VIDEO CODING USING A DEFORMATION COMPENSATION ALGORITHM BASED ON ADAPTIVE MATCHING PURSUIT IMAGE DECOMPOSITIONS

Oscar Divorra Escoda and Pierre Vandergheynst

Signal Processing Institute (ITS)
Swiss Federal Institute of Technology in Lausanne (EPFL)
Ecublens, 1015 Lausanne, Switzerland
Email: {oscar.divorra, pierre.vandergheynst}@epfl.ch

ABSTRACT

Today's video codecs employ motion compensated prediction in combination with block matching techniques. These techniques, although achieving some level of adaptivity in their latest versions [1], continue to rely on the decomposition of frames on a set of artificial primitives: blocks. This paper presents a new approach to video coding. A geometrically adaptive image decomposition scheme using an over-complete basis is used to represent the scene. Using Matching Pursuit (MP), we are able to express local features such as position, anisotropic scale and orientation in terms of a set of spatio-frequential primitives. In order to perform frame prediction, only the changes in the parameters that determine these functions from frame to frame will have to be coded. Such an approach, in addition to being able to catch displacements in images deals as well in a natural way with local scale deformations and local rotations.

1. INTRODUCTION

Classically used separable bases are not able to represent optimally 2D piecewise smooth signals such as images, therefore a growing interest has appeared in the field of approximation theory in finding better approximating bases for edge like singularities [2, 3, 4, 5]. In the field of video coding, interesting ideas have been introduced in this direction [6] by coding the residual motion prediction error with functions capable of exploiting redundancy in its structure. Anyway, this approach is limited by the fact that state of the art motion compensation does not necessarily respect the natural structures of images.

An interesting approach is to use more flexible representation techniques that would allow for more efficient coding of natural structures (i.e edges, textures or even motion). Our motivation in this paper is to use an adaptive representation of spatial primitives and correctly track its evolution through time in order to perform some higher order motion compensation, that could be called deformation compensation. One of the main interests of our technique with respect to state of the art in motion compensation based on block matching (BM) or warping procedures is its ability to efficiently represent visual primitives without imposing an ad hoc partition.

The paper is structured in the following way: In sec. 2 use of over-complete function bases for adapted image representation is reviewed followed by its possible extension to a predictive scheme for frame compensation. In sec. 3 a description of the proposed algorithm can be found followed by some result discussions and the conclusions.

This work was supported by the Swiss Federal Office for Education and Technology grant number 6044.1 KTS.

2. MATCHING PURSUIT: FROM IMAGE TO VIDEO REPRESENTATION

Expansions on general over-complete dictionaries of functions is a problem with an infinite number of solutions [7]. Finding the best N-term expansion is normally NP-hard, though for some dictionaries convex optimization procedures (Basis Pursuits) yield the correct optimum. Without restrictions on the dictionary, sub-optimal approaches have to be followed.

Matching pursuit (MP) is one of the sub-optimal approaches that iteratively approximates the solution [8]. It computes one coefficient of the expansion at every iteration such that the norm of the residual is minimized. According to this, the N-term expansion can be expressed as:

$$f = \sum_{n=0}^{N-1} \langle g_{\gamma_n}, \mathcal{R}^n f \rangle \cdot g_{\gamma_n} + \mathcal{R}^N f, \tag{1}$$

where g_{γ_n} is the function selected from the redundant dictionary and $\mathcal{R}^n f$ the remaining residual of the function being approximated at iteration n.

2.1. 2D Signal Representation Through Matching Pursuits

Studies concerning 2D edge representation show the improvement and near-optimality that the use of anisotropic oriented functions can bring in some cases [2, 3, 4]. Further studies by *Figueras et al.* [5] have as well underlined the importance of being geometrically adaptive in the choice of the set of functions that will approximate the 2D piecewise smooth signal.

According to this, it can be concluded that a dictionary of anisotropically scaled and oriented functions should be used for image expansions. In the previous work done by *Vandergheynst and Frossard* [9], this basis led to the use of MP together with a dictionary of functions (atoms) that were smooth in one direction and behaved as a wavelet in the orthogonal one. Such functions, able to represent efficiently contour like singularities in 2D, form a redundant dictionary built by applying translations, rotations and anisotropic scaling to the following generating function:

$$g(x, y) = \left(4x^2 - 2\right) \exp\left(-\left(x^2 + y^2\right)\right). \tag{2}$$

In this way our dictionary $\mathcal{D}$ will be such that:

$$g_\gamma \in \mathcal{D}, \gamma \in \Gamma, \tag{3}$$

where Γ defines all the possible allowed combinations (a priori defined for ensuring completeness) of parameters, with:

$$\gamma = (dx, dy, sx, sy, \theta), \tag{4}$$

where dx and dy determine position, sx wavelet axis scaling, sy smooth axis scaling and θ rotation.

Besides of edge like information, an image is also composed of smooth regions that correspond to low frequencies. This frequencies are more hardly represented with the use of the dictionary presented in [9]. That is why we separate the representation of the low resolution components of the image and apply MP with $\mathcal{D}$ to the remaining high frequency residual exclusively.

2.2. Extension Toward Video Representation

From sec. 2.1, it can be seen that functions have an important geometrical sense. Being conceived to represent adaptively arbitrary smoothed straight lines, they have a close relation with the local behavior of the 2D signal to be expanded. This is because although being infinite in space and frequency, the atom energy is very well localized in space as well as in frequency. Among the properties of the full dictionary generated by $g_\gamma(x, y)$ with $\gamma \in \mathbb{R}^5$, it is worth mentioning its invariance with respect to translations, rotations and isotropic scale transformations [9]. i.e. the application of an affine transformation on an image would modify in the same way the atom parameters without changing projection coefficients.

These two properties of the dictionary are the base in the assumption that two consecutive frames in a video sequence should have close expansions. Local image transformations due to translations rotations or scalings associated to common motion or even deformation of objects are expected to modify consequently parameters that define the selected atoms. Nevertheless, since $\mathcal{D}$ is redundant, interferences among atoms will occur, and coefficients will indeed be affected, although their difference with the original is assumed to be close to zero compared with the coefficient modulus.

In Fig. 1, a graphical explanation of the idea is found. The objective is to track the variation of parameters that define the functions that locally approximate the frames. By means of this tracking, a geometrical description of frames is achieved as well as their evolution through time. Such an approach is able to catch in a much more natural way geometrical image transformations that otherwise should be found through compensation of artificial image partitions like blocks that do not respect the structure of natural images.

Let $\{c_n, \gamma_n, n = 0, ..., N-1\}_t$ be the set of coefficients and atom parameters that describe a frame at time t, the problem to be solved is to find the transformation F_t such that:

$$\begin{aligned} &\{c_n, \gamma_n, n = 0, ..., N-1\}_{t+1} = ... \\ &\quad F_t\left(\{c_n, \gamma_n, n = 0, ..., N-1\}_t\right). \end{aligned} \tag{5}$$

F_t is defined such that the approximation of frame I_{t+1} is maximized in terms of energy. This leads to a formulation of the form:

$$\underset{F_t}{ArgMin} \left\{ \| I_{t+1} - \sum_n F_t^{c_n}(c_n) F_t^{\gamma_n}(g_{\gamma_n}) \| \right\}, \tag{6}$$

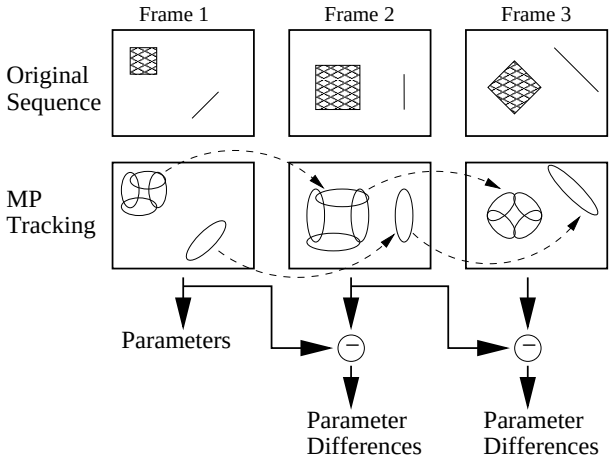

Fig. 1. Successive schematic updates of basis functions in a sequence of frames.

where F^{γ_n} and F^{c_n} correspond respectively to the transformation that concerns parameters and the transformation concerning coefficients.

Eq. 6 defines the frame prediction problem in terms of the image spatio-frequential primitives g_γ. This is again a NP-complete problem, but this time we know that the point $\{c_n, \gamma_n, n = 0, ..., N-1\}_t$ should be very close to the final solution when smooth motion is considered. From this, we consider (in the same way as it is assumed in BM techniques) that Eq. 6 can be solved as a local optimization problem. This is a non-convex, non-linear, differentiable optimization problem [10], which can be solved using various algorithms such as quasi-Newton methods, combined with line search or trust-region globalization techniques [11]. Solving for all $\gamma_n \, \forall \, n \in N$ at once, is still a very complex problem. This is the reason why a greedy approach has been chosen again to solve it, but unlike in sec. 2.1 just a local search [12] will be performed instead of the search through all the functions of the dictionary. This jointly allows us to track atoms locally as well as to reduce complexity.

3. ALGORITHM

The approach followed to perform the video base layer is based on a predictive scheme of the kind IPPP.... Intra frames are coded on the basis of their expansion on the over-complete basis by means of MP defined in sec. 2.1. As for P frames, they are predicted by locally optimizing the expansion obtained in the previous frame such that a new and better expansion on $\mathcal{D}$ is achieved, as explained in sec. 2.2. For both I and P frames, low resolution components will be separately represented and coded.

3.1. I-Frames Coding

Fig. 2 represents the algorithm applied to code Intra frames. Prior to the projection on $\mathcal{D}$ a low resolution approximation is subtracted from the image, quantized and entropy coded separately. This approximation is obtained from a R times downsampled version of the full size frame, generated with the help of an appropriate smooth kernel. The detail residual frame is then expanded through matching pursuit on the dictionary generated from Eq. 2. A limited

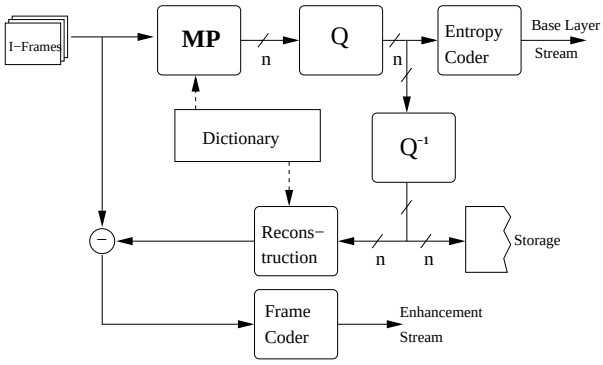

Fig. 2. Intra frame coding scheme.

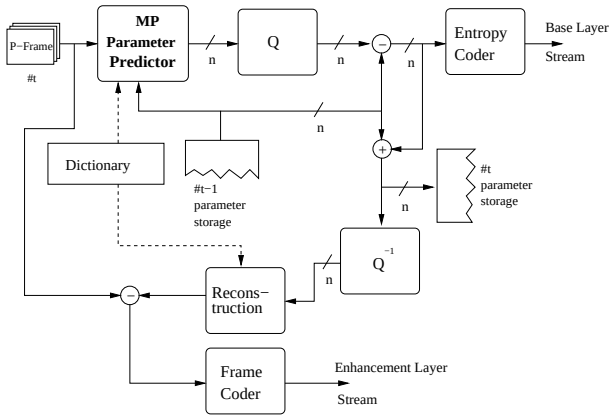

Fig. 3. Predicted frame coding scheme.

number of possible orientations N_θ is imposed as well as a maximum number of Scales N_{Smax}, the number of scales per octave is set to $N_{s/oct}$. In order to keep atoms elongated, the scale of the smooth axis must be always bigger than that of the wavelet axis, and the minimum scale $Smin$ (σ^2 in the Gaussian) is assumed to be 1.

The number of coefficients extracted from by MP will be the largest possible regarding to a coding point of view. When the decrease of distortion due to extension of the expansion is worse than using other simpler techniques, it is possible to switch and code the residual using an enhancement layer (see Fig. 2). Since the scope of this work is to propose an adaptive deformable base layer, we leave the enhancement scheme for future discussions.

Concerning the coefficient quantization to be used, a detailed discussion as well as a comparison with different techniques can be found in [13]. In Figure 2, the output of the MP process represent both coefficients and basis functions parameters. For the parameters, quantization means just the mapping of their values to symbols.

3.2. P-Frames Coding

Fig. 3 represents the algorithm applied to code Predicted frames. As in the case of I-frames, low resolution approximation are subtracted and coded separately. The projection of the frame through MP is again performed. The minimization algorithm [12] starts from the stored parameters of the precedent frame. It optimizes atom by atom in the same order as they were first found during the full search MP used for the Intra frame expansion. During the optimization of each atom, parameters are allowed to evolve in a continuous space. Since predicted atoms g'_γ are also represented with respect to $\mathcal{D}$, optimized parameters will have to be quantized such that $g'_\gamma \in \mathcal{D}$. At every predicted frame, parameters are stored in order to be used to predict the following one. Only the difference of the new parameters and coefficient will be streamed to the entropy coder. Here again, and in the same way as most popular video coding techniques, a second layer for quality enhancement could be considered in order to code the error. But this is out of the scope of this work.

4. RESULTS

Tests have been performed on the sequence foreman in QCIF format (176x144) at 30 frames per second. On the basis of the dictionary of sec. 2.1 and according to sec. 3.1 we consider $R = 4$, $N_\theta = 36$ which sets a resolution of $\frac{\pi}{36}$ due to atom symmetry, $N_{Smax} = 10$ and $N_{s/oct} = 2$. Tests have been performed on a sequence of 100 frames taking just the first frame as I-frame. All other frames have been progressively predicted by means of the procedure described in 3.2. Since no adaptive rate control is performed, the number of coefficients extracted through the MP in both Intra and Predictive coding will be fixed manually to 200.

Fig. 4. Prediction results: Original 0 frame (upper left), Coded 0 frame (upper right), Predicted 20th frame (down left) and Predicted 98th (down right).

In Fig. 4, results of prediction through 100 frames can be seen at different stages (frames 0 -I-, 20th -P-, 98th -P-) of the prediction. Where the intra frame, is coded at 0.28 bpp giving a PSNR of 30.07 dB when exponential quantization [13] is used. This shows how the algorithm is able to track spatio-frequential features through time in long sequences with just the information that concerns the prediction process (which would correspond to motion compensation in a classical hybrid video coding) without taking into account any error coding (or enhancement layer).

Fig. 5 presents the comparison between the coded sequence through the deformation compensation algorithm (visual results in Fig. 4, where the limitation of 200 atoms in the frames representa-

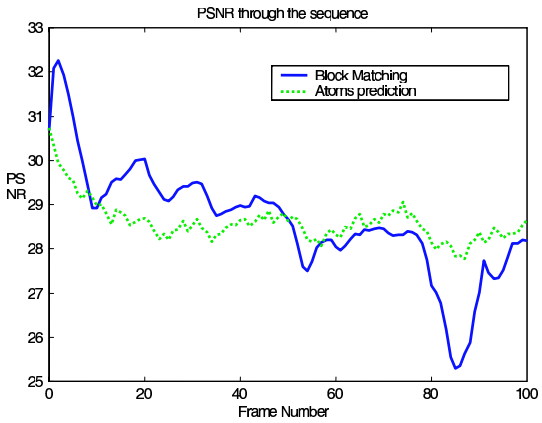

Fig. 5. PSNR comparison between an adaptive block matching scheme and the adaptive MP based compensation

tion imposes an average bit rate of 180kb/s, and between the result of the compensation with an adaptive multiscale block matching scheme where no error coding is performed. The adaptive block matching selects adaptively size blocks from 4x4 up to 16x16 and has been allowed to refine as much as it needs (in terms of better motion adaptation) the size of blocks, such that it achieves approximately the same bit budget as ours. We must notice, that in the enounced obtained bit-rate just a rounding to the closest integer has been applied to the coefficients of the frame (I and P) expansions. This implies that a much better D-R relation can be achieved for the MP case if an adapted quantization is performed. In addition, the early optimization strategy used for the atom tracking has many chances at every minimization to get stuck in local minima due to complexity of natural images. This can be seen as the addition of noise to the parameter estimation. Noise that will contribute to the increase of bit rate needs and a decrease in the approximation accuracy (increase of distortion). Moreover, since it is a completely new approach, an important amount of work has still to be done in terms of studying the appropriate coding strategies.

An important detail to observe is the fact of the stability presented by the PSNR curve of the adaptive deformable scheme in front of the BM. This can be explained due to the higher flexibility of the primitives on which images are decomposed. In fact, in the last section of both curves a temporal fall in the BM curve appears after the 90th frame. This, in the foreman sequence, corresponds to the entrance of the hand in the camera view. Unlike the block matching approach, in the adaptive deformations atoms that were representing the background may be re-used and reshaped when occlusion occurs such that a better representation is possible.

5. CONCLUSIONS

In this paper we have introduced a new way of handling temporal redundancy by means of deforming visual primitives in sequences of frames. Results show that we are able to track motion along a large amount of frames without a dramatic loss in quality. Our algorithm is also able to adapt to locally affine motions, such as rotations or scalings, unlike classical block matching strategies. Preliminary coding experiments prove that our scheme compares favorably to the state of the art for low bit rate video compression in the absence of motion compensation error coding. This technique is still in a very preliminary state and many improvements are foreseen. Possible future enhancements include a clean study of the minimization algorithm in order to avoid sub-optimalities due to local minima, a proper design of an adapted quantization strategy and a suitable coding of the parameters and the low resolution layer (for example using a wavelet codec). Furthermore, appropriate dictionaries for image representation are still an open question.

Acknowledgements

We would like to thank Prof. Michel Bierlaire, Dr. Julien Reichel, Dr. Francesco Ziliani and Dr. Markus Flierl for fruitful discussions and suggestions for future work directions.

6. REFERENCES

[1] *ITU-T Recommendation H.264/AVC*.

[2] E. J. Candès and D. L. Donoho, "Curvelets - a surprisingly effective non-adaptive representation for objects with edges.," *Curves and Surfaces, L. L. S. et al., ed., Nashville, TN, (Vanderbilt University Press)*, pp. 123–143, 1999.

[3] M. N. Do and M. Vetterli, "Contourlets: A directional multiresolution image representation," in *ICIP*, Rochester, NY, September 2002.

[4] M. N. Do, P. L. Dragotti, R. Shukla, and M. Vetterli, "On the compression of two-dimensional piecewise smooth functions," in *ICIP*, Thesalonica, October 2001.

[5] R.M. Figueras i Ventura, L. Granai, and P. Vandergheynst, "R-D analysis of adaptive edge representations," in *MMSP*, Virgin Islands, December 2002.

[6] O. Al-Shaykh, E. Miloslavsky, T. Nomura, Neff R., and A. Zakhor, "Video compression using matching pursuits," *IEEE Trans. Circuits Syst. Video Technol.*, pp. 123–143, 1999.

[7] S. Chen and D. Donoho, "Atomic decomposition by basis pursuit," in *SPIE International Conference on Wavelets*, Sant Diego, July 1995.

[8] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[9] P. Frossard and P. Vandergheynst, "Efficient image representation by anisotropic refinement in matching pursuit," in *ICASSP*, Salt Lake City, May 2001, vol. 3.

[10] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, http://www.athenasc.com/nonlinbook.html, 2nd edition, 1999.

[11] A. Conn, N. Gould, and Ph. Toint, *Trust Region Methods*, SIAM, http://www.fundp.ac.be/p̃htoint/pht/trbook.html, 2000.

[12] C. Lawrence, J. L. Zhou, and A. L. Tits, *User's Guide for CFSQP Version 2.5*, Electrical Engineering Department and Institute for Systems Research, University of Maryland.

[13] P. Frossard, P. Vandergheynst, R.M. Figueras i Ventura, and M. Kunt, "A posteriori quantization of progressive matching pursuit streams," *Submitted to IEEE Trans. on Signal Processing*, 2002.