IMAGE/VIDEO REPRESENTATION AND SCALABLE CODING USING REDUNDANT DICTIONARIES

THÈSE N° 3316 (2005)

PRÉSENTÉE À LA FACULTÉ SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

Institut de traitement des signaux

SECTION DE GÉNIE ÉLECTRIQUE ET ÉLECTRONIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Adel RAHMOUNE

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

acceptée sur proposition du jury:

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

Lausanne, EPFL
2005

Abstract

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

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

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

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

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

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

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

Version Abrégée

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

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

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

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

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

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

Contents