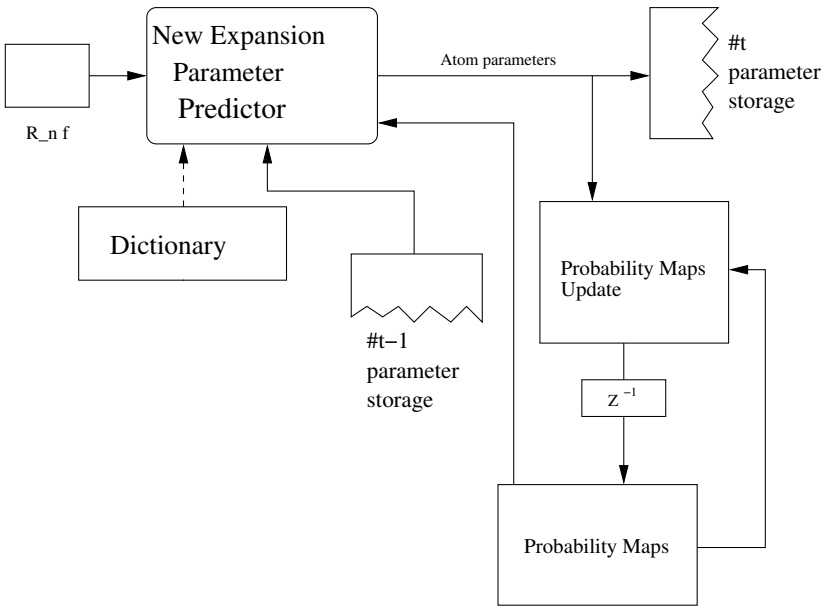


Figure 7.6: Expansion Block Scheme.

7.6.2 Atom Refresh

All the information appearing in a frame at time t can not be mapped from the previous frame and vice-versa. Indeed, we consider a forward mapping scheme where all atoms from frame at time $t-1$ try to get matched in the frame at time t. This is not always possible and sometimes the atom will not be able to find at t the feature it was representing at time $t-1$. In the present approach we consider a measure of the reliability of the prediction of a given atom evolution. At every new frame the normalized scalar product of the transformed atom is compared with the initial projection of the atom within the first frame.

$$\left| \frac{\left\| c_n^{t+1} \right\|^2}{s x_n^{t+1} s y_n^{t+1}} \right| \geq \frac{\|c_n^0\|^2}{s x_n^0 s y_n^0} \cdot \delta, \qquad \delta \in (0,1] \tag{7.28}$$

If a significant drop in the scalar product is detected the atom is canceled (the trajectory is not valid anymore). At the end of the projection process, those atoms that have been canceled are reintroduced in the frame through a full MP search. In the investigation performed in this work, the atom refresh has been set such that a very small portion of atoms can be renewed at every frame (e.g. no more than three percent).

7.6.3 Motion Initialization

The functions in use for the generation of our dictionary have a relatively simple shape. In the direction parallel to contour gradients, very likely represented by the smooth part (Gaussian) of Eq. (6.7), even very relevant atoms may slide: this is similar to the well know "aperture" problem. To avoid this, a first initialization of the expected motion maps is essential. Thus, in the case of no *a priori* indicator of the motion of a primitive, the whole pixmap of the original image included in the support of that primitive is used for a first estimate. The cross-correlation (matching) of the zero mean and normalized versions of the patch and the frame that we want to approximate is used,

i.e. the correlation between the normalized patch and the normalized frame is measured for every possible geometric transformation of the atom.

7.6.4 Motion Maps Update

A set of geometrical parameter maps are kept during the iterative decomposition of a frame. These contain the local geometrical deformations that atoms suffer in their adaption to represent a new frame. Geometry maps are updated progressively at each iteration of the greedy algorithm. After the retrieval of a function, its transformation parameters are introduced in the maps as described in Sec. 7.5.4. The information of the maps is used to introduce regularity in the selection procedure at every greedy iteration. In this way, the motion registered by the first atoms found in a certain image area will condition the geometrical deformation of posterior atoms found in that area. In Fig. 7.7 we show a representation of an atom transformation together with the associated motion. We show as well the influence area where the parameter maps will be considered. At the bottom, we show the conditioned probability (an-isotropic Gaussian) that will take part in the computation of the most likely local motion transformation given an image location.

7.7 Experimental Results

In this section we present several results corresponding to the effect of regularization on different sets of sequences, both synthetic and natural. The results shown as vector fields correspond not only to the translation of atoms but also to their deformation, i.e. the interpolated motion of atoms is represented on their whole support (see Fig. 7.7). These representations (Figs. 7.8, 7.9, 7.11, 7.12), although they may suggest an optical flow meaning, must be interpreted more in the sense of atoms deformation flow. Some examples of transformation are given for particular atoms (Figs. 7.9 and 7.12). The effects of regularization on coding performance are presented later in Sec. 7.9.

7.7.1 Synthetic Sequence Examples

In Fig. 7.8 we show an illustration of the proposed paradigm based on steering image primitives through a sequence. In this test, like in all the rest, the dictionary in use is the one proposed in Sec. 6.2.2. The flow represented in the third row shows how atoms transform to follow and match the successive transformations of the sequence. Just above in the figure, the resulting approximations and the residuals show that although primitives adapt better to shape and motion trajectory, they are subject to the lack of resolution of the dictionary. The effects can be seen in the evolution of the residual error after the approximation.

In fact, in the absence of regularity constraints, atoms try to reorganize themselves in order to reduce this residual error. Consequently, this would force many terms of the successive frames expansions to reorganize in position, angle and scale, leading to a irregular representation of the motion transformation. As part of the approximation error, a progressive blurring of the rod edges can be observed. This corresponds to the adaption of function coefficients. They try to reduce the negative effect of slight motion parameters mismatch, together with the sub-optimality of the greedy approach. Hence, a cumulation of prediction error appears through time.

Continuing with another synthetic sequence, we can see this time in Fig. 7.9 the example corresponding to the motion associated to a particular atom. The sequence corresponds to a translating and rotating square. We consider a certain atom, represented in the picture by a white mark that has the shape of its support. In both columns, we see the representation of the square by means of a 50 coefficients expansion with the footprint of the function support superimposed. In the left

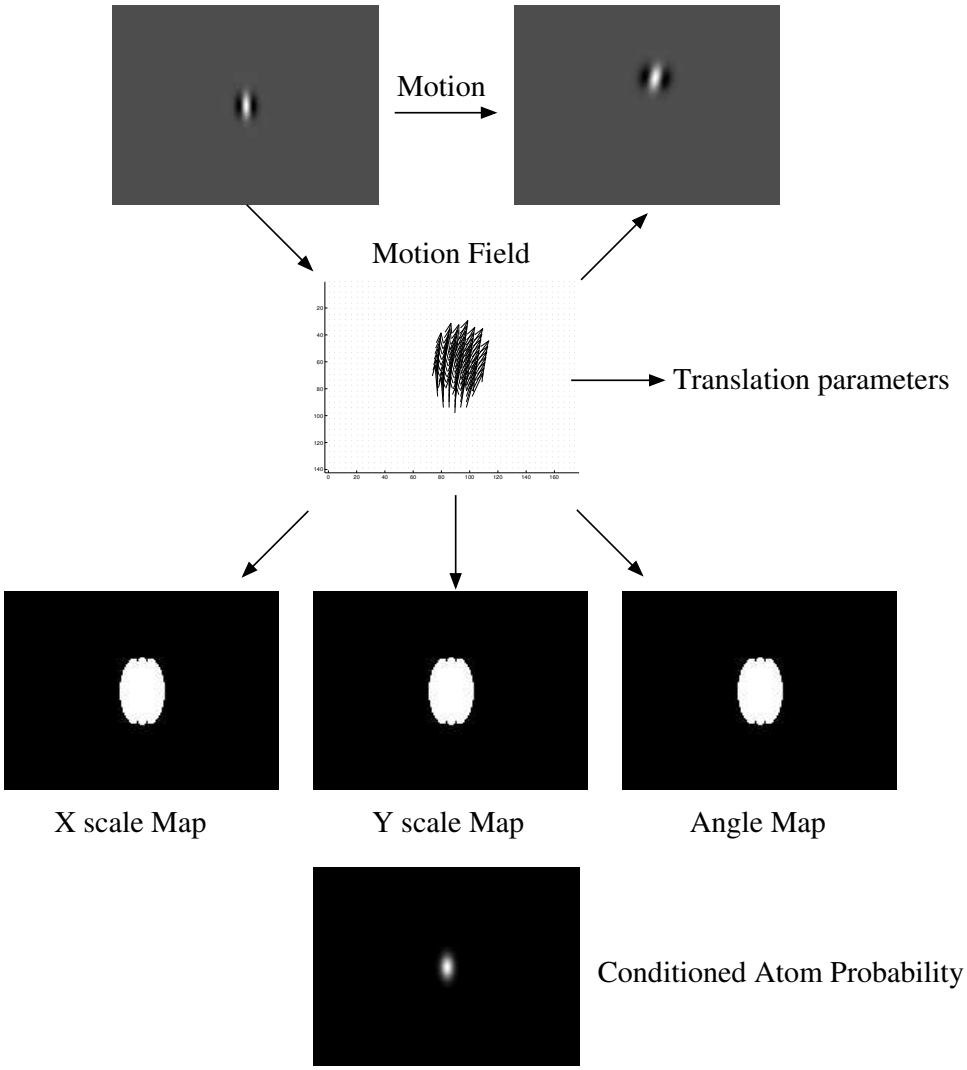


Figure 7.7: Atom transformation maps and parameters. The parameter maps (X scale, Y scale and Angle) correspond to the areas where the geometry of a selected atom will influence future greedy iterations.

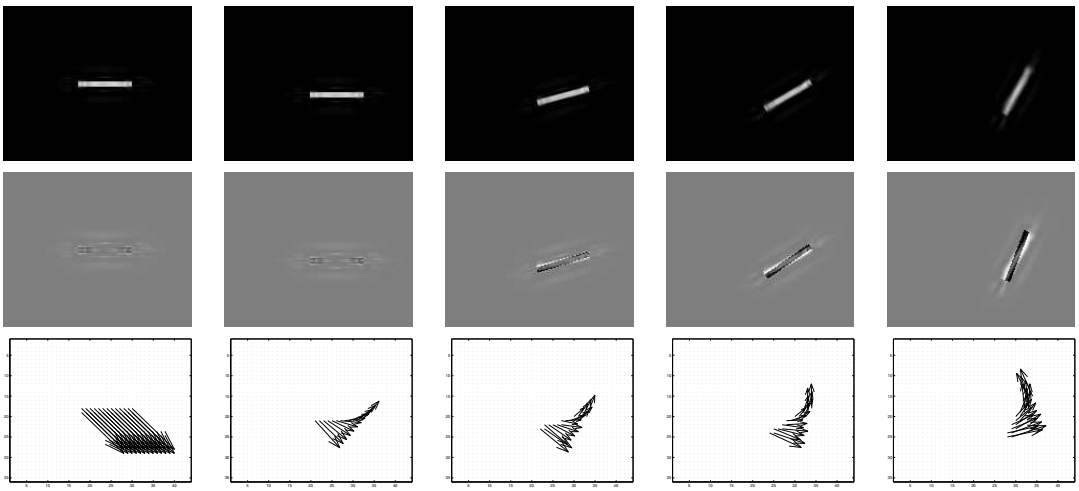


Figure 7.8: Affine motion of a synthetic model (line). From top to bottom: approximation of the line, residual with respect to the original model and motion associated to the atoms. In the second row, we clearly see the effect of parameter quantization, in this case error is induced by the limited resolution in translations and rotations.

column we display the corresponding past and present positions of the atom for the non regularized case, i.e. the selected atom is fully driven by the search of the highest projection coefficient absolute value. On the right, the atom is steered considering the a priori of rigid motion. At the bottom of Fig. 7.9 we can see the motion associated to atoms of the right upper column.

7.7.2 Natural Scene Examples

The synthetic models considered in the figures above are very simple. Fig. 7.10 shows a comparison between a non regularized result (left), and a regularized one (right). A clear influence of the regularization and motion initialization is reflected in the flow related to atoms motion. Figures on the right show that a clear relation can be established between atoms that participate in the cars approximation and their motion. In the example where the truck appears, the influence between neighboring atoms located in the wood area in the background can not be avoided, i.e. the moving atoms of the truck *push* in some measure the atoms representing the background. Interdependence among neighboring primitives is responsible for their strong interaction.

In Fig. 7.11, a set of consecutive approximated frames appears together with the motion flow. Notice how the regularized motion of atoms follow the object trajectories. However, interactions among neighboring atoms are observed, uncovered and covered parts in some situations may enter in interaction. Indeed, shrinking, dilation or slight displacement can manifest due to MP sub-optimality, dictionary granularity and the interaction among atoms.

In order to illustrate the behavior of atoms in a natural sequence, we show in Fig. 7.12 a set of pictures that represent from top to down: the original set of images, the approximation using 500 atoms, the flow of atoms, and the evolution of 3 different atoms through time. The atoms flow shows the motion of image primitives, as well as their deformation. Changes suffered by those representing the background to adapt to the motion of the head are clearly appreciated. This is because several big atoms are used to describe the background. Certainly, the building appearing behind the head can be represented very efficiently with long oriented atoms with a large scale in the parallel direction

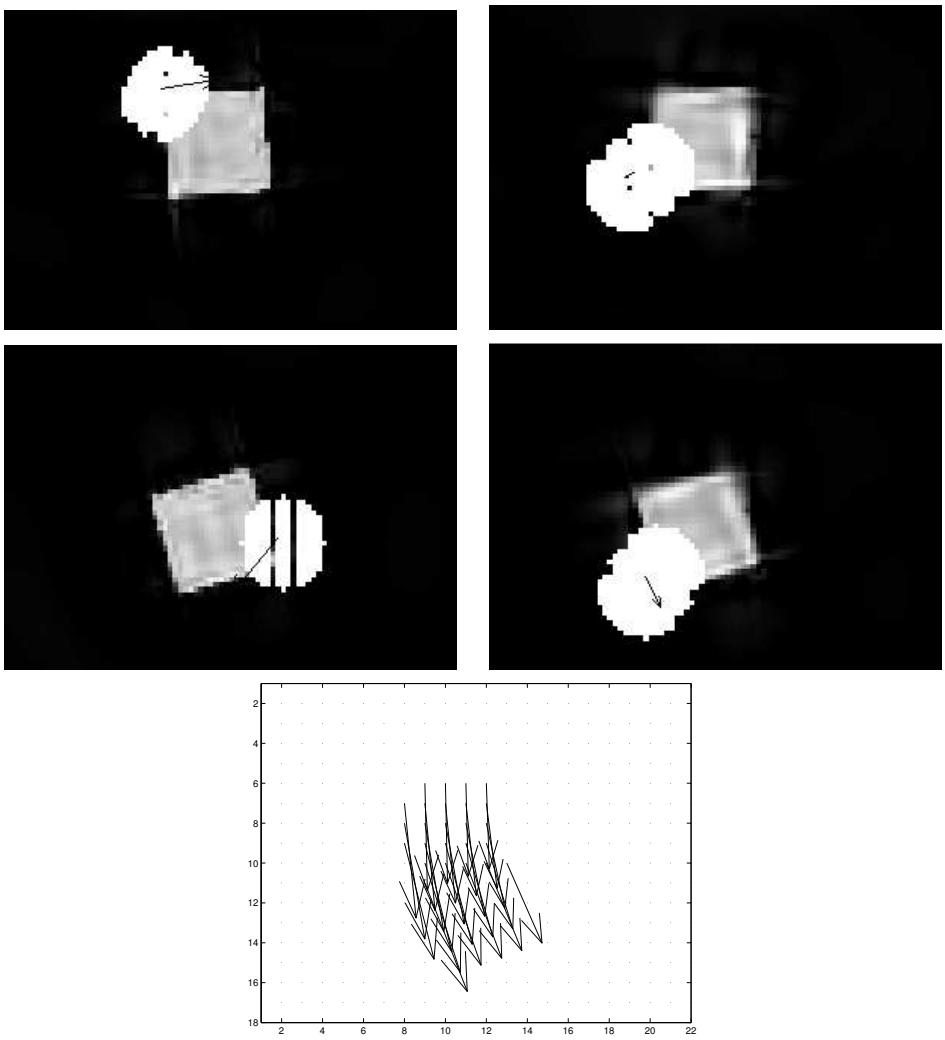


Figure 7.9: Affine motion of a synthetic model (square). The white batch corresponds to the footprint of a selected atom in two temporal instants. Left is the non-regularized prediction. Right is the regularized prediction. Bottom: most reliable motion of the regularized solution (atoms flow in the area where atoms amplitude is significant). Rotation and displacement can be appreciated.

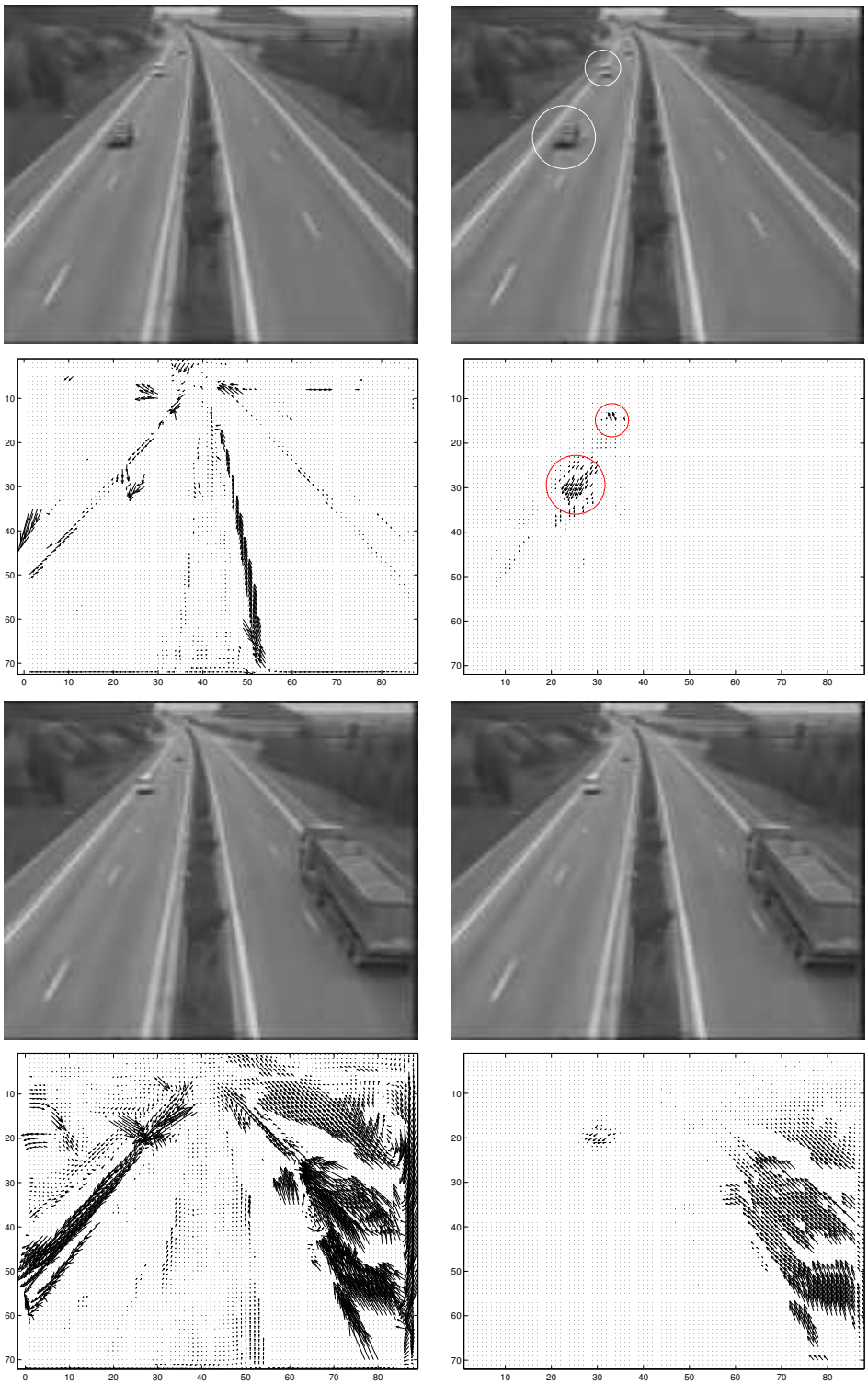


Figure 7.10: Natural sequence motorway. Left column: non-regularized solution. Right column: regularized tracking. First and third rows: Respective reconstructions with 500 atoms. Second and fourth rows: Most reliable primitives motion.

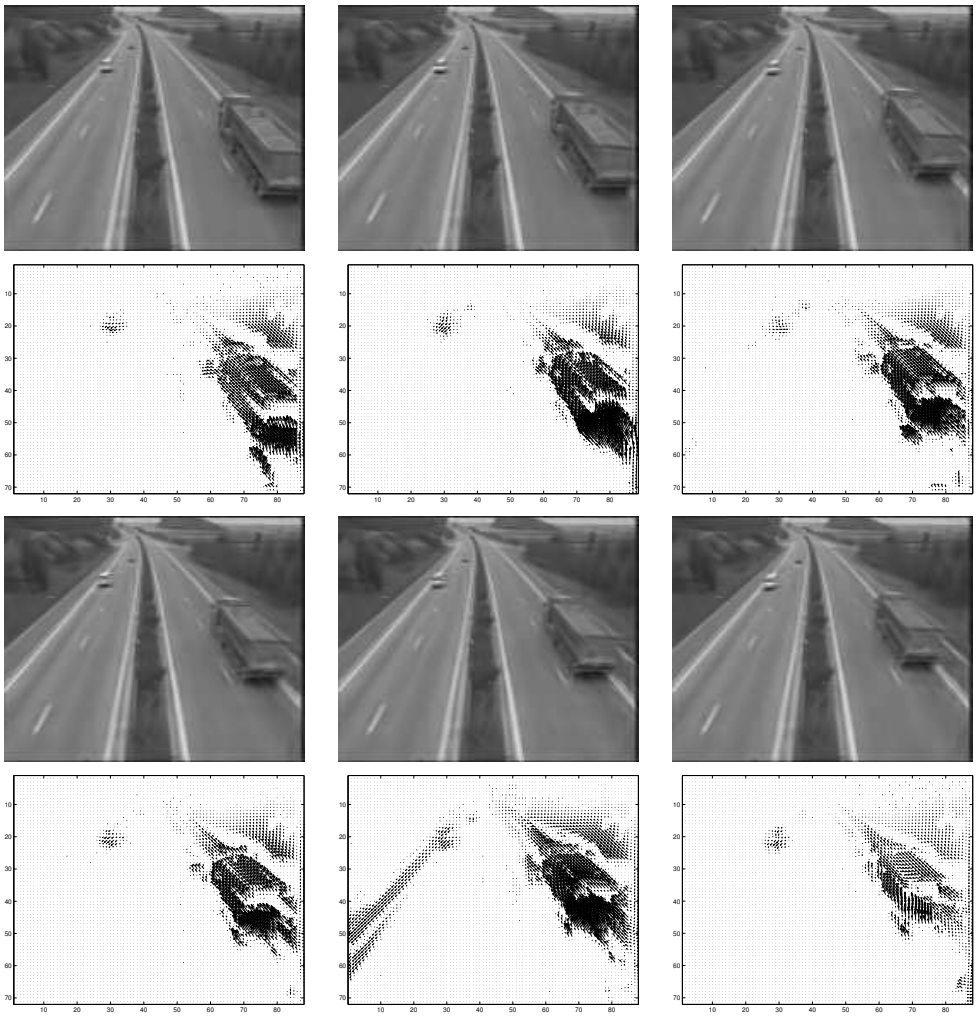


Figure 7.11: Several consecutive frames of a natural sequence showing the reconstructed signal with 500 coefficients together with the deformation suffered by atoms. The transformation of atoms from frame to frame was done using the criteria with a priori information.

to the lines. These have to re-adapt their parameters to fit as well the motion of the head. The three atom examples appearing in the last three rows are composed by the approximation pictures plus the footprint of the represented atom. They capture the motion of the head, and translation, rotation and scaling can be appreciated (Fig. 7.12). We must notice that atoms reintroduced by means of a refresh are not taken into account in the flow representation.

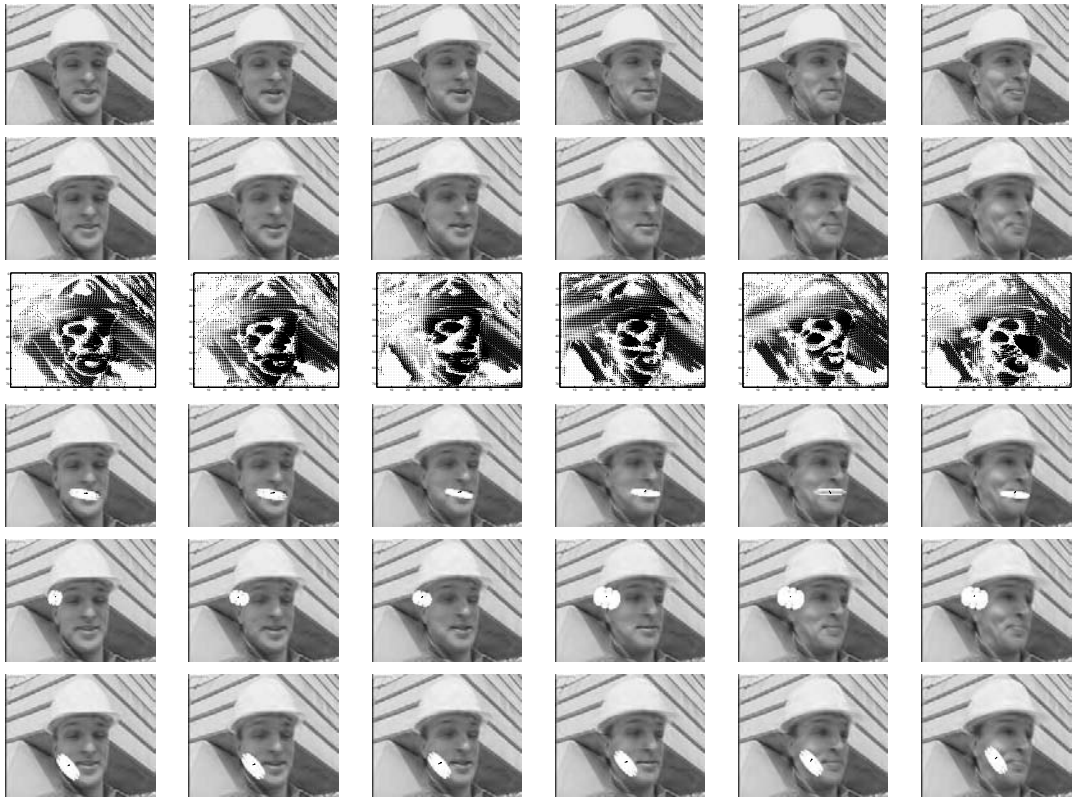


Figure 7.12: Several consecutive frames of a natural sequence showing the reconstructed signal with 500 coefficients with their associated motion. First row: the original frame; Second row: the reconstructed approximations; Third row: the deformation flow; Forth to Sixth: motion of 3 different atoms from the sequence. Their temporal evolution is indicated by the changes on the white footprint.

Evidence of the regularization effects in the sequence *foreman* can be found in Fig. 7.13. Atoms deformation becomes less instable with the regularization. Notice how important this effect is on the region where the detail of the lines of the building are. In the absence of regularization and motion initialization atoms motion is not accurate in their smooth direction. Furthermore, MP facilitates the propagation of error to neighboring atoms in the area. Motion initialization by means of matching is a key element for stability.

Weighted-MP acts as if the dictionary in use was modified depending on the spatial location and signal structure. At every atom search, approximation fidelity is constrained by the regularity model in use. Chapter 5 clearly states the limitations in approximation capabilities of Weighted-MP. In the present examples, a price is paid for the regularization of atom transformations. A cumulative loss in approximation is appreciated under the form of a progressive temporal drift in the prediction of frames (Fig. 7.14). The present loss can be explained in terms of the theoretic findings of Chapter 5. The convergence bound for Weighted-MP (see Theorem 5.5) appears to be

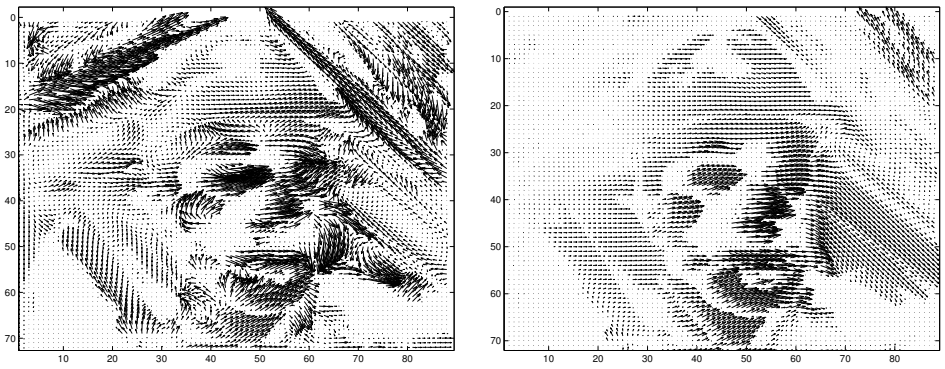


Figure 7.13: Comparison of the computed deformations (atoms associated motion) for the 2nd frame of the foreman sequence: left not regularized, right regularized.

better when a reliable *a priori* model is used, unlike in the case of pure MP. Other factors may degrade, though, approximation performance. These are appear in Eq. (5.15) and can be easily related to the different factors affecting the performance shown in the present examples. Factors affecting performance are related with Eq (5.15) as follows:

- *A priori* suitability is represented in Eq. (5.5) by ϵ_{max} and w_{Γ}^{max} terms. One can not really judge whether our regularization model is very reliable or not with the present visual results. Later, we will see in the coding results section that, at least, the model adopted can not be too bad (but not necessarily the best one). Indeed, the use of the regularity model within Matching Pursuit succeeds in exploiting signal redundancy leading to significantly better R-D results (Sec. 7.9). Hence, one may deduce that, at least in average, the assumed model somehow fits with the data. Nevertheless, if only approximation error is considered as measure, the present model may not be the most appropriate, but still is able to extract signal structure (see Sec. 7.10 results, where video decomposition is used for multi-modal analysis).
- Term η of Eq. (5.5) has a particular relevance in here. In an optimal case, the value of η would be zero. We remind that η is a penalizing factor in charge of mathematically indicating how close, from the best m-term approximation, a feasible approximation can be. The relation with the present examples comes from the parameter sampling of the dictionary. Indeed, the previous assumed quantization for translations, rotations and scaling of atoms, imposes a huge limitation in representing deformations. This limitation avoids recovering structures motion with full precision. Moreover, there is a strong assumption on the constancy of the motion field all over the spatial support of each one of the geometric atoms. The direct consequence is that approximation sparsity is reduced and, finally, a temporal drift is introduced (i.e. 3D video structures are estimated in a forward predictive fashion).

7.8 Rate-Distortion Formulation

A similar framework to Sec. 7.5 can be considered in terms of a rate-distortion (R-D) functional. As in Sec. 7.5 regularity assumptions can be taken into account to exploit redundancy in geometrical atom transformations. Regularity explain directly measured by the rate. The optimization to be solved corresponds to jointly minimizing distortion and rate:

$$\min_{F_N^t} \{D_N + \lambda R_N\}, \tag{7.29}$$

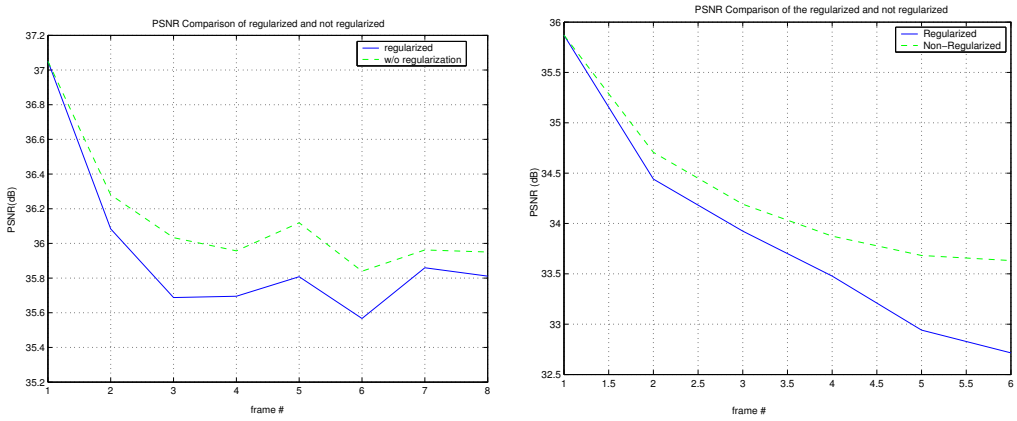


Figure 7.14: Left: Curves representing the loss from frame to frame (corresponding to those of Fig. 7.11) approximation accuracy due to the regularization of the function parameters. Right: Curves representing the loss of frame (from Fig. 7.12) approximation accuracy due to the regularization of the function parameters.

where N represents the number of terms in the expansion to approximate a given frame, D_N is the approximation distortion and R_N the invested rate to code coefficients and parameters. From the properties of Matching Pursuit representations (Sec. 4.4.2),

$$\begin{aligned} D_N &\leq \sum_{n=0}^{N-1} |\xi_n|^2 + \left\| \mathcal{R}^N I_{t+1} \right\|^2 \\ &= \sum_{n=0}^{N-1} |\xi_n|^2 + \|I_{t+1}\|^2 - \sum_{n=0}^{N-1} |c_n|^2 \\ &= \|I_{t+1}\|^2 - \sum_{n=0}^{N-1} \Delta D_n, \end{aligned} \tag{7.30}$$

where I_{t+1} is the original frame to be approximated and ΔD_n corresponds to the contribution to reduce the distortion of atom n, thus $\Delta D_n = D_n - D_{n-1}$. In Eq. (7.30) ξ_n corresponds to the quantization error of the coefficients c_n, and can be assumed to be independent of the coefficient.

If r_n is taken as the total cost needed to code the nth term of the expansion, then it follows from Eqs. (7.29) and (7.30), that Eq. (7.29) can be upperbounded as

$$\min_{F_N^t} \{D_N + \lambda R_N\} \leq \min_{F_N^t} \left\{ \|I_{t+1}\|^2 - \sum_{n=0}^{N-1} \Delta D_n + \lambda \sum_{n=0}^{N-1} r_n \right\} = \min_{F_N^t} \left\{ \hat{D}_N + \lambda R_N \right\}. \tag{7.31}$$

Considering $J_N(\lambda) = \hat{D}_N + \lambda R_N$ then

$$\begin{aligned} &\min_{F_N^t} \{J_{N-1}(\lambda) - \Delta D_N + \lambda r_n\} \\ &= \min_{F_N^t} \{J_{N-1}(\lambda) + \Delta J_N(\lambda)\} \\ &= \|I_{t+1}\|^2 + \min_{F_N^t} \left\{ \sum_{n=0}^{N-1} \Delta J_n(\lambda) \right\}. \end{aligned} \tag{7.32}$$

Thus, a compact representation of the problem is:

$$\min_{F_N^t} \left\{ \sum_{n=0}^{N-1} \Delta J_n(\lambda) \right\}. \tag{7.33}$$

Such a formulation implies a global optimization which, depending on the dictionary (e.g. non-orthogonal) and optimization constraints (e.g. non-divisibility in orthogonal smaller sub-problems), may be of overwhelming complexity (see Sec. 7.2). In the scope of MP, Eq. (7.33) turns into a suboptimal solution where every ΔJ_n is minimized at every iteration. This can be considered as the criteria for selecting $g_{\gamma_n}^{t+1}$:

$$\min_{F_N^t} \left\{ \sum_{n=0}^{N-1} \Delta J_n(\lambda) \right\} \leq \sum_{n=0}^{N-1} \min_{F_{\gamma_n^t}} \{\Delta J_n(\lambda)\} . \tag{7.34}$$

Indeed, for general redundant dictionaries, $\Delta J_n(\lambda)$ $n \in N$ are not necessarily independent among them. *Distortion* reduction and *Rate* investment at a given state of the Weighted-MP algorithm may depend on previous iterations. Closeness to optimality will be conditioned by the structure of the dictionary, how this relates to the signal to approximate and the coding structure.

7.9 Coding the Video Representation

A simple coding scheme is used to code parametric sequence representations. It is based on the predictive coding of the parameter evolution of 2D features through time. In this section, some R-D results are presented for the foreman sequence in QCIF format. First, the scheme used to code temporal evolution of geometric atoms is described. After that, several coding results are presented. In first place, results obtained by using a simple set of *rate* based constraints are shown. Then, the use of regularity based constraints and their effect on R-D results is discussed in more detail.

7.9.1 Predictive Scheme for 3D Structures Coding

The Coding Scheme

The coding algorithm is based on exploiting temporal redundancy of geometric video features. This redundancy is exploited by means of coding predictively the set of parameters that represent each spatio-temporal 3D component. The reader may have seen before in this chapter, that 3D atoms are formed by sets of temporally consecutive 2D atoms that represent the temporal evolution of a geometric primitive. Every atom is tracked through time. For this purpose, a constrained matching pursuit is used, where constraints (see Sec. 7.5.1 and Sec. 7.8) at time t are computed from the 2D expansion obtained at time $t-1$. For every nth "spatio-temporal" MP iteration, the two streams of temporal parameters and coefficients are coded based on a DPCM [156] approach (in the present case, this is based on the simplest of the predictors, i.e. difference prediction, and a uniform dead-zone quantizer). The output residual is then coded by an adaptive arithmetic coder [59, 106, 188]. Symbol statistics are independently estimated for each kind of parameter. A statistical context for every kind of parameter is reserved for "intra" coding (first DPCM sample), and another one is kept for "inter" coding (predicted samples). A module estimates whether the trajectory of an atom is at its end. If this is the case, an additional signaling is transmitted to indicate that the atom, tracked until that point, is not tracked anymore and new intra data (new geometric atom) is introduced for tracking. For sure, another possibility to code temporal parameters and coefficients evolution is to buffer a certain interval of samples and use a wavelet transform (then, efficient wavelet coding may be used, e.g. SPIHT [113]). However, this was out of the scope of this work.

The described coding scheme can be seen in Fig. 7.15. In this figure, one can also see that the estimation of *a priori* information may depend on the previous temporal states of atom trajectories, trajectory states of neighboring atoms and external motion estimators (like block matching, mesh

deformations, region correlation, optical flow, etc...). Predictive representation of spatio-temporal video components is performed on a limited length GOP basis. A maximum prediction length (L) is fixed. Every L frames, all atom trajectories are terminated and a new prediction GOP is started.

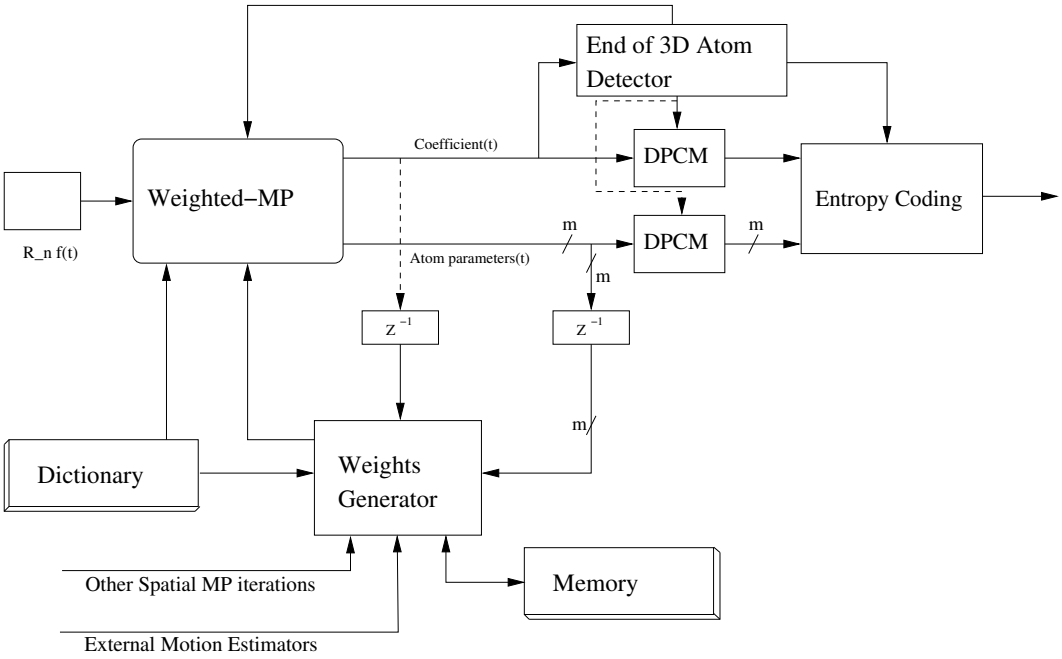


Figure 7.15: Predictive coding of geometric video structures.

As described in Chapter 6, low frequency components are represented separately under the form of a highly downsampled version of the original image. To code this information, low frequency bands are jointly coded by applying a simple temporal wavelet transform on them (spatial wavelet transformation is normally not possible since no further downscaling can be performed). Here, a simple Haar temporal transform is applied to each group of low frequency bands, belonging to the same GOP. These are quantized using a dead-zone uniform quantizer and then coded by means of arithmetic coding, following a raster-scan ordering. Dedicated m-ary adaptive statistical context estimation is used for Haar scaling functions subband and for Haar wavelet subbands.

To guarantee independence among GOPs, all adaptive arithmetic coding contexts are reset at the beginning of every GOP. This may be seen as a resynchronization point of the bit-stream.

Scalability Properties

The nature of matching pursuit allows to easily setup a progressive and scalable bit-stream (as introduced in [78]). Indeed, for every GOP, one may select at reception the reconstruction accuracy of the coded video by simply decoding up to the spatial MP iteration n. This confers to the coding scheme, in a simple fashion, the well appreciated *SNR scalability* property. Moreover, since the dictionary of spatial geometric functions in use is made of a continuous analytic expression, signal description is nothing but a continuous model of the video sequence: one thus gets *Spatial scalability* for free ([66, 78, 79]). Low frequency bands would require classic interpolation to change their size. Nevertheless, more complex fully MP based approaches may substitute these by additionally introducing, in the redundant dictionary, low frequency dedicated functions (e.g. scaled Gaussians

[66]). This kind of dictionaries are known as multi-component dictionaries. The scheme is thus both SNR and spatial scalable. *Temporal scalability* is not available due to its discrete predictive structure.

7.9.2 Results

R-D performances of the presented scheme are evaluated in this section.

Effect of the $D + \lambda R$ Greedy Based Algorithm from a Coding Point of View

The $D + \lambda R$ based Weighted-MP formulation is first evaluated here. Atom tracking is based on the formulation discussed in Sec. 7.8. This intends to impose certain regularity in the sequence description by constraining the 3D atoms extraction to an estimated bit cost of the DPCM residue that will represent the temporal variations of geometric parameters of an atom, as well as its coefficient. Motion parameters are often assumed to be exponentially distributed random variables (of course, this only considers the modulus of these). Optimal entropy coding involves the use of an adapted set of codewords for each symbol. In the exponential case, Exp-Golomb codes are appropriate. Table 7.1 shows the coding costs (codeword lengths) associated to each one of the motion symbols

Symbol to code	*Golomb codeword length*
0	1
1	3
2	3
3	5
4	5
5	5
6	5
7	7
...	...

Table 7.1: Exp-Golomb bit codeword lengths for each symbol of a set with exponential distribution.

if Huffman coding tables were used. Even if arithmetic coding is used here, this table of costs may be considered as a good approximation of the relative costs among different kind of symbols. Of course, λ balances between temporal regularity and distortion at every MP search step. The higher this is set, the lower will be the mobility allowed to atoms. Depending on λ, atoms will only move if distortion reduction is worth enough with respect to the coding cost.

Fig. 7.16 depicts the R-D behavior of the studied coding scheme when using the R-D based Weighted-MP. The reader may observe how the variation of λ is not able to produce significantly enough changes in the R-D curves. The higher the λ value, the better the performance at low bit-rates (and the worst at medium/high rates). The simple regularization law that depends on the coding costs, according to the present coding scheme (coding of the difference from the previous temporal state of a geometric atom), acts just as a motion penalty term. It virtually reduces the atom search window. If λ becomes really high, then all spatio-temporal atoms present zero motion.

The conclusion is that, unlike in the Bayesian approach presented in Sec. 7.5, the $R + \lambda D$ based *a priori* model used for Weighted-MP does not succeed at all in exploiting signal structure. In the following section, the superior R-D performance of the Bayesian regularization approach is shown .

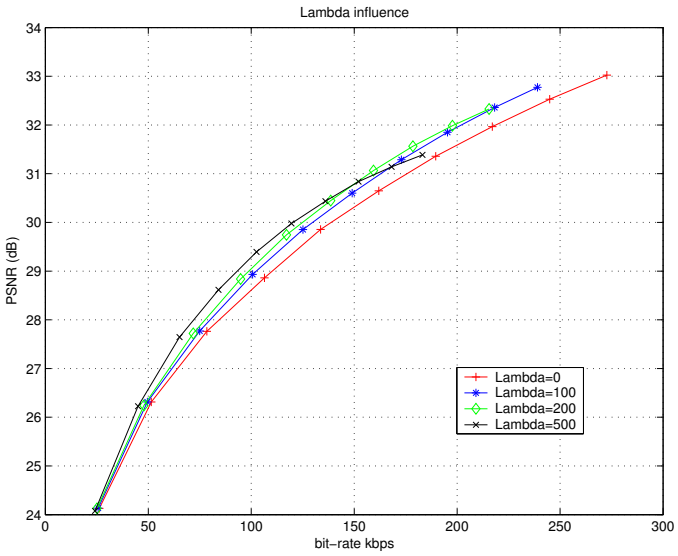


Figure 7.16: Average PSNR on the first 32 frames GOP of the foreman sequence for a given rate and several λ settings.

Effect of Regularity Constraints from a Coding Point of View

In order to have an objective measure of the regularization effects, we consider the R-D curve obtained for the simple coding scheme described in this section. Every frame is described by a set of atoms which are obtained sequentially by the iterative greedy algorithm. The criteria used in the function selection rule of our algorithm is designed such that a spatial and temporal regularity is imposed among atoms. Hence, correlation of atoms at time t with their evolved version at time $t+1$ will be exploited by only encoding the parameters and coefficients differences. When an atom is refreshed, this is obtained by doing a full search in the whole image (as described in Sec. 7.6). Atoms that have been refreshed will also be coded by just sending the difference with respect to the atom they replace in the previous frame. Finally, an arithmetic coding [59, 106, 188] of the differential data is performed.

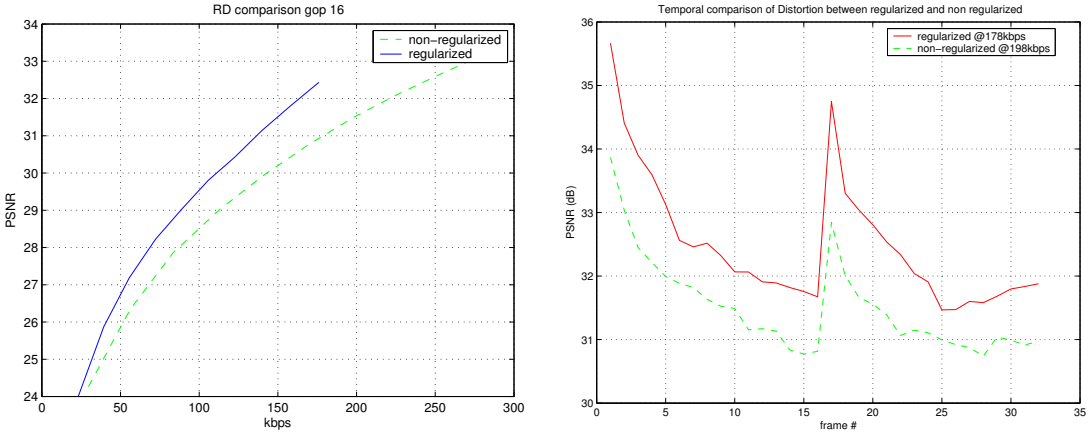


Figure 7.17: Comparison of the regularized and non-regularized foreman sequences (16 frames GOP). Left: R-D, Right: Temporal comparison at two particular rates.

Curves on Fig. 7.17 (left) show the gain obtained in terms of R-D of the regularized Bayesian matching with respect to the non-regularized one. The use of regularization in the matching criteria imposes a certain structure among the behavior of atoms in a frame. This helps reducing the instability of image primitives. A consequence of the regularization turns to be, as expected, the reduction of the amount of necessary bit-rate to represent frame to frame variations. The entropy of the parametric representation gets reduced by the low-pass filtering of parameters imposed by the MRFs criteria. Furthermore, MRFs criteria (and motion initialization when no a priori is available), reduce the propagation of error in atoms parameters, contributing to a better R-D behavior. However, as shown in Fig. 7.14, this is in exchange of a higher drift. A 3dBs loss is cumulated through the GOP. This is due to the limited dictionary resolution, the sub-optimality of MP and the effects inherent to prediction. The range of rates appearing in the curves is obtained by exploiting the natural SNR scalability that MP expansions have. For a given bit-rate, video frames are progressively reconstructed by limiting the number of atoms used per frame. In this way, coding costs may respect a pre-selected bit-rate.

Figure 7.17 (right) shows the effect of the regularization in terms of rate distortion for the foreman sequence. Both curves show the common drift behavior appearing from the predictive nature of the representation. Notice that the regularized version has a gain between 0.5-1.5 dBs over the non-regularized with 20kbps less. Notice the difference between Fig. 7.14 and Fig. 7.17. The weakened matching criteria of the greedy algorithm produces a loss in the regularized approximation with respect to the non-regularized one when the same number of atoms per frame is used in both cases (Fig. 7.14). In the regularized case, atoms are not able to freely move and place themselves to compensate errors done by earlier atoms of the iterative decomposition. Indeed, parameters quantization introduce motion mismatch in the atoms deformation. This is however the price to pay for having some coding gains (Fig. 7.17).

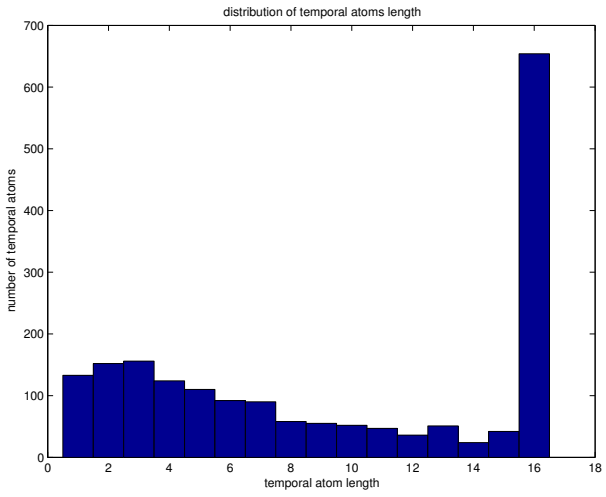


Figure 7.18: Distribution of length for the temporal atoms. The length is determined by the atom refresh criteria of Sec. 7.6 where atoms loosing 80% of their amplitude are refreshed

3D geometric atoms do not necessarily last all along a GOP. Fig. 7.18 shows the histogram of temporal lengths for atoms prediction that are determined by the criteria described in Sec. 7.6. In this example the 48 first frames (3 GOPS) of the sequence foreman have been taken into account for the generation of the statistics. The total number of spatio-temporal atoms (sets of atoms that are predicted from frame to frame without being refreshed) within this 3 GOPs is 1876. There are

about 35 per cent of atoms that succeed in being predicted from frame to frame during all the GOP. However, a relevant number need to be refreshed quite often, common temporal lengths are from 1 to 8 frames. Sequence changes (occlusions, uncoverings and simple interaction among atoms) force their refresh. Atom refresh is a natural manner to introduce components to represent new information that appeared in the signal. However, atom interaction as well as mismatch due to the lack of resolution on the function parameters, contributes to the unnecessary rising of the refresh rate. Hence coding efficiency is reduced.

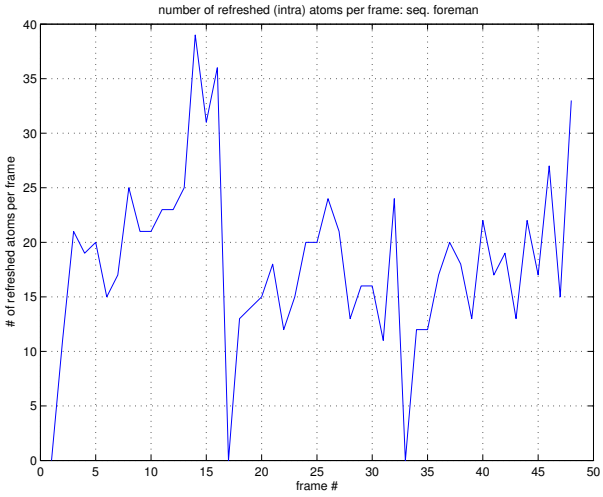


Figure 7.19: Number of new atoms introduced in the refresh procedure in each frame of the sequence.

Atom refresh rate is not the same for every temporal frame, depending on the motion of the picture and the drift propagation through MP iterations, a different number of coefficient primitives pass the threshold fixed to determine whether an atom can still be considered worth to be kept or not. In Fig. 7.19, we monitor the number of refresh atoms introduced at every frame. As shown, intra frames (the first at every GOP) are not considered in this graph. In the graph a zero every 16 frames appears even if all atoms are *refreshed*, i.e. none of them is predicted from the previous frame. In Fig. 7.19 several maximal pics more relevant than others can be identified. Close relations can be found between these and sets of frames of the sequence where most relevant changes appear. In the first GOP, the head appearing in the sequence turns from left to right. This movement progressively occludes a part of the face. At frame 12, the face starts to turn back to its initial position, uncovering in the procedure areas that were not visible before. This requires the insertion of additional information to cope with the change of topology of the contours and shape in the frame 13 (Fig. 7.20)

Finally, coding results of the scheme presented in Sec. 7.9 are compared to a set of scalable coding schemes (These are: MPEG-4 with spatial scalability, MPEG-4 FGS, SPIHT-3D and MP3D. For a list of references, see Chapter 2). Results achieved by the algorithm presented in this chapter appear to be comparable or better than these. The most comparable scheme in performance is that corresponding to MP3D [152]. At very low bit-rates, average representation of structures without taking into account temporal motion is more interesting from a R-D point of view. Indeed, at very-low bit-rates, motion information (in full precision) becomes too expensive to code. For middle and higher rates, the use of motion advantages the approach presented in this chapter. As one can expect, motion adapted 3D geometric primitives can exploit spatio-temporal geometric structures

Figure 7.20: Reconstructed frames 12,13 from the foreman sequence. In them we observe the uncovering of the left region (right in the picture) of the man's face.

in order to reduce signal dimensionality. The video representation achieved in this chapter supplies a signal model which is cheaper to code. The motion information of neighboring atoms should be jointly coded to achieve significantly better results.

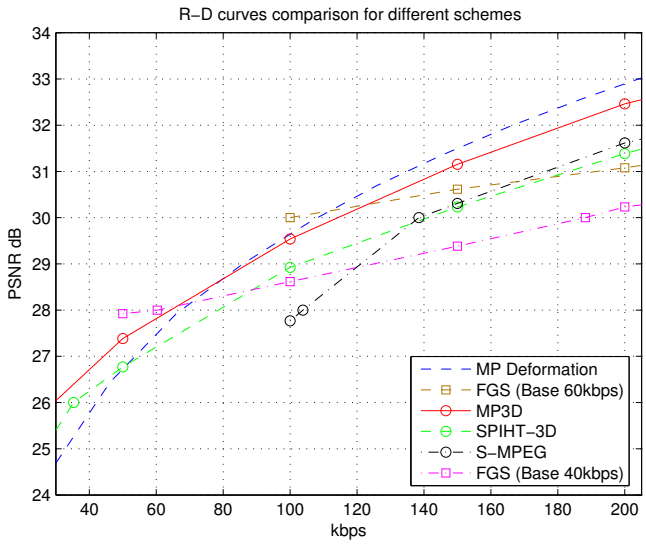


Figure 7.21: Comparison of R-D performance for the sequence Foreman of different scalable coding schemes. R-D data, other than that obtained from the coding scheme of Sec. 7.9, has been obtained from [152].

7.10 Multimodal Analysis Using Redundant Parametric Decompositions

Flexible video representations may be useful for other tasks than just compression. Indeed, efficient video modeling means also efficient video description. Even though descriptions used for a given application are not necessarily adapted to other applications, in this section we test the presently

proposed video representation in the framework of signal analysis. Indeed, the spatio-temporal evolution of geometric video features are used as features in a multi-modal audio-visual analysis experiment.

7.10.1 Motivation

Computer systems have progressed in all their aspects since they first appeared several decades ago. Nowadays, simple personal computers have turned into very complex computation platforms able to perform a large range of tasks. Apart from their basic computational and data ordering purposes, computers are now sophisticated multimedia platforms used for communications, automatic surveillance, efficient (and/or friendly) human-machine interaction, data analysis and many other tasks. Audio and video processing is often the main ingredient of many of these tasks. Signal analysis has usually the final goal of emulating cognitive human capacities. However, this is not always easy to implement and, for the moment, there is still a long research way to go.

Concerning audio and visual data modalities, human beings have a special ability in understanding and analyzing what is happening in an audio-visual scene. Like in many other kinds of signals, interrelations appear often between audio and video. Indeed, given a particular sound event, we, humans, have not much difficulty to locate its origin in a scene, if this has been generated by some visible mechanical action. Humans also exploit audio and visual stimulus correlation for other purposes. Once known the visual location of a sound source, its visual information may be used to enhance sound perception as well. Indeed, multimodal correlation helps "focusing" on a particular sound and eases source separation.

Provide computers with multimodal analysis capabilities is not an easy task. Given the already challenging nature of "monomodal" analysis itself, it has not been until the last five or six years that multimodal analysis has started finding a significant interest among scientists [19, 23, 68, 101, 165, 166]. Although research in this area is quite young and still at a preliminary stage, the increasing engagement of the scientific community is a guarantee that the study of multimodal signal processing is and will be, in the next years, a field in vogue.

In this section, we present the applicability of spatio-temporal geometric video representations within the field of audio-visual multimodal analysis. Monaci *et al.* [131] have used the video representation technique presented in this chapter to extract salient signal geometric features and their associated regular motion in order to combine these with audio features, to retrieve in scenes the location from where sound originates.

7.10.2 Modality Features Extraction & Fusion

The retrieval of correlations between audio and video signals is a problem with a very high dimensionality. For a given distance measure, one desires to locate those spatio-temporal video regions that are interrelated with a certain audio track. In order to make this problem feasible, audio-visual data needs to be modeled such that dimensionality gets reduced and only relevant signal information is used. Data modeling is, thus, supposed to capture the main characteristics of each signal modality that may contain information about the other modality.

Most of the investigated approaches present two main ways of looking for multimodal correlations:

- The most commonly used is the approach that first extracts independently features from each kind of data and latter fuses both modalities in order to determine where data correlation is present.

- Less common (but not less interesting) are those approaches that intend to find, in a jointly manner, an optimal modeling and fusion criteria of data (e.g. [166]). This approach, clearly much more challenging than the usual one, has many chances to supply more powerful solutions in a near future. However, until the moment, results like those of [166] are not able to deal with dynamic scenes yet.

The first approach, which in any case is quite unexplored yet, is considered in here. In the following, feature extraction is described together with some comments comparing this to the state of the art. Next to that, the use of a data fusion criteria is proposed.

Audio Feature Extraction

Audio signals have a rich variety of components that human auditive system is able to perceive (Fig. 7.22). This is the reason why high sampling frequencies are, thus, normally required to preserve all the useful signal frequencies. Correlations of the wide diversity of sounds with the also large variety of geometric configurations of the visual stimulus of a mouth are possible. Indeed, this is the main basis for *lip reading*. A positional model of lips may be assigned to each sound and transitional models between sounds can be established.

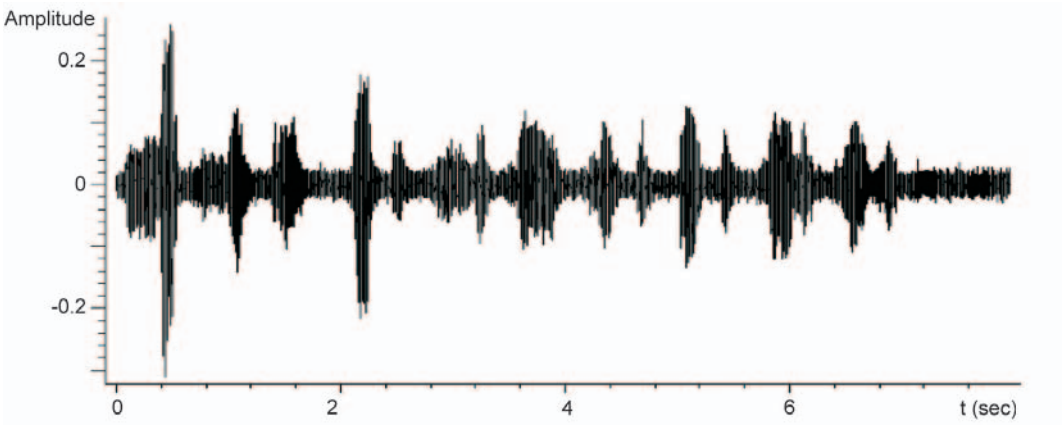


Figure 7.22: Audio signal of a subject uttering the first ten digits (in Italian) [131].

In this section we consider a much simpler and generic approach. We look for the location in video of simple vibrations that can be found at the same time in audio and visual motion. Video temporal sampling, and consequently motion, are usually of very low rate. Hence, audio features of interest for us are in a similar variability range. Higher dimensional audio features are unnecessary and these may even be misleading.

In literature, one may find different kinds of audio features used. [101] use the instantaneous energy of the audio track as feature. [165] bases its audio analysis on cepstral representations [150]. [23] and [68] use a linear combination of spectral power coefficients through time. This combinations are such that they maximize, respectively, feature entropy and mutual information with video. In here, an estimate of audio energy contained per frame is taken into account (see Fig. 7.23). A direction to investigate is the use of subsets of time-frequency sparse representations (like signal projections on redundant Gabor dictionaries [131]) such that more efficient audio-visual data correlations can be achieved.

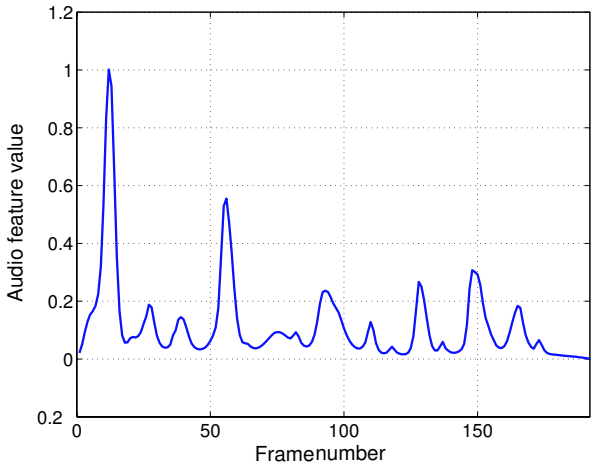


Figure 7.23: 1D Audio feature based on the normalized measure of the instantaneous audio energy for a sampling rate equal to the video frame rate [131]. The original signal may be seen in Fig. 7.22.

Video Feature Extraction

Clearly, video features need to capture temporal variations. To date, video features used for multimodal audio-visual fusion are often based on pixel-wise intensity difference measures [68, 101, 165, 166] or regularized pixel intensity measures [23]. Some approaches look forward exploiting local motion information by means of optical flow measures [19]. In any case, none of the actual approaches try to exploit the real structural nature of video signals.

All investigations carried out in this thesis have as fundamental purpose to take into account the geometric nature of video. Efficient signal modeling and representation requires the use of methods able to capture particular characteristics of each signal kind. A question that arises at this point is: Why should we use a representation of video based on a basis of *deltas* (i.e. pixel wise features), if video is made of moving regions surrounded by contours with high geometrical content? For particular applications, one may consider the use of adapted template based approaches (in order to model particular objects and their trajectories: lips, faces, etc...). However, for generic non-application constrained approaches, the answer seems to be that we should, indeed, use a signal modeling capable to exploit video structural properties while keeping generic and flexible enough.

In a tentative to introduce such properties into the video feature extraction process, the use of the spatio-temporal video approximations using geometric primitives is considered. In particular, the approach adopted is the one presented in this chapter. Video is decomposed in 3D video components intended to capture geometric components (like oriented edges) and the temporal evolution of their geometry.

For example, potential features to correlate with audio are: the temporal variation of position, orientation, scaling or coefficient projection of 2D geometric primitives (i.e. each of the g_γ^t and F_γ^t $\forall t \in [0, T]$.

Data Fusion

Once features are available, a measure is required to determine how much these are related among them. In the literature, different fusion criteria may be found. These are selected depending on the assumptions done to formulate the multimodal analysis problem.

Information theoretic formulations often use the *Mutual Information* measure [36]. However,

when few data samples are available, there may be some problems concerning probability density estimations. Other approaches consider the formulation of the problem from a projective point of view. Assuming that all extracted features (audio and video in our case) are data vectors that belong to a same *projective space**, a distance can be established among these. Correlation measure determines the distance between two given feature vectors. Those vectors having a maximum correlation (i.e. maximum scalar product) are those which are situated at a minimum distance. When mean is removed systematically from each feature vector and then this are normalized, the distance measure is known as the *Pearson* ρ correlation coefficient [11]:

$$\hat{\rho} = \frac{\sum_{i=1}^{n} (x_i - E\{\mathbf{x}\})(y_i - E\{\mathbf{y}\})}{\sqrt{\sum_{i=1}^{n} (x_i - E\{\mathbf{x}\})^2 \cdot \sum_{i=1}^{n} (y_i - E\{\mathbf{y}\})^2}}, \tag{7.35}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two feature data vectors. In [131], *Pearson* correlation showed to be the most appropriate criteria to fuse the audio and video features discussed in this section. Hence, this is the measure used to generate the results presented in the next point.

As shown in [131], a significance test on the result of (7.35) needs to be performed in order to determine which 3D video components have a correlation with audio which is relevant enough.

The following function of $\hat{\rho}$:

$$\hat{f}_t(\hat{\rho}) = \frac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}},$$

where n is the length of feature vectors, belongs to a *Student's* distribution with $n-2$ degrees of freedom if audio and visual feature correlation is zero. If the probability that the measured $\hat{f}_t(\hat{\rho})$, for a given 3D video atom, belongs to a *Student's* distribution is small enough, then this atom is considered to be correlated with sound (the reader is referred to [131] for further details).

7.10.3 Results

In here, two illustrative results are shown of the usage of geometric video representations for audio-visual data correlation. The examples presented are based on two video sequences where in each, two subjects pronounce a certain text one after the other. Fig. 7.24 and Fig. 7.25 clearly show the location of geometric atoms found to have a temporal evolution correlated with sound. This perfectly coincides with those primitives that take part in mouth, nose and chin structures. As a relevant detail, in Fig. 7.24 the left subject continues to move the mouth while the second subject speaks. Only the source of sound is detected. The colored location indicators are the energy envelops of the selected atoms. Notice how these nicely adapt their orientation according to the geometric characteristics of the structures they represent.

The presented spatio-temporal geometric video representation based on atoms deformation, has shown to be capable to recover 3D video components that represent geometry and their motion through time. This representation has shown to be an interesting alternative for video feature extraction in multimodal signal analysis. Results are promising and encourage for further research. The interested reader will find additional results and future development of this research in [130].

*A *projective space* is understood here as the set of all one-dimensional subspaces spanned by each of the possible real vectors of dimension n (i.e. n is the size of our feature vectors). Each one of the feature vectors generates one of such one-dimensional subspaces.

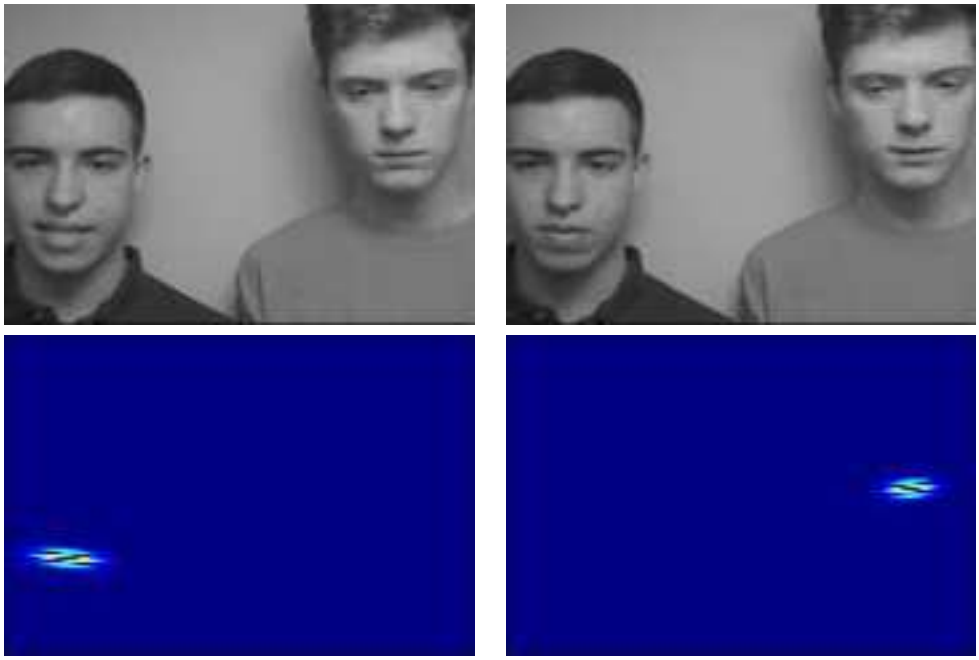


Figure 7.24: Up: Original sequence frames. Down: Spatio-temporal geometric atoms with significant correlation with audio track [130].

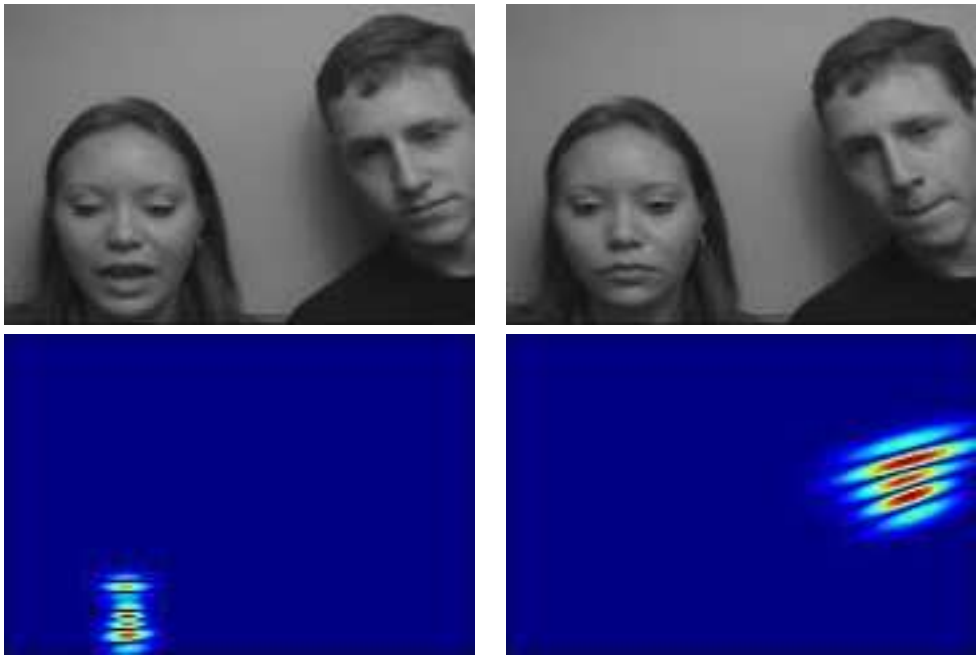


Figure 7.25: Up: Original sequence frames. Down: Spatio-temporal geometric atoms with significant correlation with audio track [130].

In general, results show to be very positive, but far from being totally robust. In effect, the tests put into relevance several weak points of the present implementation for 3D geometric structures extraction:

- First of all, the heuristic solution supplied in Sec. 7.6 to handle occlusions, appearance and disappearance effects is not the optimal one. However, there is not much choice since geometric structure tracking is performed in a forward predictive way. This produces an undesired drift effect that affects the robustness of the algorithm.

- Another point to underline is the simplicity of the *a priori* regularity model used for the Weighted-MP. Moreover, at this research stage, this still requires parameter training (though parameters could be self optimized by some non-linear iterative procedure).

- Finally, let us remind that temporal evolution of the estimated geometric structures is subject to the granularity of the dictionary in use. This limits motion resolution, generating noisy estimates of geometric features motion.

7.11 Conclusions

In this chapter, a method for the decomposition of video signal on 3D geometric structures has been presented. The purpose of this is the recovery of sparse video approximations where atoms jointly represent spatial geometry and temporal trajectories. The problem has been formulated taking into account lessons learned in previous chapters for the use of redundant dictionaries and *a priori* information. Indeed, the use of *a priori* information, appears of key importance for the recovery of video structures due to dictionary coherence. 3D video structures are extracted in a forward predictive way. Atoms retrieved in a frame at time t are progressively deformed to match the successive transformations of posterior frames. The set of temporal deformations associated to a given atom g_γ^t together with the initial description of g_γ^t form one of the spatio-temporal atoms used to describe the sequence of frames. The obtained description shows to have encouraging results in coding applications as well as in video analysis applications.

The forward predictive strategy used to solve the problem, makes occlusions, appearing and disappearing objects to be poorly handled. Dictionary parameters granularity reduce the resolution of motion representations. Even if this is necessary for coding purposes, video signal analysis requires as much resolution as possible to reduce noise in motion estimates of structures. The present approach assumes that motion is constant all over the support of a 2D geometric feature. However this is not necessarily true.

This chapter has clearly unmasked most of the unknowns concerning the representation of video signals by means of the superposition of 3D atoms, able to represent spatial geometry as well as temporal transformations. From the observed advantages and drawbacks of the present implementation, we can state the following requirements for future research and applications implementation:

- If applications are not limited by delay constraints, fully 3D atoms extraction needs to be adopted instead of the forward predictive one. This typically introduces in a natural fashion handling with occlusions, appearing and disappearing objects. In the 3D extraction, the same principles on the use of Weighted-MP discussed in this chapter can be used. The difference with respect to the forward prediction scheme is that signal does not need to be processed respecting its causal nature. Forward prediction toward future and past (according to the video time line), considering any temporal instant as possible starting point, can be used.

- A better compromise between complexity and motion resolution recovery is required. Video analysis requires above all representations able to give parametric descriptions of sequences as accurate as possible. For the case of video compression, as far as rate is kept limited, motion resolution is of great importance as well.
- Investigation on better and more structured *a priori* models is required. From a video analysis point of view, a better *a priori* will supply better parameter estimation accuracy. From a video coding point of view, an appropriate *a priori* may be the necessary tool to increase R-D performance.

An interesting example of a fully 3D approach (non-predictive based), that uses an *a priori* based exclusively on a BM translational motion estimation, conceived for video coding, can be found in [151]. In this, however, spatial geometry modeling is conditioned by the BM translational motion estimation, which does not take the real nature of spatial video structures into account.

Chapter 8

Conclusions

8.1 Summary

In this dissertation, several aspects of adaptive models for video approximations have been investigated. Efficient approximation methods for flexible video coding approaches must be capable to adapt as much as possible to the signal. In video approximations, this implies that signal geometry must be efficiently modeled. Video is a 3D signal with geometric content in spatial structures as well as in their temporal evolution.

Temporal geometry information, classically exploited by motion estimation, is typically embedded within separable temporal transforms (like wavelets) in order to extract spatio-temporal signal components which exploit temporal geometry in a multi-scale fashion. Since video is an example of piecewise-smooth signal, the effect of adaptive temporal representations has been studied. Indeed, in the case of motion compensated temporal filtering, temporal wavelet decompositions must be adapted to the length of object motion trajectories. In the simpler case of non motion compensated 3D subband video coding, motion is assumed to be equal to zero. Hence, when this assumption fails, temporal transforms must be adapted in length such that wavelet transforms avoid crossing the temporal edges generated in the signal by moving objects.

Present video representation techniques model temporal geometry without taking into account the real nature of spatial video structures. Even though recent approaches try to decompose video signals on a set of multi-scale motion adaptive 3D wavelet functions [12, 160], which already use the multi-scale spatial nature of video frames, spatial geometry is not taken into account at all.

Ideally, given the 3D geometric nature of video signals, these should be approximated by models capable to accurately jointly represent space and time geometry. Considering the modeling of video signals like the superposition of basis functions from a dictionary, basis functions should be capable to model 3D video components describing a geometric feature and its evolution through time. Very coherent redundant dictionaries seem to be required for this task. Moreover, the use of frame based signal decompositions appears to be of no use due to their lack of sparsity preservation. In this case, the use of highly non-linear decomposition algorithms to achieve sparse video representations is required. Nevertheless, due to the computationally demanding task to decompose video on dictionaries made of 3D spatio-temporal geometry based functions. Rich enough dictionaries (in order to accurately model all sorts of geometry) are so big that special strategies to extract spatio-temporal geometric video components must be considered. Moreover, usable highly non-linear decomposition algorithms such as Matching Pursuits do not work as desired when dealing with highly coherent

dictionaries of functions.

In this thesis, the extraction of spatio-temporal 3D geometric video components has been studied in detail. The strategy investigated is based on the retrieval of 2D geometric components in a video frame, and then looking for their evolution through time. The set of spatial parameters of the 2D geometric functions together with their variation through time, form the desired 3D geometric video components. The approach presented in this work adopts a predictive strategy in order to look for the temporal evolution of 2D geometric components, respecting the causality of the signal. 2D components are tracked from frame to frame to determine their temporal variation. Retrieving a *sparse* representation is the main concern of this work. Hence, as aforementioned, we must use some highly non-linear decomposition algorithm with our redundant dictionaries (i.e. with the 2D geometric dictionaries too). A greedy algorithm has been adopted for the features tracking task. This extracts the trajectory of a geometric feature at every iteration. Indeed, global optimization of all 2D atom trajectories at once would result into a too complex task.

However, the combination of highly coherent dictionaries with matching pursuits usually does not work as one would like. Greedy algorithms are a sub-optimal highly non-linear optimization algorithm capable to find the optimal solution to signal decompositions in particular occasions. One can ensure the good behavior of matching pursuits when the dictionary in use is incoherent enough. Though, interesting dictionaries for the decomposition of images and video seem to be quite coherent. In order to solve this problem, which also concerns a very long list of other applications (not just the ones concerned by video approximations), a new paradigm for highly non-linear decomposition algorithms is proposed and deeply analyzed in this work. In order to make highly non-linear algorithms to work better with coherent dictionaries, *a priori* models relating the internal structure of dictionaries with the signal must be considered. A particular example of these proposed algorithms is Weighted-MP/OMP. One can significantly enhance signal expansions sparsity by modifying the algorithm functions selection rule such that the MP/OMP algorithm becomes signal adaptive.

Weighted-MP/OMP within the video expansion problem can be seen as a greedy algorithm which considers additional information about probable motion trajectories of video components. In our case, this was implemented by imposing regularity constraints on the motion of neighboring geometric components as well as by computing pre-estimates of local motion with some pixel-based matching technique. Geometric video decompositions have been tested also in terms of performance of video compression as well as for multi-modal audio-visual analysis purposes. The results obtained have shown to be promising, encouraging to prosecute research on the subject.

In order to have a clear sketch of the main contributions of the thesis, these are recalled in the following.

Localized Temporal Adaptivity in 3D Wavelet Video Coding

A model based R-D theoretical analysis of different wavelet decomposition strategies has been performed for motion compensation free subband video coding. This analysis has given a better understanding of coding performance of non-linear video approximations with 3D separable wavelet basis. A locally adaptive temporal decomposition strategy has been suggested in order to improve the R-D performance of coding applications.

Intra-Adaptive Motion-Compensated Lifted Wavelets for Video Coding

A piecewise-smooth model concerning the temporal behavior of motion compensated video data has been described. According to this, and the theoretical background on R-D performance of wavelet

and oracle based coding schemes, an intra-adaptive scheme of the MCTF extension of H.263++ has been proposed in order to allow a better modeling of motion trajectories in video signals. This has allowed to achieve better R-D performances on MCTF video codecs.

Use of *a Priori* Models within Highly Non-linear Algorithms for Sparse Repersentations and Approximations

A detailed theoretical analysis has been performed on the influence of using accessory *a priori* models in highly non-linear signal expansion algorithms together with coherent dictionaries. Practical examples validate the theoretical findings as well as they show how *a priori* information, dictionary and signals must be related within highly non-linear decomposition algorithms. Very redundant dictionaries for the approximation of signals with particular properties, often include sets of functions specialized in representing some kind of feature. The good behavior of decomposition algorithms is directly related with their capacity to assign these specialized functions to the right signal features.

An Efficient Full Search Matching Pursuit Image Decomposition Algorithm

A feasible efficient full search matching pursuit strategy for image decompositions using 2D geometric dictionaries has been proposed. This substitutes previous sub-optimal strategies based on genetic algorithms, which delivered worse approximation results and showed to be slower. The proposed decomposition technique allows to obtain efficiently sparse geometric MP based image approximations.

A Full Geometry Based Video Approximation Scheme

An *a priori* based predictive greedy strategy to extract 3D primitives from video signals has been proposed. This has supplied a spatio-temporal geometric video representation scheme based on the superposition of these video primitives. The obtained representation, when used for signal scalable compression, has shown to be, in average, more efficient than some *state of the art* techniques. This representation approach shows to be also quite interesting as a source of video features for multi-modal data fusion.

8.2 Future Research

Apparently, it seems to be quite usual that at the end of a PhD thesis one feels like ready to, at last, start serious research. Hence, plenty of ideas come in mind and a frustrated voice coming from inside tells you: *Why didn't I think about that before??!!* (well, better not to think about that...).

Reached this point, last but not least, let us review a very little set of possible future research directions to follow; just in case these are of interest for the reader.

Subband Video Coding

Lifting based adaptive subband video decompositions select in a level by level fashion the coding modes used to tune the intra-adaptive and frame-adaptive steps. This is completely sub-optimal since its effect on the global signal transform is not taken into account. Instead of locally optimizing the parameters of the video transform, these should be selected in a global optimization fashion. While this is tractable by prune-join tree decompositions in the non motion compensated

3D Wavelet case, it is tremendously demanding in terms of computational complexity for the MCTF case. Indeed, the optimization problem becomes combinatorial with many dimensions. Sub-optimal approaches, where at least more than one decomposition level is taken into account to select the lifting steps modes, could be a compromise between complexity and R-D performance.

Applications of *A Priori* based Weighted Highly Non-linear Algorithms for Sparse Representations and Approximations

Weighted highly non-linear algorithms for signals decompositions, like Weigthed-MP/OMP, can be applied to any application requiring the use of coherent redundant dictionaries as well as sparse signal representations or approximations. General under-determined inverse problems can use this paradigm as far as a certain model can be established in order to relate signal and internal dictionary structure.

Good signals modeling involves the retrieval of independent components in signals. This can be applied to audio analysis and compression, video analysis and compression, image analysis and compression, data mining and content indexing, "blind" source separation, multi-modal data processing, etc. For all this purposes, particularly adapted dictionaries have to be investigated as well as signal/dictionary based *a priori* models. *A priori* models should be carefully studied depending on the class of signals to approximate in order to determine the guarantees that a particular dictionary together with a particular *a priori* model have to achieve good sparse signal decompositions. Much better application performances than those obtained by using orthonormal basis are expected for the future.

Efficient Full Search Matching Pursuit Enhancements

Efficient full search approaches for Matching Pursuit decompositions must profit from the implicit structure of redundant dictionaries and the way basis functions are constructed. The use of steerability properties of some functions may allow much faster computation strategies of scalar products between all functions of the dictionary and the signal. Moreover, a deep investigation of appropriate dictionaries for natural images approximations should be performed.

Full Geometry Based Video Approximations

The spatio-temporal video representation presented in this thesis is based on a predictive causal strategy for the extraction of 3D video geometric components. To avoid difficulties handling objects occlusions and apparitions, a 3D approach should be considered instead. A fully 3D decomposition based on a GOP should be taken into account. In this way, spatio-temporal geometric components can be extracted from any spatio-temporal location in the GOP. This approach would be more respectful with the video structure. Moreover, the general problem formulation should be defined in a slightly different way. 3D components extraction should be performed in a even higher non-linear optimization way in order to jointly optimize the selection of the 2D geometric shapes together with the motion parameters retrieval. In addition, this could be done such that, the same motion parameters are imposed to a neighborhood of nearby atoms. The neighborhood should be decided within the optimization procedure leading to a complex chicken-egg segmentation and classification procedure. However, it is not clear whether such an approach is computationally feasible. In addition to all that, other geometric dictionaries can also be investigated.

Applications of 3D Full Geometric Video Descriptions

As demonstrated in this PhD thesis, 3D video geometric components may be used as basic building blocks for compression models as well as features for video analysis applications. For each one of this applications, one may investigate different dictionaries in order to decide which may be more useful in each case. Appropriate coding strategies as well as appropriate data fusion approaches must be investigated.

APPENDIX

Appendix A

Performance Proofs of 3D Schemes on the Moving Horizon Model

A.1 Proof of Theorem 3.2

Proof: The R-D analysis of classic Packet Wavelet for video coding (2D+1D separable wavelet decomposition) is analogous to the one performed in Sec. 3.4.2. The main change, concerns the generated number of wavelet coefficients. Due to the 2D+1D condition of the present transform, now, decomposition levels will refer to the 2D ones. Hence, in the remaining of the proof, index j will involve all spatio-temporal coefficients at spatial scale 2^{-j}.

The number of coefficients per spatial decomposition level for the *Moving Horizon* model are such that:

$$n_j \sim 3 \cdot 2^{2j} \left((j+2) + \sum_{j'=1}^{J-j} 2^{j'} \right). \tag{A.1}$$

Consequently, the total number of coefficients involved in the approximation of our synthetic signal is

$$N_J \sim \sum_{j=0}^{J} n_j + \frac{n_o}{3} = 4^J \left(4J + \frac{32}{3} \right) - 4 \cdot 2^J + \frac{4}{3}. \tag{A.2}$$

The use of the fixed packet scheme, for the 2D+1D dyadic decomposition, does not modify the overall maximum decay of coefficients depending on the level. In effect, temporally diagonal subbands are the same for the present scheme and for the isotropic 3D wavelet decomposition. In this case, the amplitude decay of coefficients, the rate needed to code them and distortion introduced for every coefficient are also:

$$\begin{aligned} |c_{j,\mathbf{k}}| &\sim 2^{-\frac{3j}{2}}, \\ R_{c_{j,\mathbf{k}}} &\sim \log_2 \left(\frac{1}{\Delta} \right) = \frac{3J}{2}, \\ D_{c_{j,\mathbf{k}}} &\sim \Delta^2 \sim 2^{-3J}. \end{aligned}$$

Following from Eq. A.2, we establish the relation between rate, distortion and the finest level (J) of

detail involved in the signal approximation as

$$\begin{aligned} R &\sim 4^J\left(6J^2+48J+\frac{256}{3}\right)-2^J\left(6J+32\right)+2J+\frac{8}{3}, \\ D &\sim 2^{-J}\left(4J+\frac{53}{3}-4\cdot 2^{-J}+\frac{4}{3}\cdot 4^{-J}\right). \end{aligned} \tag{A.3}$$

The value of J as a function of R is required to determine the global behavior of $D(R)$.

At this point, a lower bound on the distortion may be computed. For this, we need to approximate the value of R by a lower bound such that J can be easily solved:

$$\begin{aligned} R &\gtrsim 4^J\left(6J^2+48J+\frac{256}{3}\right)-2^J\left(6J+32\right)+\frac{8}{3} \\ &= 4^J\left(3J+16\right)\left(3J+8\right)\frac{2}{3}-2^J\left(3J+16\right)2+\frac{8}{3} \\ &\geq 4^J\left(3J+11\right)^2\frac{2}{3}-2^J\left(3J+11\right)2+\frac{8}{3} \\ &= 4^J\left(J+\frac{11}{3}\right)^2 6-2^J\left(J+\frac{11}{3}\right)6+\frac{8}{3} \end{aligned} \tag{A.4}$$

In order to solve (A.4), we consider the change of variables $J'=J+11/3$ and $y=\sqrt{J'}2^{J'}$. This, turns (A.4) into a simple one variable second order equation:

$$R \gtrsim 6\cdot 4^{-11/3}y^2-6\cdot 2^{-11/3}y+\frac{8}{3}.$$

Taking the solution that implies a positive relation between y and R results into:

$$y \lesssim 4\cdot 2^{2/3}+\frac{4}{3}\sqrt{-14\cdot 2^{1/3}+12\cdot 2^{1/3}R}.$$

Combining this with the change of variable $y=\sqrt{J+11/3}2^{(J+11/3)}$ and solving for J, one obtains an upper bound on the dependency of J on R. That is,

$$J \lesssim \frac{\mathrm{W}\left(\frac{4}{3}\log(2)\left(3\cdot 2^{2/3}+\sqrt{2}\sqrt{2^{1/3}\left(-7+6R\right)}\right)\right)-\frac{11\log(2)}{3}}{\log(2)}. \tag{A.5}$$

The desired lower bound on the $D(R)$ is, thus, obtained from the combination of (A.3) and (A.5). Hence,

$$\begin{aligned} D(R) &\gtrsim \frac{32}{3\log(2)}\frac{\left(3\mathrm{W}\left(E(R)\right)-\frac{11}{\log(2)}\right)2^{2/3}\mathrm{W}\left(E(R)\right)}{E(R)} \\ &\quad +\frac{424\cdot 2^{2/3}}{3}\frac{\mathrm{W}\left(E(R)\right)}{E(R)}+\frac{8192}{3}\left(\frac{\mathrm{W}\left(E(R)\right)}{E(R)}\right)^3 \\ &\quad -512\cdot 2^{1/3}\left(\frac{\mathrm{W}\left(E(R)\right)}{E(R)}\right)^2 \end{aligned} \tag{A.6}$$

where $E(R)=\frac{4}{3}\log(2)\left(3\cdot 2^{2/3}+\sqrt{2}\sqrt{2^{1/3}\left(-7+6R\right)}\right)$.

At high rates, $D(R)$ can be approximated by a much simpler expression that corresponds to its asymptotic behavior. In effect, the three last terms of (A.6) become insignificant while rate increases.

$E(R)$ behaves linearly with $\sqrt{R}$ and $\mathrm{W}(E(R))$ can be approximated by $\log(E(R))$. Hence, $D(R)$, at high rates, can be expressed as:

$$D(R) \sim \frac{\log^2(R)}{\sqrt{R}}, \tag{A.7}$$

which proves Theorem 3.2.

■

A.2 Proof of Theorem 3.3

Proof: We follow again the procedure adopted in Sec. 3.4.2 and 3.4.3 to determine an upper bound on the $D(R)$ as well as its asymptotic behavior at high rates.

This time, the number of coefficients per spatial decomposition level for the *Moving Horizon* model is:

$$n_j \sim 3 \cdot 2^{2j} \left(2 + \sum_{j'=j+1}^{J} 2^{j'-j} \right), \tag{A.8}$$

which gives a the total of non zero coefficients:

$$N_J \sim \sum_{j=0}^{J} n_j + \frac{n_o}{3} = 4^J 12 - 2^J 4. \tag{A.9}$$

The same characterization used for $|c_{j,\mathbf{k}}|$, $R_{c_{j,\mathbf{k}}}$ and $D_{c_{j,\mathbf{k}}}$ in sections 3.4.2 and 3.4.3 applies in here. Thus, the distortion and rate behavior with respect to J can be expressed as:

$$\begin{aligned} R &\sim 4^J (18J + 96) - 2^J (6J + 32) - 8, \\ D &\sim 2^{-J} 19 - 4^{-J} 4. \end{aligned} \tag{A.10}$$

An upper bound on R can be obtained in the following way:

$$R \sim 4^J (18J + 96) - 2^J (6J + 32) - 8 \tag{A.11}$$

$$= 4^J \left(J + \frac{32}{6} \right) 18 - 2^J 6 \left(J + \frac{32}{6} \right) - 8 \tag{A.12}$$

$$\leq 4^J \left(J + \frac{32}{6} \right) 18 - 2^J 6 \sqrt{J + \frac{32}{6}} - 8. \tag{A.13}$$

From this, we can easily solve J if two changes of variable are applied to convert (A.13) into a second order equation, i.e. $J' = J + 32/6$ and $y = \sqrt{J'} 2^{J'}$. We follow the same procedure to that of Sec. 3.4.3. Hence,

$$R \lesssim 4^{J'} J' 4^{-32/6} 18 - 2^{J'} 2^{-32/6} 6 \sqrt{J'} - 8 \tag{A.14}$$

$$= y^2 4^{-32/6} 18 - 2^{-32/6} 6y - 8, \tag{A.15}$$

and

$$y \gtrsim \frac{16}{3} 2^{1/3} + \frac{16}{3} \sqrt{17 \cdot 2^{2/3} + 2 \cdot 2^{2/3} R}. \tag{A.16}$$

It just remains to apply backward the previous variable changes to (A.16) to get the final bound on J:

$$J \gtrsim \frac{3\ \mathrm{W}\left(2\log(2) \left(\frac{16}{3} 2^{1/3} + \frac{16}{3} \sqrt{17 \cdot 2^{2/3} + 2 \cdot 2^{2/3} R} \right)^2 \right) - 32\log(2)}{6 \cdot \log(2)}. \tag{A.17}$$

An upper bound on $D(R)$ follows from (A.10) and (A.17):

$$D(R) \lesssim 304 \cdot 2^{5/6} \frac{\sqrt{\mathrm{W}\left(2\log(2)E(R)^2\right)}}{\sqrt{\log(2)}E(R)} - 2048 \cdot 2^{2/3} \frac{\mathrm{W}\left(2\log(2)E(R)^2\right)}{\log(2)E(R)^2}, \tag{A.18}$$

where $E(R) = 16/3(2^{1/3} + \sqrt{17 \cdot 2^{2/3} + 2 \cdot 2^{2/3}R})$. Thus, $D(R)$, at high rates, can be expressed as:

$$D(R) \sim \sqrt{\frac{\log(R)}{R}}, \tag{A.19}$$

which proves Theorem 3.3.

■

Appendix B

Proofs on the Use of *A Priori* Models in Greedy Algorithms

B.1 Proof of Theorem 5.1

Proof: The inner product between the residual at a certain iteration of MP and an atom from the dictionary can be interpreted as the probability of that atom to be selected. According to [174], we see that, at every iteration, the following should be satisfied for a $Weak(\alpha)$ greedy algorithm:

$$\rho\left(r_k\right) = \frac{\left\|W_{\overline{\Gamma}}\left(D_{\overline{\Gamma}}^T r_k\right)\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma^T r_k\right)\right\|_\infty} < \alpha, \tag{B.1}$$

where, as stated previously, W_Γ, $W_{\overline{\Gamma}}$ are two diagonal sub-matrices of $W(f,\mathcal{D})$ containing the weights $w_i \in (0,1]$ corresponding to D_Γ and $D_{\overline{\Gamma}}$. According to the assumption that $r_k \in span(D_\Gamma)$ and that the columns of D_Γ are linearly independent, then $r_k = \left(D_\Gamma W_\Gamma\right)\left(D_\Gamma W_\Gamma\right)^+ r_k = P_\Gamma r_k = P_\Gamma^T r_k$, where P_Γ is the orthogonal projector on the space spanned by D_Γ. This gives:

$$\frac{\left\|W_{\overline{\Gamma}}\left(D_{\overline{\Gamma}}^T r_k\right)\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma^T r_k\right)\right\|_\infty} = \frac{\left\|W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left(D_\Gamma W_\Gamma\right)\left(D_\Gamma W_\Gamma\right)^+ r_k\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma^T r_k\right)\right\|_\infty} =$$

$$\frac{\left\|W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left(\left(D_\Gamma W_\Gamma\right)^+\right)^T \left(D_\Gamma W_\Gamma\right)^T r_k\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma^T r_k\right)\right\|_\infty}.$$

This quantity can be bounded by:

$$\frac{\left\|W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left(\left(D_\Gamma W_\Gamma\right)^+\right)^T \left(W_\Gamma D_\Gamma^T\right) r_k\right\|_\infty}{\left\|W_\Gamma\left(D_\Gamma r_k\right)\right\|_\infty} \leq$$

$$\left\|W_{\overline{\Gamma}} D_{\overline{\Gamma}}^T \left(\left(D_\Gamma W_\Gamma\right)^+\right)^T\right\|_{\infty,\infty} = \left\|\left(D_\Gamma W_\Gamma\right)^+ \left(D_{\overline{\Gamma}} W_{\overline{\Gamma}}\right)\right\|_{1,1}.$$

Given that $\|\cdot\|_{1,1}$ is the maximum ℓ_1 norm of the columns of a matrix, and that the weighting

matrices are diagonal, then the *Exact Recovery Condition* is

$$\sup_{g_i \in D_{\overline{\Gamma}}} \left\| \left(D_\Gamma W_\Gamma\right)^+ g_i \cdot w_i \right\|_1 < \alpha, \tag{B.2}$$

where w_i is the corresponding *a priori* factor of g_i from the diagonal of $W_{\overline{\Gamma}}$.

■

B.2 Proof of Theorem 5.3

Proof: Theorems 5.1 and 5.2 give the conditions under which Weighted *Weak*-MP and WBP recover the optimal set of atoms. In this proof the factor α is conserved independently of the algorithm in use. Note that for the particular results of WBP and Weighted-MP/OMP this value equals 1.

Starting from (B.2) and following the procedure suggested in [174] an upper bound based on μ_1^w can be obtained:

$$\sup_{g_i \in D_{\overline{\Gamma}}} \left\| \left(D_\Gamma W_\Gamma\right)^+ g_i \cdot w_i \right\|_1 =$$

$$\sup_{g_i \in D_{\overline{\Gamma}}} \left\| \left(\left(D_\Gamma W_\Gamma^T\right)^T \left(D_\Gamma W_\Gamma^T\right) \right)^{-1} \left(W_\Gamma D_\Gamma^T\right) g_i \cdot w_i \right\|_1 \leq \tag{B.3}$$

$$\left\| \left(\left(W_\Gamma D_\Gamma^T\right) \left(W_\Gamma D_\Gamma^T\right)^T \right)^{-1} \right\|_{1,1} \cdot \sup_{g_i \in D_{\overline{\Gamma}}} \left\| \left(W_\Gamma D_\Gamma^T\right) g_i \cdot w_i \right\|_1 .$$

The first term on the right hand side of the inequality corresponds to the $1,1$-norm of the inverse Gram matrix of the weighed sub-dictionary of optimal functions. This can be expressed as:

$$\left(\left(W_\Gamma D_\Gamma^T\right) \left(W_\Gamma D_\Gamma^T\right)^T \right)^{-1} = \left(I + A_w\right)^{-1}, \tag{B.4}$$

where I denotes the identity matrix and A_w is a symmetric matrix. Due to the diagonal weight matrices W_Γ, the matrix A_w is not composed only of the off-diagonal elements. Adding and subtracting the identity matrix, we can rewrite (B.4) in the following way:

$$\left(I + A_w\right)^{-1} = \left(I + \left(\left(\left(W_\Gamma D_\Gamma^T\right) \left(W_\Gamma D_\Gamma^T\right)^T \right) - I \right) \right)^{-1}.$$

Akin to [174] this can be expanded by means of *Neumann* series [105] and, if $\|A_w\|_{1,1}<1$, we have:

$$\begin{aligned} \left\| \left(I + A_w\right)^{-1} \right\|_{1,1} &= \left\| \sum_{k=0}^{\infty} \left(-A_w\right)^k \right\|_{1,1} \\ &\leq \sum_{k=0}^{\infty} \|A_w\|_{1,1}^k = \frac{1}{1 - \|A_w\|_{1,1}}. \end{aligned}$$

Thus,

$$\left\| \left(\left(W_\Gamma D_\Gamma^T\right) \left(W_\Gamma D_\Gamma^T\right)^T \right)^{-1} \right\|_{1,1} \leq \frac{1}{1 - \|A_w\|_{1,1}}. \tag{B.5}$$

The $1,1$-norm of A_w can be expressed as:

$$\|A_w\|_{1,1} = \sup_{g_\gamma \in D_\Gamma} \left[\sum_{l \neq \gamma} | < g_l, g_\gamma > | \cdot w_l \cdot w_\gamma + |1 - w_\gamma^2| \right], \tag{B.6}$$

where the summation comes from the off-diagonal elements and the last term comes from the diagonal part. Note that for convergence of the *Neumann* series we need $\|A_w\|_{1,1}<1$. This is ensured by hypothesis since $\|A_w\|_{1,1} \leq \mu_1^w(m-1)+\epsilon_{max}$ and

$$\mu_1^w(m-1)+\epsilon_{max} < 1$$

by (5.10) and (5.11). From (B.5) it follows that:

$$\begin{aligned} &\left\| \left(\left(W_\Gamma D_\Gamma^T\right) \left(W_\Gamma D_\Gamma^T\right)^T \right)^{-1} \right\|_{1,1} \\ &\leq \frac{1}{1-(\mu_1^w(m-1)+\epsilon_{max})}. \end{aligned} \tag{B.7}$$

Coming back to Eq. (B.3), the second term can be bounded as

$$\sup_{g_i \in D_\Gamma} \left\| \left(W_\Gamma D_\Gamma^T\right) g_i \cdot w_i \right\|_1 \leq \mu_1^w(m). \tag{B.8}$$

Finally, from (B.7) and (B.8) we obtain

$$\frac{\mu_1^w(m)}{1-(\mu_1^w(m-1)+\epsilon_{max})} < \alpha, \tag{B.9}$$

and this proves the theorem.

■

B.3 Proof of Theorem 5.4

Let us consider first two preliminary lemmas that will be used in the proof of this theorem. These correspond to those used in the methodology appearing in [95], but are adapted to the case where *a priori* information is used.

Lemma B.1 *Let Γ be an optimal set, with $|\Gamma| = m$, associated to the exact sparse expansion of signal f and the reliable* a priori *knowledge weighting matrix $W(f,\mathcal{D})$ (and the W_Γ sub-matrix). Then, the square singular values ($\sigma^2_{min_w}$) of the matrix $(D_\Gamma W_\Gamma)$ are such that:*

$$\sigma^2_{min_w} \geq 1-\mu_1^w(m-1)-\epsilon_{max}. \tag{B.10}$$

Note that if $\epsilon_{max} \ll 1$, then $\sigma^2_{min_w} \gtrsim 1-\mu_1^w(m-1)$, which mimics the result of classic Weak-*MP [95].*

Proof: Consider the Gram matrix $G \triangleq (D_\Gamma W_\Gamma)^T (D_\Gamma W_\Gamma)$, then the singular values $\sigma^2_{k_w}$ are the eigenvalues (λ_k) of G. From the Geršgorin Disk Theorem [105] an upper bound on the eigenvalues of λ_k can be drawn in the way performed in [54, 82, 94, 95, 142]. This shows that every eigenvalue of G lies in one of the m disks

$$\Delta_k = \left\{ z : |G_{kk}-z| \leq \sum_{j\neq k} |G_{jk}| \right\}. \tag{B.11}$$

Hence, since $\sum_{j\neq k}|G_{jk}| = \sum_{j\neq k} |< w_j \cdot g_j, w_k \cdot g_k >| \leq \mu_1^w(m-1)$ then:

$$|G_{kk}-\lambda_k| \leq \mu_1^w(m-1), \tag{B.12}$$

where $G_{kk} \geq 1 - \epsilon_{max}$ and since $\mu_1^w(m-1) + \epsilon_{max} < 1$,

$$\sigma^2_{min_w} \geq 1 - \epsilon_{max} - \mu_1^w(m-1). \tag{B.13}$$

■

Lemma B.2 *For any index set Γ of $|\Gamma| = m$, the corresponding data dependent weighting matrix W_Γ and a coefficients vector $\mathbf{b}$,*

$$\sup_{\gamma\in\Gamma} |< D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma >| \geq \frac{\|D_\Gamma \mathbf{b}\|^2}{\left\|W_\Gamma^{-1}\mathbf{b}\right\|_1}. \tag{B.14}$$

Moreover, given a residual $r_k = f - f_k$ such that $r_k \in span\,(g_\gamma, \gamma \in \Gamma)$ and the smallest square singular value of D_Γ ($\sigma^2_{min_w}$) then,

$$\frac{\sup_{\gamma\in\Gamma} |< D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma >|}{\|r_k\|} \geq \sqrt{\frac{\sigma^2_{min_w}}{m}}. \tag{B.15}$$

For the sake of clarity of the section, the proof of this lemma is included in the Appendix.

Proof: To prove this lemma, we just need to follow the procedure appearing in [95, 142] which uses results from DeVore and Temlyakov [45].

$$\begin{aligned}
\|D_\Gamma \mathbf{b}\|^2 = & \left\langle D_\Gamma \mathbf{b}, D_\Gamma W_\Gamma W_\Gamma^{-1} \mathbf{b} \right\rangle \\
= & \sum_{\gamma\in\Gamma} \frac{b_\gamma}{w_\gamma} \left\langle g_\gamma \cdot w_\gamma, D_\Gamma \mathbf{b} \right\rangle \\
\leq & \sum_{\gamma\in\Gamma} \left|\frac{b_\gamma}{w_\gamma}\right| \left|\langle D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma \rangle\right| \\
\leq & \left\|W_\Gamma^{-1}\mathbf{b}\right\|_1 \sup_{\gamma\in\Gamma} \left|\langle D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma \rangle\right|.
\end{aligned} \tag{B.16}$$

For the final proof of the Lemma two additional results are needed.

- By the *Jensen's Inequality* [90] $\left\|W_\Gamma^{-1}\mathbf{b}\right\|_1$ can be bounded as

$$\left\|W_\Gamma^{-1}\mathbf{b}\right\|_1^2 \leq m \cdot \left\|W_\Gamma^{-1}\mathbf{b}\right\|_2^2. \tag{B.17}$$

  In fact:

$$\begin{aligned}
\left\|W_\Gamma^{-1}\mathbf{b}\right\|_1^2 = & \left(\sum_{i=0}^{m-1} \left|\frac{b_i}{w_i}\right|\right)^2 = m^2 \left(\sum_{i=0}^{m-1} \frac{|b_i|}{m \cdot w_i}\right)^2 \\
\leq & \; m^2 \sum_{i=0}^{m-1} \left|\frac{b_i}{w_i}\right|^2 \frac{1}{m} \leq m \cdot \left\|W_\Gamma^{-1}\mathbf{b}\right\|_2^2.
\end{aligned}$$

- By means of the *Singular Value Decomposition* [87] any $\left\|W_\Gamma^{-1}\mathbf{b}_k\right\|_2^2$ (where k indicates the iteration number) can be bounded as

$$\frac{\|r_k\|^2}{\sigma^2_{min_w}} \geq \left\|W_\Gamma^{-1}\mathbf{b}_k\right\|_2^2. \tag{B.18}$$

This is proved by:

$$\begin{aligned}
\|r_k\|^2 = & \|D_\Gamma \mathbf{b}_k\|^2 \\
= & \mathbf{b}_k^T \left(D_\Gamma W_\Gamma W_\Gamma^{-1}\right)^T D_\Gamma W_\Gamma W_\Gamma^{-1} \mathbf{b}_k \\
= & \mathbf{b}_k^T \left(W_\Gamma^{-1} W_\Gamma D_\Gamma^T\right) D_\Gamma W_\Gamma W_\Gamma^{-1} \mathbf{b}_k \\
= & \mathbf{b}_k^T W_\Gamma^{-1} \left(D_\Gamma W_\Gamma\right)^T D_\Gamma W_\Gamma W_\Gamma^{-1} \mathbf{b}_k \\
= & \mathbf{b}_k^T W_\Gamma^{-1} \left(U_{\Gamma_W} \Sigma_{\Gamma_W} V_{\Gamma_W}^T\right)^T ... \\
& \left(U_{\Gamma_W} \Sigma_{\Gamma_W} V_{\Gamma_W}^T\right) W_\Gamma^{-1} \mathbf{b}_k \\
= & \mathbf{b}_k^T W_\Gamma^{-1} \left(V_{\Gamma_W} \Sigma_{\Gamma_W}^T U_{\Gamma_W}^T\right) ... \\
& \left(U_{\Gamma_W} \Sigma_{\Gamma_W} V_{\Gamma_W}^T\right) W_\Gamma^{-1} \mathbf{b}_k,
\end{aligned}$$

where U_{Γ_W} and V_{Γ_W} are orthonormal matrices and Σ_{Γ_W} is a diagonal matrix such that

$$diag\left(\Sigma_{\Gamma_W}\right) = \left(\sigma_{0_w}, \sigma_{1_w}, ..., \sigma_{k_w}, ..., \sigma_{m_w}\right).$$

From now on consider $\mathbf{y} = V_{\Gamma_W}^T W_\Gamma^{-1} \mathbf{b}_k$. Therefore,

$$\begin{aligned}
\|r_k\|^2 = & \mathbf{b}_k^T W_\Gamma^{-1} V_{\Gamma_W} \Sigma_{\Gamma_W}^2 V_{\Gamma_W}^T W_\Gamma^{-1} \mathbf{b}_k \\
= & \mathbf{y}^T \Sigma_{\Gamma_W}^2 \mathbf{y} = \sum_{k=0}^{m-1} \sigma_{k_w}^2 \cdot y_k^2 \\
\geq & \sigma_{min_w}^2 \|\mathbf{y}\|^2 = \sigma_{min_w}^2 \left\|W_\Gamma^{-1} \mathbf{b}_k\right\|^2.
\end{aligned}$$

Thus, finally from (B.17) and (B.18) it follows

$$\frac{\|r_k\|^2}{\sigma_{min_w}^2} \geq \frac{\|W_\Gamma^{-1} \mathbf{b}_k\|_1^2}{m}, \tag{B.19}$$

that jointly with (B.16) gives the result stated by Lemma B.2:

$$\frac{\sup_{\gamma \in \Gamma} |< D_\Gamma \mathbf{b}, g_\gamma \cdot w_\gamma >|}{\|r_k\|} \geq \sqrt{\frac{\sigma_{min_w}^2}{m}}.$$

■

Finally, Theorem 5.4 can be proved as follows:

Proof: Let $r_{k+1} = f - f_k$ be the residual of the Weighted-MP/OMP algorithm at the kth iteration, then it is known that:

$$\|r_{k+1}\|^2 \leq \|r_k\|^2 - \left|< r_k, g_{\gamma_{k+1}} >\right|^2, \tag{B.20}$$

where the inequality applies for OMP, while for the case of MP the equality holds. In our case the selection of $g_{\gamma_{k+1}}$ is driven by $W(f, \mathcal{D})$, i.e.

$$\left|\left\langle r_k, g_{\gamma_{k+1}} \right\rangle\right| = \alpha \cdot \frac{1}{w_\gamma} \sup_\gamma \left|\langle r_k, g_\gamma \cdot w_\gamma \rangle\right|. \tag{B.21}$$

Hence,

$$\begin{aligned} \|r_{k+1}\|^2 \leq & \|r_k\|^2 - \alpha^2 \cdot \frac{1}{w_\gamma^2} \sup_\gamma |< r_k, g_\gamma \cdot w_\gamma >|^2 \\ \leq & \|r_k\|^2 \left(1 - \alpha^2 \frac{\frac{1}{w_\gamma^2} \sup_\gamma |< r_k, g_\gamma \cdot w_\gamma >|^2}{\|r_k\|^2} \right). \end{aligned} \tag{B.22}$$

Then, from Eqs. (B.15), (B.10) and given $w_\gamma \leq 1$, it follows that:

$$\begin{aligned} \|r_{k+1}\|^2 \leq & \|r_k\|^2 \left(1 - \alpha^2 \frac{\sigma_{min_w}^2}{w_\gamma^2 m} \right) \\ \leq & \|r_k\|^2 \left(1 - \alpha^2 \frac{\sigma_{min_w}^2}{m} \right) \\ \leq & \|r_k\|^2 \left(1 - \alpha^2 \frac{1 - \mu_1^w (m-1) - \epsilon_{max}}{m} \right) \\ \leq & \|f\|^2 \left(1 - \alpha^2 \frac{1 - \mu_1^w (m-1) - \epsilon_{max}}{m} \right)^{k+1}. \end{aligned} \tag{B.23}$$

■

B.4 Proof of Theorem 5.5

Proof: To demonstrate the result of Theorem 5.5, we follow the steps of the original proofs by *Tropp* [175] and *Gribonval and Vandergheynst* [95]. This time however, *a priori* knowledge is taken into account. First of all, let us remind the following statements:

- $f_m^{opt} \in span(\Gamma_m)$
- $r_k = f - f_k$
- $r_m^{opt} = f - f_m^{opt}$ is such that $r_m^{opt} \perp (f_m^{opt} - f_k) \ \forall \ 0 \leq k < m$, hence $\|r_k\|_2^2 = \|f_m^{opt} - f_k\|_2^2 + \|f - f_m^{opt}\|_2^2$.

In order to ensure the recovery of any atom belonging to the optimal set $\Gamma = \Gamma_m$, the following needs to be satisfied:

$$\rho^w(r_k) = \frac{\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot r_k\|_\infty}{\|D_\Gamma \cdot W_\Gamma \cdot r_k\|_\infty} < \alpha, \tag{B.24}$$

where $\alpha \in (0,1]$ is the weakness factor [173]. To establish (5.15), the previous expression has to be put in terms of f_m^{opt} and f_k. Hence,

$$\begin{aligned}
\rho^w(r_k) &= \frac{\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot r_k\|_\infty}{\|D_\Gamma \cdot W_\Gamma \cdot r_k\|_\infty} = \frac{\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot (f - f_k)\|_\infty}{\|D_\Gamma \cdot W_\Gamma \cdot (f - f_k)\|_\infty} \\
&= \frac{\left\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot \left(f - f_m^{opt}\right) + D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f - f_m^{opt}\right) + D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty} \\
&= \frac{\left\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot \left(f - f_m^{opt}\right) + D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty} \\
&\leq \frac{\left\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot \left(f - f_m^{opt}\right)\right\|_\infty}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty} + \frac{\left\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty} \\
&= \frac{\left\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot \left(f - f_m^{opt}\right)\right\|_\infty}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty} + \rho^w\left(f_m^{opt} - f_k\right),
\end{aligned} \tag{B.25}$$

where the second term can be upper bounded since $(f_m^{opt} - f_k) \in span(\Gamma)$ [49],

$$\rho^w\left(f_m^{opt} - f_k\right) \leq \frac{\mu_1^w(m)}{1 - (\mu_1^w(m-1) + \epsilon_{max})}. \tag{B.26}$$

The first term of the last equality in (B.25) can be upper bounded in the following way:

$$\frac{\left\|D_{\overline{\Gamma}} \cdot W_{\overline{\Gamma}} \cdot \left(f - f_m^{opt}\right)\right\|_\infty}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty} = \frac{\sup\limits_{\gamma \in \overline{\Gamma}} \left|\left\langle g_\gamma \cdot w_\gamma, \left(f - f_m^{opt}\right)\right\rangle\right|}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty}, \tag{B.27}$$

and by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\frac{\sup\limits_{\gamma \in \overline{\Gamma}} \left|\left\langle g_\gamma \cdot w_\gamma, \left(f - f_m^{opt}\right)\right\rangle\right|}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty} &\leq \frac{\sup\limits_{\gamma \in \overline{\Gamma}} \|g_\gamma \cdot w_\gamma\|_2 \cdot \left\|f - f_m^{opt}\right\|_2}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty} \\
= \frac{\sup\limits_{\gamma \in \overline{\Gamma}} |w_\gamma| \cdot \left\|f - f_m^{opt}\right\|_2}{\left\|D_\Gamma \cdot W_\Gamma \cdot \left(f_m^{opt} - f_k\right)\right\|_\infty} &= \frac{\sup\limits_{\gamma \in \overline{\Gamma}} |w_\gamma| \cdot \left\|f - f_m^{opt}\right\|_2}{\sup\limits_{\gamma \in \overline{\Gamma}} \left|\left\langle g_\gamma \cdot w_\gamma, \left(f_m^{opt} - f_k\right)\right\rangle\right|}.
\end{aligned} \tag{B.28}$$

In order to further upper bound the expression above, the denominator can be lower bounded, as shown in [49]. Indeed, by the singular value decomposition:

$$\sup_{\gamma \in \overline{\Gamma}} \left|\left\langle g_\gamma \cdot w_\gamma, \left(f_m^{opt} - f_k\right)\right\rangle\right| \geq \sqrt{\frac{\sigma^2_{min_w}}{m}} \left\|f_m^{opt} - f_k\right\|_2, \tag{B.29}$$

where $\sigma^2_{min_w}$ is the minimum of the squared singular values of $G \triangleq \left(D_\Gamma W_\Gamma\right)^T \left(D_\Gamma W_\Gamma\right)$, and can be bounded as $\sigma^2_{min_w} \geq 1 - \epsilon_{max} - \mu_1^w(m-1)$. Moreover, in (B.28), $\|f - f_m^{opt}\|_2$ can be defined as $\|f - f_m^{opt}\|_2 = (1+\eta) \cdot \|r_m^{opt}\|_2$, where $\eta \geq 0$ stands for a sub-optimality factor which indicates whether f_m^{opt} can be reached and, if not possible (i.e. $\eta \neq 0$), sets the best possible reachable approximation error. Hence, (B.28) can be rewritten as:

$$\frac{\sup\limits_{\gamma \in \overline{\Gamma}} |w_\gamma| \cdot \left\|f - f_m^{opt}\right\|_2}{\sup\limits_{\gamma \in \overline{\Gamma}} \left|\left\langle g_\gamma \cdot w_\gamma, \left(f_m^{opt} - f_k\right)\right\rangle\right|} \leq \frac{\sup\limits_{\gamma \in \overline{\Gamma}} |w_\gamma| \cdot (1+\eta) \cdot \left\|r_m^{opt}\right\|_2}{\sqrt{\frac{1 - \mu_1^w(m-1) - \epsilon_{max}}{m}} \left\|f_m^{opt} - f_k\right\|_2}. \tag{B.30}$$

Thus, from (B.30) and (B.26), a sufficient condition for the recovery of a correct atom can be expressed as:

$$\begin{aligned}\rho^w(r_k) &\leq \frac{\sup_{\gamma\in\overline{\Gamma}} |w_\Gamma| \cdot (1+\eta) \cdot \left\|r_m^{opt}\right\|_2}{\sqrt{\frac{1-\mu_1^w(m-1)-\epsilon_{max}}{m}}\left\|f_m^{opt}-f_k\right\|_2} + \frac{\mu_1^w(m)}{1-\mu_1^w(m-1)-\epsilon_{max}} \\ &= \frac{w_{\overline{\Gamma}}^{max}\sqrt{(1-\mu_1^w(m-1)-\epsilon_{max})\,m}\cdot(1+\eta)\cdot\left\|r_m^{opt}\right\|_2 + \left\|f_m^{opt}-f_k\right\|_2\mu_1^w(m)}{(1-\mu_1^w(m-1)-\epsilon_{max})\left\|f_m^{opt}-f_k\right\|_2} < \alpha.\end{aligned} \tag{B.31}$$

Considering that $\|f_m^{opt}-f_k\|_2^2 = \|r_k\|_2^2 - \|r_m^{opt}\|_2^2$, it easily follows that

$$\frac{w_{\overline{\Gamma}}^{max}\sqrt{(1-\mu_1^w(m-1)-\epsilon_{max})\,m}\cdot(1+\eta)\cdot\left\|r_m^{opt}\right\|_2 + \sqrt{\|r_k\|_2^2-(1+\eta)^2\left\|r_m^{opt}\right\|_2^2}\mu_1^w(m)}{(1-\mu_1^w(m-1)-\epsilon_{max})\sqrt{\|r_k\|_2^2-(1+\eta)^2\left\|r_m^{opt}\right\|_2^2}} < \alpha. \tag{B.32}$$

Then, we solve for $\|r_k\|_2^2$:

$$\|r_k\|_2^2 > (1+\eta)^2\left\|r_m^{opt}\right\|_2^2\left(1+\frac{\left(w_{\overline{\Gamma}}^{max}\right)^2(1-\mu_1^w(m-1)-\epsilon_{max})}{(\alpha\,(1-\mu_1^w(m-1)-\epsilon_{max})-\mu_1^w(m))^2}\right). \tag{B.33}$$

For simplicity, let us consider the case where a full search atom selection algorithm is available. Thus, replacing $\alpha = 1$ in (B.33) proves Theorem 5.5.

■

B.5 Proof of Corollary 5.3

Proof: For simplicity, let us use an upper bound on the left hand side of Eq. (5.16). Indeed, the factor $0 < (w_{\overline{\Gamma}^{max}})^2 \leq 1$ is removed:

$$\left(1+\frac{m\,(1-(\mu_1^w(m-1)+\epsilon_{max}))}{(1-(\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max}))^2}\right) \leq \left(1+\frac{m\,(1-(\mu_1^w(m-1)+\epsilon_{max}))}{(1-(\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max}))^2}\right).$$

Let us suppose the *a priori* knowledge in use is reliable. Then the following relations can be assumed:

$$\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max} \leq \mu_1(m-1)+\mu_1(m) < 1 \tag{B.34}$$

and

$$\mu_1^w(m-1)+\epsilon_{max} \leq \mu_1(m-1). \tag{B.35}$$

Now we can proof the inequality. Let us make the hypothesis that the following is true:

$$\frac{(1-(\mu_1^w(m-1)+\epsilon_{max}))}{(1-(\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max}))^2} \leq \frac{(1-\mu_1(m-1))}{(1-(\mu_1(m-1)+\mu_1(m)))^2}.$$

Then,

$$1 \leq \frac{(1-\mu_1(m-1))}{(1-(\mu_1^w(m-1)+\epsilon_{max}))}\cdot\frac{(1-(\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max}))^2}{(1-(\mu_1(m-1)+\mu_1(m)))^2}. \tag{B.36}$$

According to (B.34) and (B.35), the following can be considered:

$$\tau_1 \triangleq \mu_1(m-1) - (\mu_1^w(m-1) + \epsilon_{max}) \tag{B.37}$$

$$\tau_2 \triangleq \mu_1(m-1) + \mu_1(m) - (\mu_1^w(m) + \mu_1^w(m-1) + \epsilon_{max}) \tag{B.38}$$

where $0 \leq \tau_1 \ll \mu_1(m)$ and $0 \leq \tau_2 \ll \mu_1(m-1)+\mu_1(m)$. Hence, being $\delta \triangleq \tau_2/\left(1-(\mu_1(m-1)+\mu_1(m))\right)$ the second fraction in (B.36) can be substituted:

$$1 \leq \frac{(1-\mu_1(m-1))}{(1-(\mu_1^w(m-1)+\epsilon_{max}))} \cdot (1+\delta)^2 . \tag{B.39}$$

Moreover, being $\delta' \triangleq \tau_1/\left(1-\mu_1(m-1)\right)$, the remaining fractional term of (B.39) may be considered such that

$$1 \leq \frac{1}{1+\delta'} \cdot (1+\delta)^2 = (1+\delta) \cdot \frac{1+\delta}{1+\delta'}. \tag{B.40}$$

From this, clearly $(1+\delta) \geq 1$. So, if $(1+\delta) \geq (1+\delta')$ then Corollary 5.3 is proved.

Hence, let us check, finally, if this last condition holds. Inserting in $(1+\delta) \geq (1+\delta')$ the definitions of δ and δ' we find:

$$\frac{\tau_2}{(1-(\mu_1(m-1)+\mu_1(m)))} \geq \frac{\tau_1}{(1-\mu_1(m-1))},$$

which will be always true if

$$\frac{\tau_2}{\tau_1} \geq 1.$$

Let us assume, then, that $\tau_2 \geq \tau_1$. This, together with (B.37) and (B.38), yields:

$$(\mu_1(m) - \mu_1^w(m)) \geq 0,$$

which asserts all hypothesis and concludes the whole proof.

■

B.6 Proof of Theorem 5.6

Proof: Let us consider k such that $\|r_k\|_2^2$ satisfies Eq. (5.15) of Theorem 5.5. Then, it is known that for *Weak*-MP:

$$\|r_{k-1}\|_2^2 - \|r_k\|_2^2 \geq |\langle r_k, g_{\gamma_k}\rangle|^2 , \tag{B.41}$$

where the inequality applies for OMP, while in the case of MP the equality holds. Moreover, considering the weighted selection, then

$$\|r_{k-1}\|_2^2 - \|r_k\|_2^2 \geq \alpha \sup_{\gamma\in\Gamma} |\langle r_k, g_\gamma \cdot w_\gamma\rangle|^2 \frac{1}{w_\gamma^2} = \alpha \sup_{\gamma\in\Gamma} \left|\left\langle f_m^{opt} - f_k, g_\gamma \cdot w_\gamma\right\rangle\right|^2 \frac{1}{w_\gamma^2}, \tag{B.42}$$

where the last equality follows from the assumption that Eq. (5.15) of Theorem 5.5 is satisfied and because $(f - f_m^{opt}) \perp span(\Gamma)$. And by (B.29),

$$\|r_{k-1}\|_2^2 - \|r_k\|_2^2 \geq \frac{\alpha}{w_\gamma^2} \frac{\sigma_{min_w}^2}{m} \left\|f_m^{opt} - f_k\right\|_2^2 . \tag{B.43}$$

As stated before, $\|f_m^{opt} - f_k\|_2^2 = \|r_k\|_2^2 - \|r_m^{opt}\|_2^2$, hence $\|f_m^{opt} - f_{k-1}\|_2^2 - \|f_m^{opt} - f_k\|_2^2 = \|r_{k-1}\|_2^2 - \|r_k\|_2^2$, which together with (B.43) gives:

$$\left\|f_m^{opt} - f_k\right\|_2^2 \leq \left\|f_m^{opt} - f_{k-1}\right\|_2^2 \left(1 - \frac{\alpha}{w_\gamma^2}\frac{\sigma_{min_w}^2}{m}\right) \leq \left\|f_m^{opt} - f_{k-1}\right\|_2^2 \left(1 - \alpha\frac{\sigma_{min_w}^2}{m}\right). \tag{B.44}$$

Finally, by simply considering $0 \leq l \leq k$ by recursion it follows:

$$\|r_k\|_2^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2 \leq \left(1 - \alpha \frac{\sigma^2_{min_w}}{m}\right)^{k-l} \left(\|r_l\|_2^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2\right), \tag{B.45}$$

and the Theorem is proved.

■

B.7 Proof of Theorem 5.7

In order to prove Theorem 5.7, several intermediate results are necessary. These follow easily by taking into account the use of *a priories* in the proofs of Theorem 7 in [95]. The detailed procedure is described in the following.

Lemma B.3 *Let $W(f, \mathcal{D})$ be a reliable* a priori *information and $\{r_k\} : k \geq 0$ a sequence of residuals produced by Weighted-MP/OMP, then as long as $\|r_k\|_2^2$ satisfies Eq.* (5.15) *of Theorem 5.5, for any $1 \leq k < m$ such that $N_k < N_m$,*

$$N_m - N_k < 1 + \frac{m}{1 - \mu_1^w (m-1) - \epsilon_{max}} \left[\log\left(\frac{\left\|r_k^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2}\right) + \log\left(\frac{1+\lambda_k^w}{1+\lambda_m^w}\right)\right], \tag{B.46}$$

where

$$\lambda_l^w \triangleq \frac{l\left(1 - (\mu_1^w(l-1) + \epsilon_{max})\right) \cdot \left(w_{\Gamma_l}^{max}\right)^2}{\left[1 - (\mu_1^w(l-1) + \mu_1^w(l) + \epsilon_{max})\right]^2}, \tag{B.47}$$

in which l corresponds to the size of a particular optimal set of atoms ($l = |\Gamma_l|$).

Proof: From Theorem 5.6, it follows that for $l = N_k$, $k = N_m - 1$, defining

$$\beta_l^w \triangleq 1 - \frac{1 - \mu_1^w(l-1) - \epsilon_{max}}{l}, \tag{B.48}$$

where l is defined as in (B.47), and starting from the condition in the residual $\left\|r_{N_{m-1}}\right\|_2^2$ as defined in the Definition 5.4, the following is accomplished if $\alpha = 1$:

$$\begin{aligned}
\lambda_m^w (1+\eta)^2 \left\|r_m^{opt}\right\|_2^2 &< \left\|r_{N_{m-1}}\right\|_2^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2 \\
&\leq (\beta_m^w)^{N_m - 1 - N_k} \cdot \left(\|r_{N_k}\|_2^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2\right) \\
&\leq (\beta_m^w)^{N_m - 1 - N_k} \cdot \left((1+\lambda_k)\left\|r_k^{opt}\right\|_2^2 (1+\eta)^2 - \left\|r_m^{opt}\right\|_2^2 (1+\eta)^2\right)
\end{aligned} \tag{B.49}$$

Operating on (B.49) as in [95], it easily follows that:

$$\left(\frac{1}{\beta_m^w}\right)^{N_m - 1 - N_k} < \frac{\left\|r_k^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2} \frac{1+\lambda_k^w}{1+\lambda_m^w},$$

thus,

$$N_m - 1 - N_k \log\left(\frac{1}{\beta_m^w}\right) < \log\left[\frac{\left\|r_k^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2} \frac{1+\lambda_k^w}{1+\lambda_m^w}\right].$$

If $t \geq 0$ then $\log(1-t) \leq -t$ and so

$$\frac{1}{\log\left(\frac{1}{\beta_m^w}\right)} \leq \frac{m}{1-\mu_1(m-1)-\epsilon_{max}}.$$

This proves the result presented in (B.46) and so the Lemma.

■

In order to use Lemma B.3 in Theorem 5.7, an estimate of the argument of the second logarithm in (B.46) is necessary. This can be found in the following Lemma.

Lemma B.4 *For all m such that $\mu_1^w(m-1)+\mu_1^w(m)+\epsilon_{max}<1$ and $1 \leq k < m$, we have:*

$$\lambda_m^w \geq m \cdot \left(w_{\overline{\Gamma}}^{max}\right)^2 \tag{B.50}$$

$$\frac{\lambda_k^w}{\lambda_m^w} \leq \frac{k}{m} \cdot \frac{(1-\mu_1^w(k-1)-\epsilon_{max_k}) \cdot \left(w_{\overline{\Gamma}_k}^{max}\right)^2}{(1-\mu_1^w(m-1)-\epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max}\right)^2} \tag{B.51}$$

Proof: Consider the definition of λ_m^w of (B.47). Then since $\mu_1^w(l-2)+\mu_1^w(l-1)+\epsilon_{max} \leq \mu_1^w(l-1)+\mu_1^w(l)+\epsilon_{max}$ for $2 \leq l \leq m$, the following can be stated:

$$\frac{\lambda_{l-1}^w}{\lambda_l^w} \leq \frac{l-1}{l} \cdot \frac{\left(1-\mu_1^w(l-2)-\epsilon_{max_{l-1}}\right) \cdot \left(w_{\overline{\Gamma}_{l-1}}^{max}\right)^2}{(1-\mu_1^w(l-1)-\epsilon_{max_l}) \cdot \left(w_{\overline{\Gamma}_l}^{max}\right)^2}. \tag{B.52}$$

By assuming $k+1 \leq l \leq m$ the Lemma is proved.

■

Finally, building on the results obtained from Theorem 5.6 and Lemmas B.3 and B.4, Theorem 5.7 can be proved.

Proof: To prove Theorem 5.7 we need to upper bound the factor $\frac{1+\lambda_k^w}{1+\lambda_m^w}$ in Eq. (B.46). For this purpose let us consider the following:

$$\frac{(1+\lambda_k^w)}{(1+\lambda_m^w)} \leq \frac{(1+\lambda_k^w)}{(\lambda_m^w)} \leq \frac{1}{\lambda_m^w} + \frac{\lambda_k^w}{\lambda_m^w}. \tag{B.53}$$

Together with the results of Lemma B.4, it gives:

$$\log\left(\frac{1+\lambda_k^w}{1+\lambda_m^w}\right) \leq \log\left(\frac{1}{m} + \frac{k}{m} \cdot \frac{(1-\mu_1^w(k-1)-\epsilon_{max_k}) \cdot \left(w_{\overline{\Gamma}_k}^{max}\right)^2}{(1-\mu_1^w(m-1)-\epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max}\right)^2}\right). \tag{B.54}$$

Hence, using Eq. (B.46) we obtain

$$\begin{aligned} N_m - N_k < 1 + \frac{m}{1-\mu_1^w(m-1)-\epsilon_{max}} \Bigg[& \log\left(\frac{\left\|r_k^{opt}\right\|_2^2}{\left\|r_m^{opt}\right\|_2^2}\right) + \cdots \\ & \log\left(\frac{1}{m} + \frac{k}{m} \cdot \frac{(1-\mu_1^w(k-1)-\epsilon_{max_k}) \cdot \left(w_{\overline{\Gamma}_k}^{max}\right)^2}{(1-\mu_1^w(m-1)-\epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max}\right)^2}\right)\Bigg]. \end{aligned} \tag{B.55}$$

Theorem 5.7 is thus proved by particularizing the previous expression for the case where $k = 1$. For the case of $N_m \geq N_1 + 1 = 2$, this yields that

$$\begin{aligned} N_m - N_1 &< 1 + \frac{m}{1 - \mu_1^w (m-1) - \epsilon_{max}} \cdot \\ &\quad \left[\log \left(\frac{\left\| r_1^{opt} \right\|_2^2}{\left\| r_m^{opt} \right\|_2^2} \right) + \log \left(\frac{1}{m} + \frac{(1 - \epsilon_{max_1}) \cdot \left(w_{\overline{\Gamma}_1}^{max} \right)^2}{m \left(1 - \mu_1^w (m-1) - \epsilon_{max_m}\right) \cdot \left(w_{\overline{\Gamma}_m}^{max} \right)^2} \right) \right] \cdot \\ &< 1 + \frac{m}{1 - \mu_1^w (m-1) - \epsilon_{max}} \cdot \\ &\quad \left[\log \left(\frac{\left\| r_1^{opt} \right\|_2^2}{\left\| r_m^{opt} \right\|_2^2} \right) + \log \left(\frac{1}{m} + \frac{1}{m \left(1 - \mu_1^w (m-1) - \epsilon_{max_m}\right) \cdot \left(w_{\overline{\Gamma}_m}^{max} \right)^2} \right) \right] \end{aligned} \tag{B.56}$$

Which, since $\mu_1^w (m-1) < \frac{1 - \epsilon_{max}}{2}$ and $1 - \mu_1^w (m-1) - \epsilon_{max} > \frac{1 - \epsilon_{max}}{2}$, then

$$\begin{aligned} N_m &< 2 + \frac{m}{1 - \mu_1^w (m-1) - \epsilon_{max}} \left[\log \left(\frac{\left\| r_1^{opt} \right\|_2^2}{\left\| r_m^{opt} \right\|_2^2} \right) + \log \left(\frac{2 + (1 - \epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max} \right)^2}{m \left(1 - \epsilon_{max_m}\right) \cdot \left(w_{\overline{\Gamma}^m}^{max} \right)^2} \right) \right] \\ &< 2 + \frac{m}{1 - \mu_1^w (m-1) - \epsilon_{max}} \left[\log \left(\frac{\left\| r_1^{opt} \right\|_2^2}{\left\| r_m^{opt} \right\|_2^2} \right) + \log \left(\frac{3}{m \left(1 - \epsilon_{max_m}\right) \cdot \left(w_{\overline{\Gamma}_m}^{max} \right)^2} \right) \right] \\ &< 2 + \frac{m}{1 - \mu_1^w (m-1) - \epsilon_{max}} \log \left(\frac{3 \left\| r_1^{opt} \right\|_2^2}{m \cdot \left\| r_m^{opt} \right\|_2^2 (1 - \epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max} \right)^2} \right) \\ &< 2 + \frac{2 \cdot m}{1 - \epsilon_{max}} \log \left(\frac{3 \left\| r_1^{opt} \right\|_2^2}{m \cdot \left\| r_m^{opt} \right\|_2^2 (1 - \epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max} \right)^2} \right). \end{aligned} \tag{B.57}$$

This is only possible if $3 \left\| r_1^{opt} \right\|_2^2 \geq m \cdot \left\| r_m^{opt} \right\|_2^2 (1 - \epsilon_{max_m}) \cdot \left(w_{\overline{\Gamma}_m}^{max} \right)^2$.

■

Appendix C

Analysis of Block Dictionaries Influence on Video Representations

C.1 Proof of Theorem 7.1

Consider the situation posed in sec. 7.4.1 where the dictionary $\mathcal{D}$ is the union of several sub-dictionaries such that:

$$\mathcal{D} = \bigcup_{i=0}^{N-1} D_{B_i}. \tag{C.1}$$

Let f be a function such that

$$f \in span\left(g_{\gamma_{B_i}} : i \in [0, m-1],\ g_{\gamma_{B_i}} \in \mathcal{D}_{B_i}\right), \tag{C.2}$$

i.e. f can be expressed as a linear combination of atoms $g_{\gamma_{B_i}}$ where no more than one primitive is taken from each dictionary block $\mathcal{D}_{B_i}$. This is a very restrictive situation. However several examples can be found in practice as depicted in Sec. 7.4.1 where this situation may apply. Given the additional constraints imposed to the dictionary and the signal f, a refinement of the exact recovery condition (Stability Condition) defined in theorem 4.1 can be established.

A new measure on the coherence can thus be introduced where the block based division of the dictionary is taken into account. Borrowing ideas from [142] where the same coherence measure was used for a similar situation, we define the *Babel block* function $\mu_{1_B}(m)$.

Definition C.1 *([142]) Let $\mathcal{D} = \bigcup_{i=0}^{N-1} D_{B_i}$ denote a block dictionary and Γ the set of sub-blocks from where, at most a function is taken from each, then the cumulative coherence function $\mu_{1_B}(m)$ is*

$$\mu_{1_B}(m) \triangleq \max_{\Gamma|\ \|\Gamma\|_0 = m} \max_{j \notin \Gamma, l} \sum_{i \in \Gamma} \max_{k} \left| < g_k^{B_i}, g_l^{B_j} > \right|, \tag{C.3}$$

This measure defines the worst cumulative dot product possible among two functions of different blocks for the worst selection of the Γ set of blocks.

We can now prove Theorem 7.1.

Proof: From (4.21) and following the procedure suggested in [174], it follows:

$$\begin{aligned}\sup_{g_\gamma \notin D_\Gamma} \left\| (D_\Gamma)^+ g_\gamma \right\|_1 &= \sup_{g_\gamma \notin D_\Gamma} \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} D_\Gamma^T g_\gamma \right\|_1 \\ &\leq \left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{1,1} \sup_{\gamma \notin D_\Gamma} \left\| D_\Gamma^T g_\gamma \right\|_1\end{aligned} \tag{C.4}$$

The first term that corresponds to the $1,1-norm$ of the inverse of the Gram matrix can be expressed as:

$$\left(D_\Gamma^T D_\Gamma\right)^{-1} = (I + A)^{-1}, \tag{C.5}$$

where I denotes the identity matrix and A all the off-diagonal components of the dictionary Gram matrix. Expanding (C.5) by means of Neumann series and using $\|A\|_{1,1}<1$ we have:

$$\left\| \left(D_\Gamma^T D_\Gamma\right)^{-1} \right\|_{1,1} = \left\| \sum_{k=0}^{\infty} (-A)^k \right\|_{1,1} \leq \sum_{k=0}^{\infty} \|A\|_{1,1}^k = \frac{1}{1 - \|A\|_{1,1}}. \tag{C.6}$$

$\|A\|_{1,1}$ is the biggest ℓ_1 norm column of matrix A and

$$\|A\|_{1,1} = \max_k \sum_{j \neq k} | < g_k, g_j > | \leq \mu_{1_B}(m-1). \tag{C.7}$$

The second term of (C.4) term can be upper bounded as follows:

$$\sup_{g_\gamma \notin D_\Gamma} \left\| D_\Gamma^T g_\gamma \right\|_1 = \sup_{g_\gamma \notin D_\Gamma} \sum_{g_l \in D_\Gamma} | < g_\gamma, g_l > | \leq \mu_{D_B} + \mu_{1_B}(m-1), \tag{C.8}$$

where μ_{D_B} denotes the maximum coherence (inner product) between two functions into a block. In order to be more explicit, (C.8) represents the worst possible case for the addition of cross products between the optimal set of m atoms and an atom not belonging to this set. This is, the $m-1$ worst possible cross products between atoms belonging to different blocks of $\mathcal{D}$ plus the worst possible coherence between two atoms belonging to the same block.

Hence, exact recovery of the the right set of atoms is ensured when:

$$\frac{\mu_{D_B} + \mu_{1_B}(m-1)}{1 - \mu_{1_B}(m-1)} < \alpha \tag{C.9}$$

■

Bibliography

[1] (1990;1993). *Video codec for audiovisual services at $p \times 64$ kbit/s.* Int. Telecommun. Union-Telecommun. (ITU-T), Recommendation H.261, version 1, 1990, version 2, 1993.

[2] (1993). *Coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s-Part 2: Video.* Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC) JTC 1, ISO/IEC 11 172-2 (MPEG-1).

[3] (1994). *Generic coding of moving pictures and associated audio information-Part 2: Video.* Int. Telecommun. Union-Telecommun. (ITU-T) and Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC) JTC1, Recommendation H.262 and ISO/IEC 13818-2 (MPEG-2 Video),.

[4] (1995;1998;2000). *Video coding for low bit rate communication.* Int. Telecommun. Union-Telecommun. (ITU-T), Recommendation H.263, version 1, 1995; version 2, 1998; version 3, 2000.

[5] (1999-2003). *Coding of audio-visual objects-Part 2: Visual.* Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC) JTC 1, ISO/IEC 14496-2 (MPEG-4 visual version 1),.

[6] (2003). *Advanced video coding for generic audiovisual services.* Int. Telecommun. Union-Telecommun. (ITU-T) and Int. Standards Org./Int. Electrotech. Comm. (ISO/IEC) JTC 1, Recommendation H.264 and ISO/IEC 14496-10 (MPEG-4) AVC.

[7] (2003). ITU-T Recommendation H.264 - ISO/IEC 14496-10 AVC: *Advanced Video Coding for Generic Audiovisual Services.* ITU-T and ISO/IEC JTC1.

[8] I. 15444-1:2000 (2000). *Information technology* - JPEG2000 *image coding system.*

[9] T. Aach, A. Kaup, R. Mester (1990). Combined displacement estimation and segmentation of stereo image pairs based on gibbs random fields. In *ICASSP.*

[10] Y. Altunbasak, A. M. Tekalp (1997). Closed-form connectivity-preserving solutions for motion compensation using 2-d meshes. *IEEE Trans. Image Processing* **6**(9):1255–1269.

[11] T. W. Anderson (1984). *An Introduction to Multivariate Statistical Analysis.* John Wiley & Sons, 2nd edn.

[12] Y. Andreopoulos et al. (2004). In-band motion compensated temporal filtering. *Signal Processing: Image Communication* **19**(7):653–673.

[13] J.-P. Antoine, R. Murenzi, P. Vandergheynst, S. Twareque Ali (2004). *Two-Dimensional Wavelets and their Relatives.* Cambridge, United Kingdom.

[14] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies (1992). Image coding using the wavelet transform. *IEEE Trans. Image Processing* **1**(2):205–220.

[15] J. L. Barron, D. J. Fleet, S. S. Beauchemin (1992). Performance of optical flow techniques. In *IEEE Int. Conf. of Computer Vision and Pattern Recognition*.

[16] A. O. Barut, R. R. (1987). *Theory of Group Representations and Applications*. ISBN: 997150216X. World Scientific Pub Co.

[17] D. Bernier, K. Taylor (1996). Wavelets from square-integrable representations. *SIAM Journal on Mathematical Analysis* **27**(2):594–608.

[18] D. P. Bertsekas (1999). *Nonlinear Programming*. Athena Scientific, http://www.athenasc.com/nonlinbook.html, 2nd edn.

[19] P. Besson, M. Kunt (2005). *Information theoretic optimization of audio features for multimodal speaker detection*. Tech. Rep. ITS-2005.008, LTS/ITS - EPFL.

[20] R. Bracewell (1999). *The Fourier Transform and Its Applications*, chap. Convolution Theorem, pp. 108–112. McGraw-Hill, New York, 3rd edn.

[21] S. Brofferio, F. Rocca (1977). Interframe redundancy reduction of video signals generated by translating objects. *IEEE Trans. Commun.* **25**:448–455.

[22] C. S. Burrus, T. W. Parks (1985). *DFT/FFT and Convolution Algorithms: Theory and Implementation*. John Wiley and Sons, New York.

[23] T. Butz, J.-P. Thiran (2002). Feature space mutual information in speech-video sequences. In *IEEE ICME*, vol. 2, pp. 361–364.

[24] M. Campani, A. Verri (1992). Motion analysis from first-order properties of optical. In *CVGIP: Image Understanding*.

[25] S. L. Campbell, C. D. Meyer (1979). *Generalized Inverses of Linear Transformations*. Pitman, London, U.K.

[26] E. J. Candès, D. Donoho (1999). Ridgelets: a key to higher dimensional intermittency? *Phil. Trans. R. Soc. Lond.* pp. 2495–2509.

[27] E. J. Candès, D. L. Donoho (1999). Curvelets - a surprisingly effective nonadaptive representation for objects with edges. In A.Cohen, C.Rabut, L.L.Schmaker (eds.), *Curve and Surface Fitting*, Vanderbilt University Press.

[28] S. S. Chen, D. L. Donoho, M. A. Saunders (1999). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comp.* **20**(1):33–61.

[29] S. Cho, W. Pearlman (2002). A full-featured, error-resilient, scalable wavelet video codec based on the set partitioning in hierarchical trees (spiht) algorithm. *IEEE Trans. Circuits Syst. Video Technol.* **12**(3):157 – 171.

[30] S.-J. Choi, J. Woods (1999). Motion-compensated 3-d subband coding of video. *IEEE Trans. Image Processing* **8**(2):155 – 167.

[31] A. Cohen, W. Dahmen, I. Daubechies, R. DeVore (2001). Tree approximation and optimal encoding. *Applied and Computational Harmonic Analysis* **11**(2):192–226.

[32] A. Cohen, I. Daubechies, O. Guleryuz, M. Orchard (2002). On the importance of combining wavelet-based non-linear approximation in coding strategies. *IEEE Trans. Inform. Theory* **48**(7):1895–1921.

[33] R. Coifman, M. Wickerhauser (1992). Entropy-based algortihms for best basis selection. *IEEE Trans. Inform. Theory* **38**(2):713–718.

[34] A. Conn, N. Gould, P. Toint (2000). *Trust Region Methods*. SIAM.

[35] R. Corless et al. (1996). On the lambert w function. *Advances in Computational Mathematics* (5):329–359.

[36] T. M. Cover, J. A. Thomas (1991). *Elements of Information Theory*. John Wiley & Sons, New York.

[37] A. Z. D. Taubman (1994). Multi-rate 3-d subband coding of video. *IEEE Trans. Image Processing* **3**(5):572–588.

[38] I. Daubechies (1988). Time-frequency localization operators: A geometric phase space approach. *IEEE Trans. Inform. Theory* **34**(4):605–612.

[39] I. Daubechies, W. Sweldens (1998). Factoring wavelet transforms into lifting steps. *J. Fourier Anal. Appl.* .

[40] M. Davies, L. Daudet (2003). Sparsifying subband decompositions. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, IEEE, New Paltz, NY.

[41] M. Davies, L. Daudet (2004). Fast sparse subband decomposition using firsp. In *EUSIPCO*, EURASIP.

[42] G. Davis, S. Mallat, M. Avellaneda (1997). Adaptive greedy approximations. *Journal of Constructive Approximations* **13**:57–98.

[43] L. Davis (1991). *Handbook of Genetic Algorithms*. Van Nostrand.

[44] C. De Vleeschouwer, A. Zakhor (2003). In-loop atom modulus quantization for matching pursuit and its application to video coding. *IEEE Trans. Image Processing* **12**(10):1226–1242.

[45] R. DeVore, V. Temlyakov (1996). Some remarks on greedy algorithms. *Adv. Comput. Math.* **2-3**(5):173–187.

[46] O. Divorra Escoda, M. Flierl, P. Vandergheynst (2004). *Intra-Adaptive Motion-Compensated Lifted Wavelets for Video Coding*. Internal Report 27.2004, EPFL, LTS-2/ITS - EPFL.

[47] O. Divorra Escoda, M. Flierl, P. Vandergheynst (2005). Intra-adaptive motion-compensated lifted wavelets for video coding. In *EUSIPCO*.

[48] O. Divorra Escoda, L. Granai, P. Vandergheynst (2004). *On the Use of* a Priori *Information for Sparse Signal Approximations*. Tech. rep., ITS/LTS-2 EPFL.

[49] O. Divorra Escoda, L. Granai, P. Vandergheynst (2004). *On the Use of* a Priori *Information for Sparse Signal Representations*. Tech. Rep. 18.2004, ITS/LTS-2 EPFL.

[50] O. Divorra Escoda, P. Vandergheynst, M. Bierlaire (2004). *Video Representation Using Greedy Approximations Over Redundant Parametric Dictionaries*. Technical Report 19.2004, LTS-2/ITS - EPFL.

[51] M. Do, P. Dragotti, R. Shukla, M. Vetterli (2001). On the compression of two-dimensional piecewise smooth functions. In *IEEE International Conference on Image Processing (ICIP)*, Thessaloniki, Greece.

[52] M. N. Do (2001). *Directional Multiresolution Image Representations*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne.

[53] D. L. Donoho (1999). Wedgelets: Nearly-minimax estimation of edges. *Annals of Statistics* **27**(3):859–897.

[54] D. L. Donoho, M. Elad (2003). Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization. *Proc. Nat. Aca. Sci.,* **100**(5):2197–2202.

[55] D. L. Donoho, X. Huo (2001). Uncertainty principles and ideal atom decomposition. *IEEE Trans. Inform. Theory* **47**(7):2845–2862.

[56] D. L. Donoho, I. M. Johnstone (1994). Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **81**:425–455.

[57] D. L. Donoho, I. M. Johnstone (1998). Minimax estimation via wavelet shrinkage. *Annals of Statistics* (26):879–921.

[58] P. Dragotti, M. Vetterli (2003). Wavelet footprints: Theory, algorithms and applications. *IEEE Trans. Signal Processing* **51**(5):1306–1323.

[59] D. L. Duttweiler, C. Chamzas (1995). Probability estimation in arithmetic and adaptive huffman entropy coders. *IEEE Trans. Image Processing* **4**(3):237–246.

[60] M. Effros, H. Feng, K. Zeger (2004). Suboptimality of the karhunen-loève transform for transform coding. *IEEE Trans. Inform. Theory* **50**(8):1605–1619.

[61] R. M. Figueras i Ventura, O. Divorra Escoda, P. Vandergheynst (2004). *A Matching Pursuit Full Search Algorithm for Image Approximations*. Tech. Rep. ITS-2004.031, ITS-STI/EPFL.

[62] R. M. Figueras i Ventura, L. Granai, P. Vandergheynst (2002). R-d analysis of adaptive edge representations. In IEEE (ed.), *Workshop on Multimedia Signal Processing*, Virgin Islands.

[63] R. M. Figueras i Ventura, L. Granai, P. Vandergheynst (2002). *R*-D *Analysis of Adaptive Edge Representations*. Tech. rep., Virgin Islands.

[64] R. M. Figueras i Ventura, P. Vandergheynst (2001). *Matching Pursuit with Genetic Algorithms*. Tech. Rep. 02/2001, F-Group, LTS, EPFL.

[65] R. M. Figueras i Ventura, P. Vandergheynst (2002). Evolutive multiresoltion matching pursuit and its relation with the human visual system. In *EUSIPCO*, vol. 2, pp. 395–398.

[66] R. M. Figueras i Ventura, P. Vandergheynst, P. Frossard (to appear). Low rate and scalable image coding using non-linear representations. *IEEE Trans. Image Processing* .

[67] R. M. Figueras i Ventura, P. Vandergheynst, P. Frossard, A. Cavallaro (2004). Color image scalable coding with matching pursuit. In *ICASSP*, vol. 3, pp. 53–56, Montreal.

[68] J. W. Fisher III, T. Darrell (2004). Speaker association with signal-level audiovisual fusion. *IEEE Trans. Multimedia* **6**(3):406–413.

[69] D. J. Fleet (1994). Disparity from local weighted phase-correlation. In *IEEE Int. Conf. on Systems, Man and Cybernetics: Humans, Information and Technology*, pp. 48–54.

[70] M. Flierl (2003). Video coding with lifted wavelet transforms and frame-adaptive motion compensation. In *Visual Content Processing and Representation*, vol. 2849 of *Lecture Notes in Computer Science*, pp. 243–251, Springer, Madrid.

[71] M. Flierl, B. Girod (2003). Generalized b pictures and the draft h.264/avc video compression standard **13**(7):587–597.

[72] M. Flierl, B. Girod (2003). Investigation of motion-compensated lifted wavelet transforms. In *Proceedings of the Picture Coding Symposium*, pp. 59–62, Saint-Malo, France.

[73] M. Flierl, B. Girod (2004). Video coding with motion-compensated lifted wavelet transforms. *EURASIP Journal on Image Communication* **19**(7):561–575. Special Issue on Subband/Wavelet Interframe Video Coding.

[74] M. Flierl, P. Vandergheynst (2005). An improved pyramid for spatially scalable video coding. In *IEEE International Conference on Image Processing*, Genova, Italy.

[75] M. Flierl, P. Vandergheynst, B. Girod (2004). Video coding with lifted wavelet transforms and complementary motion-compensated signals. In *SPIE Conference on Visual Communications and Image Processing*, vol. 5308, pp. 497–508, San Jose, CA.

[76] M. Flierl, T. Wiegand, B. Girod (1998). A locally optimal design algorithm for block-based multihypothesis motion-compensated prediction. In *Data Compression Conf. (DCC)*, pp. 239–248.

[77] W. T. Freeman, E. H. Adelson (1991). The design and use of steerable filters. *IEEE Trans. Pattern Anal. Machine Intell.* **13**(9):891–906.

[78] P. Frossard (2000). *Robust and Multiresolution Video Delivery: From H.26x to Matching Pursuit Based Technologies.* Ph.D. thesis, EPFL, LTS.

[79] P. Frossard, P. Vandergheynst, R. M. Figueras i Ventura (2003). High flexibility scalable image coding. In *VCIP*, Lugano.

[80] P. Frossard, P. Vandergheynst, R. M. Figueras i Ventura, M. Kunt (2004). A posterior quantization of progressive matching pursuit streams. *IEEE Trans. Signal Processing* **52**(2):525 – 535.

[81] S. Geman, D. Geman (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* **6**:721–741.

[82] A. Gilbert, S. Muthukrishnan, M. J. Strauss (2003). Approximation of functions over redundant dictionaries using coherence. In *14th ACM-SIAM Symposium on Discrete Algorithms (SODA'03)*.

[83] F. Giorda, A. Racciu (1975). Bandwidth reduction of video signals via shift vector transmission. *IEEE Trans. Commun.* **23**(9):1002–1004.

[84] B. Girod (1987). The efficiency of motion-compensating prediction for hybrid coding of video sequences. *IEEE J. Select. Areas Commun.* **5**(7):1140–1154.

[85] B. Girod (1993). Motion-compensating prediction with fractional-pel accuracy. *IEEE Trans. Commun.* **41**(4):604–612.

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

[87] G. Golub, C. V. Loan (1996). *Matrix Computations*. Johns Hopkins University Press, Baltimore,MD.

[88] I. F. Gorodnitsky, J. S. George, B. D. Rao (1995). Neuromagnetic source imaging with FOCUSS: A recursive weighted minimum norm algorithm. *J. Electroenceph. Clinical Neurophysiol.* **95**(4):231–251.

[89] I. F. Gorodnitsky, B. D. Rao (1997). Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm. *IEEE Trans. Signal Processing* **45**(3):600–616.

[90] I. S. Gradshteyn, I. M. Ryzhik (2000). *Tables of Integrals, Series, and Products*. Academic Press, San Diego, CA.

[91] L. Granai, P. Vandergheynst (2004). Sparse decomposition over multi-component redundant dictionaries. In *Proc. of Multimedia Signal Processing, Workshop on. MMSP04*, pp. 494–497.

[92] R. Gray (1984). Vector quantization. *IEEE Signal Processing Magazine* **1**(2):4 – 29.

[93] R. M. Gray, D. L. Neuhoff (1998). Quantization. *IEEE Trans. Inform. Theory* **44**(6):2325 – 2383.

[94] R. Gribonval, E. Bacry (2003). Harmonic decomposition of audio signals with matching pursuit. *IEEE Trans. Signal Processing* **1**(51).

[95] R. Gribonval, P. Vandergheynst (2004). *On the Exponential Convergence of Matching Pursuits in Quasi-Incoherent DictionariesDono*. Tech. Rep. 1619, IRISA, Rennes, France.

[96] O. G. Guleryuz (to appear). Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part i - theory. *IEEE Trans. Image Processing* .

[97] O. G. Guleryuz (to appear). Nonlinear approximation based image recovery using adaptive sparse reconstructions and iterated denoising: Part ii - adaptive algorithms. *IEEE Trans. Image Processing* .

[98] R. M. Haralick, J. S. Lee (1993). The facet approach to optical flow. In *Image Understanding Workshop*.

[99] Y. Hel-Or, P. C. Teo (1996). *A Common Framework for Steerability, Motion Estimation and Invariant Feature Detection*. Tech. Rep. CS-TN-96-28, Stanford University.

[100] Y. Hel-Or, P. C. Teo (1998). Canonical decomposition of steerable functions. *Journal of Mathematical Imaging and Vision* **9**(1).

[101] J. Hershey, J. Movellan (1999). Audio-vision: Using audio-visual synchrony to locate sounds. In *NIPS*.

[102] J. Holland (1992). Genetic algorithms. *Sci. Amer.* pp. 44–50.

[103] E. S. Hong, R. E. Ladner (2002). Group testing for image compression. *IEEE Trans. Image Processing* **11**(8).

[104] B. K. P. Horn, B. G. Shunck (1981). Determining optical flow. *Artif. Intell.* **17**:185–203.

[105] R. Horn, C. Johnson (1985). *Matrix Analysis*. Cambridge Univ. Press.

[106] P. G. Howard, J. S. Vitter (1994). Arithmetic coding for data compression. *Proceedings of the IEEE* **82**(6):857–865.

[107] J. R. Jain, A. K. Jain (1981). Displacement measurement and its applicaion in interframe image coding. *IEEE Trans. Commun.* **29**(12):1799–1808.

[108] N. S. Jayant, P. Noll (1984). *Digital Coding of Waveforms*. Prentice-Hall, Inc.

[109] H. Jozawa (1996). Motion compensated video coding using rotation and scaling models. In *Picture Coding Symposium (PCS)*, vol. 1, pp. 309–312, Melbourne.

[110] G. Karlsson, M. Vetterli (1988). Three dimensional sub-band coding of video. In *ICASSP*, vol. 2, pp. 1100 – 1103, IEEE, New York.

[111] A. K. Katsaggelos et al. (2005). Advances in efficent resource allocation for packet-based real-time video transmission. *Proceedings on the IEEE* **93**(1):135–147.

[112] R. D. Kell (1929). Improvements relating to electric picture transmission systems. UK Patent No. 12805/30.

[113] B.-J. Kim, Z. Xiong, W. A. Pearlman (2000). Low bit-rate scalable video coding with 3-d set partitioning in hierarchical trees (3-d spiht). *IEEE Trans. Circuits Syst. Video Technol.* **10**(8):1374 – 1387.

[114] R. Kinderman, J. L. Snell (1980). *Contemporary Mathematics: Markov Random Fields and Their Applications*. American Mathematical Society.

[115] A. P. Korostelev, A. B. Tsybakov (1993). *Minimax Theory of Image Reconstruction*, vol. 82 of *Lecture Notes in Statistics*. Springer-Verlag.

[116] S. Krüger (1998). *Motion Analysis and Estimation Using Multiresolution Affine Models*. Ph.D. thesis, Department of Computer Science - University of Bristol.

[117] C. Kuglin, D. Hines (1975). The phase correlation image alignment method. In *IEEE Int. Conf. Cybern. Soc.*, pp. 163–165.

[118] E. Le Pennec, S. Mallat (2005). Sparse geometric image representations with bandelets. *IEEE Trans. Image Processing* **to appear**.

[119] M. Lewicki, T. Sejnowski (2000). Learning overcomplete representations. *Neural Comp.* **12**:337–365.

[120] Y. Lu, C. Lu, Z. Li (2003). A modified space frequency decomposition algorithm for visual motion. In *ICME*, Baltimore.

[121] S. Mallat (1998). *A Wavelet Tour of Signal Processing*. Academic Press.

[122] S. G. Mallat, Z. Zhang (1993). Matching pursuits with time-frequency dictionaries. *IEEE Trans. Signal Processing* **41**(12):3397–3415.

[123] R. Manduchi, P. Perona, D. Shy (1997). *Efficient Implementation of Deformable Filter Banks*. Tech. Rep. CNS-TR-97-04, California Institute of Technology.

[124] S. Mann, R. Picard (1997). Video orbits of the projective group: A simple approach to features estimation of parameters. *IEEE Trans. Image Processing* **6**(9):1281–1295.

[125] X. Marichal (1998). *Motion Estimation and Compensation for Very Low Bit-Rate Video Coding*. Ph.D. thesis, Université catholique de Louvain, Louvain la Neuve.

[126] D. Marr (1982). *Vision: a computational investigation into the human representation and processing of visual information*. W. H. Freeman, San Francisco.

[127] S. Martello, P. Toth (1990). *Knapsack problems: algorithms and computer implementations*. Wiley, New York.

[128] C. Mayer (2003). Motion compensated in-band prediction for wavelet-based spatially scalable video coding. In *IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, Hong Kong (cancelled).

[129] Y. Meyer (2002). Oscillating patterns in image processing and in some nonlinear evolution equations. In *AMS*, vol. 22 of *University Lecture Series*.

[130] G. Monaci (2005). Multimodal analysis using redundant parametric decompositions. on-line: http://lts2www.epfl.ch/∼monaci/multimodal.html.

[131] G. Monaci, O. Divorra Escoda, P. Vandergheynst (2004). *Multimodal Analysis Using Redundant Parametric Decompositions*. Tech. Rep. ITS-2004.024, LTS-2/ITS EPFL.

[132] R. Neff, A. Zakhor (1997). Very low bit-rate video coding based on matching pursuits **7**(1):158–171.

[133] H. J. Nussbaumer (1982). *Fast Fourier Transform and Convolution Algorithms*. Springer-Verlag, Berlin.

[134] J.-R. Ohm (1994). Three-dimensional subband coding with motion compensation. *IEEE Trans. Image Processing* **3**(5):559 – 571.

[135] J. R. Ohm (2005). Advances in scalable video coding. *Proceedings of the IEEE* **93**(1):42 – 56.

[136] B. Olshausen, D. Field (1997). Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research* **37**:3311–3325.

[137] B. A. Olshausen (2003). Learning sparse, overcomplete representations of time-varying natural images. In *IEEE ICIP*, Barcelona, Catalonia.

[138] A. Papoulis (1991). *Probability, Random Variables, and Stochastic Processes*. McGrawHill, 3rd edn.

[139] Y. Pati, R. Reziifar, P. S. Krishnaprasad (1993). Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*.

[140] G. Pau, C. Tillier, B. Pesquet-Popescu, H. Heijmans (2004). Motion compensation and scalability in lifting-based video coding. *Signal Processing: Image Communication* **19**(7):577 – 600.

[141] L. Peotta, L. Granai, P. Vandergheynst (2003). Very low bit rate image coding using redundant dictionaries. In *Proc. of 48th SPIE Annual Meeting*, SPIE, San Diego, CA.

[142] L. Peotta, P. Vandergheynst (2003). *MP in Block Quasi-Incoherent Dictionaries*. Tech. rep., Signal Processing Institute, EPFL, Lausanne, Switzerland.

[143] P. Perona (1992). Steerable-scalable kernels for edge detection and junction analysis. *IVC* **10**:663–672.

[144] B. Pesquet-Popescu, V. Bottreau (2001). Three-dimensional lifting schemes for motion compensated video compression. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1793 – 1796.

[145] J. D. Pintér (1996). *Global Optimization in Action*, vol. 6 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

[146] C. I. Podilchuk, N. S. Jayant, N. Favardin (1995). Three-dimensional subband coding of video **4**(2):125 – 139.

[147] J. Portilla, E. P. Simoncelli (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision* **40**(1):49–71.

[148] P. Prandoni, M. Vetterli (1999). Approximation and compression of piecewise smooth functions. *Phil. Trans. R. Soc. Lond.* **357**(1760):2573–2591.

[149] M. P. Queluz (1996). *Multiscale Motion Estimation and Video Compression*. Ph.D. thesis, Laboratoire de Telecommunications et Teledetection, UCL, Louvain la Neuve, Belgique.

[150] L. Rabiner, B. H. Juang (1993). *Fundamentals of speech recognition*. Prentice Hall, Englewood Cliffs, New Jersey.

[151] A. Rahmoune, P. Vandergheynst, P. Frossard (2004). *Flexible Motion-Adaptive Video Coding with Redundant Expansions*. Technical Report 17.2004, LTS-2/ITS - EPFL.

[152] A. Rahmoune, P. Vandergheynst, P. Frossard (2004). Mp3d: Highly scalable video coding scheme based on matching pursuit. In *ICASSP*.

[153] K. Ramchandran, M. Vetterli (1993). Best wavelet packet bases in a rate-distortion sense. *IEEE Trans. Image Processing* **2**(2):160–175.

[154] K. Ramchandran, Z. Xiong, K. Asai, M. Vetterli (1996). Adaptive transforms for image coding using spatially varying wavelet packets. *IEEE Trans. Image Processing* **5**(7):1197–1204.

[155] S. Sardy, A. Bruce, P. Tseng (2000). Block coordinate relaxation methods for nonparametric wavelet denoising. *Journal of Computational and Graphical Statistics* **9**(2):361–379.

[156] K. Sayood (2000). *Introduction to Data Compression*. Academic Press, 2nd edn.

[157] H. Schwarz, D. Marpe, T. Wiegand (2004). *Scalable Extension of H.264/AVC (Proposal) - MPEG04/M10569/S03*. ISO/IEC JTC1/SC29/WG11, Munich.

[158] A. Secker, D. Taubman (2001). Motion-compensated highly scalable video compression using an adaptive 3d wavelet transform based on lifting. In *IEEE International Confernce on Image Processing (ICIP)*, vol. 2, pp. 1029–1032, Thessaloniki, Greece.

[159] A. Secker, D. Taubman (2003). Lifting-based invertible motion adaptive transform (limat) framework for highly scalable video compression. *IEEE Trans. Image Processing* **12**(12):1530–1542.

[160] A. Secker, D. Taubman (2004). Highly scalable video compression with scalable motion coding. *IEEE Trans. Image Processing* **13**(8):1029 – 1041.

[161] V. Seferidis, M. Ghanbari (1993). General approach to block matching motion estimation. *Optical Engineering* **32**:1464–1474.

[162] M. Servais, G. de Jager (1997). Video compression using the three dimensional discrete cosine transform (3d-dct). In *Sout African Symposium on Communications and Signal Processing*, pp. 27–32, IEEE.

[163] R. Shukla (2004). *Rate-Distortion Optimized Geometrical Image Processing*. Ph.D. thesis, Ecole Polythecnique Fédérale de Lausanne, Lausanne, Switzerland.

[164] R. Shukla, P. Dragotti, M. Do, M. Vetterli (2004). Rate distortion optimized tree structured compression algorithms for piecewise smooth images. *IEEE Trans. Image Processing* .

[165] M. Slaney, M. Covell (2000). Facesync: A linear operator for meaasuring synchronization of video facial images and audio tracks. In *NIPS*.

[166] P. Smaragdis, M. Casey (2003). Audio/visual independent components. In *ICA*, pp. 709–714.

[167] J. L. Starck, M. Elad, D. L. Donoho (2004). *Image Decomposition Via the Combination of Sparse Representations and a Variational Approach*. Tech. rep., CEA-Saclay, DAPNIA/SEDI-SAP.

[168] G. J. Sullivan, R. L. Baker (1991). Motion compensation for video compression using control grid interpolation. In *Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*.

[169] G. J. Sullivan, R. L. Baker (1991). Rate-distortion optimized motion compensation for video compression using fixed or variable size blocks. In *IEEE Global Telecommunications Conf. (GLOBECOM)*, pp. 85–90.

[170] G. J. Sullivan, T. Wiegand (2005). Video compression - from concepts to the h.264/avc standard. *Proceedings of the IEEE* **93**(1):18 – 30.

[171] W. Sweldens (1996). Wavelets and the lifting scheme: A 5 minute tour. *Z. Angew. Math. Mech.* **76**(2):41–44.

[172] Y. Taki, M. Hatori, S. Tanaka (1974). Interframe coding that follows the motion. In *Institute of Electronics and Communication Engineers Jpn. Annu. Conv. (IECEJ)*, p. 1263.

[173] V. N. Temlyakov (2000). Weak greedy algorithms. *Advances in Computational Mathematics* **12**(2-3):213–227.

[174] J. A. Tropp (2003). *Greed is Good: Algorithmic Results for Sparse Approximation*. Tech. rep., ICES, University of Texas at Austin, Austin, USA.

[175] J. A. Tropp (2004). Greed is good : Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory* **50**(10):2231–2242.

[176] J. A. Tropp (2004). *Just Relax: Convex Programming Methods for Subset Selection and Sparse Approximation*. Tech. rep., ICES, University of Texas at Austin, Austin, USA.

[177] D. S. Turaga, M. van der Schaar (2003). Content-adaptive filtering in the umctf framework. In *IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP)*, Hong Kong (cancelled).

[178] P. Vandergheynst, P. Frossard (2001). Efficient image representation by anisotropic refinement in matching pursuit. In *IEEE ICASSP*, vol. 3, pp. 1757–1760.

[179] L. A. Vese, S. J. Osher (2003). Modeling textures with total variation minimization and oscillating patterns in image processing. *Journal of Scientific Computing* **19**(1-3):553–572.

[180] M. Vetterli (2001). Wavelets, approximation and compression. *IEEE Signal Processing Mag.* **18**(5):59 – 73.

[181] M. Vetterli, J. Kovačević (1995). *Wavelets and Subband Coding*. Prentice-Hall PTR.

[182] M. Wakin, J. Romberg, H. Choi, R. Baraniuk (2004). Wavelet-domain approximation and compression of piecewise smooth images. *IEEE Trans. Image Processing* **submitted**. Submitted.

[183] Y. Wang, J. Ostermann, Y. Zhang (2001). *Digital Video Processing and Communications*. Prentice Hall.

[184] Y. Wang, A. R. Reibman, L. Shunan (2005). Multiple description coding for video delivery. *Proceedings on the IEEE* **93**(1):57– 70.

[185] E. W. Weisstein. (1999-2004). Lambert w-function. From MathWorld–A Wolfram Web Resource. "http://mathworld.wolfram.com/LambertW-Function.html". (last accessed on May 2005).

[186] T. Wiegand, G. Sullivan, G. Bjntegaard, A. Luthra (2003). Overview of the h.264/avc video coding standard. *IEEE Trans. Circuits Syst. Video Technol.* **13**(7):557 – 559.

[187] T. Wiegand, X. Zhang, B. Girod (1999). Long-term memory motion compensated prediction **9**(1):70–84.

[188] I. Witten, R. Neal, J. Cleary (1987). Arithmetic coding for data compression. *Commun. ACM* **30**(6):520–540.

[189] J. Xu, Z. Xiong, S. Li, Y.-Q. Zhang (2001). Three-dimensional embedded subband coding with optimized truncation (3-d escot). *Applied and Computational Harmonic Analysis* **10**.

[190] M. Zibulevsky, B. A. Pearlmutter (2001). Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation* **13**(4):863–882.

Curriculum Vitae

Oscar DIVORRA ESCODA

Citizenship:	Spanish, EU	Signal Processing Institute (LTS2/ITS)
Date of Birth:	November 21st, 1977	École Polytechnique Fédérale de Lausanne (EPFL)
Place of Birth:	Móra d'Ebre, Catalonia	CH-1015 Lausanne, Switzerland
Email:	oscar.divorra@ieee.org	Web: http://lts2www.epfl.ch/∼divorra

Work Experience

Mar. 2001 – Mar. 2005	**École Polytechnique Fédérale de Lausanne (EPFL)** Lausanne, Switzerland, School of Engineering, Signal Processing Institute, Research and Teaching Assistant.
Oct. 2003	**Université catholique de Louvain (UCL)** Louvain la Neuve, Belgium, Telecommunications Laboratory, Visiting Researcher.
Oct. 2000 – Dec. 2000	**École Polytechnique Fédérale de Lausanne (EPFL)** Lausanne, Switzerland, School of Engineering, Signal Processing Institute, Research Internship.
Jul. 1999 – Sep. 1999	**Robert Bosch GmbH** Hildesheim, Germany, Industrial Research Practicum

Education

Mar. 2001 – Jul. 2005	**École Polytechnique Fédérale de Lausanne (EPFL)** Lausanne, Switzerland, School of Engineering, Signal Processing Institute, Ph.D. Program
Feb. 2000 – Aug. 2000	**École Polytechnique Fédérale de Lausanne (EPFL)** Lausanne, Switzerland, School of Engineering, Signal Processing Institute, Exchange Student (Erasmus program), M.Sc. Thesis.

Sep. 1995 – Nov. 2000	**Universitat Politècnica de Catalunya (UPC)**, Barcelona, Catalonia, Spain, School of Telecommunications Engineering, M.Sc. Degree on Electrical and Electronics Engineering.

Languages

Catalan:	Mother tongue.
Spanish:	Bilingual.
English:	Fluent oral, read and written.
French:	Fluent oral, read and written.
German:	Basic knowledge.
Italian:	Basic knowledge.

Awards & Honors

- UCLA Grant to attend to the "Multiscale Geometry in Image Processing and Coding" graduate course at IPAM, UCLA, Los Angeles, USA (2004).
- 2^{nd} awarded on the Chemistry Olympiads of Catalonia in 1995.
- 3^{rd} awarded on the Physics Olympiads of Catalonia in 1995.

Hobbies

- Music: I like to listen to and to play music. I hold a *Piano Diploma* from *Liceu Superior de la Música de Barcelona*.
- Sports: Hiking, Team sports, Skiing, Badminton...
- Other: Cinema, Traveling, Reading, Cooking...

Personal Publications

Journal Papers

- G. Monaci, O. Divorra Escoda and P. Vandergheynst, **Analysis of Multimodal Sequences Using Geometric Video Representations**, Submitted to Signal Processing: Image Communication on June 2005
- O. Divorra Escoda, L. Granai and P. Vandergheynst, **On the Use of A Priori Information for Sparse Signal Approximations**, Submitted to IEEE Transactions on Signal Processing on February 2005
- A. Petrovic, O. Divorra Escoda and P. Vandergheynst, **Multiresolution Segmentation of Natural Images: From linear to Non- Linear Scale-Space Representations**, IEEE Transactions on Image Processing, Vol. 13, No 8, pp. 1104-1114, August 2004

Conference Papers

- G. Monaci, O. Divorra Escoda, P. Vandergheynst, **Analysis of Multimodal Signals Using Redundant Representations**, Proc. of IEEE, ICIP, September 2005
- O. Divorra Escoda, M. Flierl, P. Vandergheynst, **Intra-Adaptive Motion-Compensated Lifted Wavelets for Video Coding**, Proc. of EURASIP, EUSIPCO, September 2005
- O. Divorra Escoda and P. Vandergheynst, **A Bayesian Approach to Video Expansions on Parametric Over-Complete 2-D Dictionaries**, Proc. of IEEE, MMSP, September 2004
- O. Divorra Escoda and P. Vandergheynst, **Video Coding Using a Deformation Compensation Algorithm Based on Adaptive Matching Pursuit Image Decompositions**, Proc. of IEEE, ICIP, September 2003
- O. Divorra Escoda and P. Vandergheynst, **Locally Temporal Adaptive Transform Scheme for Sub-band Video Coding**, Proc. of IEEE, ICASSP, April 2003
- O. Divorra Escoda, A. Petrovic and P. Vandergheynst, **Segmentation of Natural Images Using Scale-Space Representations: A Linear and a Non-Linear Approach**, In Proceedings of EUSIPCO, September 2002

- T. Ebrahimi, Y. Abdeljaoued, R. M. Figueras i Ventura and O. Divorra Escoda, **MPEG-7 CAMERA**, In Proceedings of IEEE, ICIP, Vol. 3, pp. 600-603, October 2001

- O. Divorra Escoda, R. M. Figueras i Ventura, E. Debes and T. Ebrahimi, **Influence of a Large Image Watermarking Scheme Parallelization on Possible Attacks**, Proc. of SPIE's 46th Annual Meeting on Optical Science and Technology, Vol. 4472, pp. 175-186, August 2001

Patents

- T. Butz, E. Dati, O. Divorra Escoda, M. Kunt and P. Vandergheynst, **Temperature Image Reconstruction**. EU Patent Application N. 05405418.4, Ref. N. B-4867-EP Filed July 2005

- O. Divorra Escoda, P. Vandergheynst, M. Bierlaire, J. Reichel and F. Ziliani, **Video Coding Method of Exploiting the Temporal Redundancy Between Successive Frames in a Video Sequence**. EU Patent PCT/IB03/06141 Filed December 17th, 2003

Technical Reports

- O. Divorra Escoda and P. Vandergheynst, **An Analysis of Temporal Adaptivity in 3D Wavelet Video Coding**, ITS-TR 06/2005, École Polytechnique Fédérale de Lausanne, Signal Processing Institute, January 2005

- R. M. Figueras i Ventura, O. Divorra Escoda P. Vandergheynst, **A Matching Pursuit Full Search Algorithm for Image Approximations**, ITS-TR 31/2004, École Polytechnique Fédérale de Lausanne, Signal Processing Institute, December 2004

- O. Divorra Escoda, M. Flierl and P. Vandergheynst, **Intra-Adaptive Motion-Compensated Lifted Wavelets for Video Coding**, ITS-TR 27/2004, École Polytechnique Fédérale de Lausanne, Signal Processing Institute, November 2004

- O. Divorra Escoda, L. Granai and P. Vandergheynst, **On the Use of A Priori Information for Sparse Signal Approximations**, ITS-TR 23/2004, École Polytechnique Fédérale de Lausanne, Signal Processing Institute, November 2004

- G. Monaci, O. Divorra Escoda and P. Vandergheynst, **Multimodal Analysis Using Redundant Parametric Decompositions**, ITS-TR 24/2004, École Polytechnique Fédérale de Lausanne, Signal Processing Institute, October 2004

- O. Divorra Escoda, L. Granai and P. Vandergheynst, **On the Use of A Priori Information for Sparse Signal Representations**, ITS-TR 18/2004, École Polytechnique Fédérale de Lausanne, Signal Processing Institute, September 2004

- O. Divorra Escoda, P. Vandergheynst and M. Bierlaire, **Video Representation Using Greedy Approximations Over Redundant Parametric Dictionaries**, ITS-TR 19/2004, École Polytechnique Fédérale de Lausanne, Signal Processing Institute, September 2004

- O. Divorra Escoda and P. Vandergheynst, **Segmentation of Natural Images Using Scale-Space Representation with Multi-Scale Edge Supervised Hierarchical Linking**, ITS-TR 04/2001, École Polytechnique Fédérale de Lausanne, Signal Processing Institute, August 2001

Master Thesis

- O. Divorra Escoda, **Motion Detection & Segmentation for Audio-Visual Source Separation**, École Polytechnique Fédérale de Lausanne, Signal Processing Institute, August 2000