BIDIMENSIONAL DICTIONARY AND CODING SCHEME FOR A VERY LOW BITRATE MATCHING PURSUIT VIDEO CODER

Fulvio Moschetti†, Kazuo Sugimoto†, Sadaatsu Kato†,Minoru Etoh‡

†NTT DoCoMo, Inc. Multimedia Signal Processing Laboratory, 3-5 Hikarinooka, Yokosuka, Japan
‡DoCoMo Communications Laboratories USA, Inc., 181 Metro Drive, San Jose, CA 95110

ABSTRACT

In this paper we present a video coder based on a redundant dictionary that employs non separable bi-dimensional functions. The algorithm used for the representation of the residual of motion estimation is Matching Pursuit (MP). The proposed dictionary has new features capable of catching curved and oriented contours typical of the residual of motion estimation in natural video sequences. An adaptive grid approach is adopted for the coding of atoms and this is combined with an innovative coding method employing an arithmetic encoder.

The proposed coder shows improvements over H.264 in terms of compression efficiency, with up to 20% gains for very low bitrates.

1. INTRODUCTION

Mobile video peer-to-peer communications and mobile video streaming are dramatically changing the way we think about communications. In this environment, especially in 3G networks for which error resiliency is directly provided at the data link layer, the saving of bandwidth is the key issue. Since the late 80ies and the appearance of H.261[1], video compression efficiency has kept on evolving at a steady rate. Remarkable improvements have been achieved for each component of the coding engine, though the basic architecture, based on block matching-dct-vlc remained almost the same, including the very recent baseline profile version of H.264 [2].

In this paper we propose a video coder based on MP. Differently from the previous MP dictionaries, the one introduced here adopts 2d functions (also called atoms) able to catch oriented edges in the displaced frame difference (dfd) images, namely those images that are the residual signal of the motion estimation. Atoms are arranged on a variable grid and coded using an arithmetic coder whose context adapts to the characteristics of the structure of the source. Results show an improvement of up to 20% when compared with H.264 for bitrates in the range of 20-40 kbs.

2. MATCHING PURSUIT

A detailed and complete explanation of the theory of Matching Pursuit can be found in [3]. Here we briefly recall the basics of the iterative process used for the selection of the waveforms that represent the signal structures.

Let $D = \{g_\gamma\}_{\gamma=\Gamma}$ be a dictionary of unitary norm vectors g_γ called atoms and Γ represent the set of all possible indexes. A Matching Pursuit begins projecting f on a vector $g_{\gamma_0} \in D$ and computing the residual Rf:

$$f = \langle g_{\gamma_0}, f \rangle g_{\gamma_0} + Rf \tag{1}$$

Since Rf and g_{γ_0} are orthogonal, it follows that

$$\|f\|^2 = \left|\langle g_{\gamma_0}, f \rangle\right|^2 + \|Rf\|^2 . \tag{2}$$

To minimize $\|Rf\|$, we must choose g_{γ_0} such that the $\left|\langle g_{\gamma_0}, f \rangle\right|$ projection is maximal. Applying iteratively such a procedure, after N iterations we obtain:

$$f = \sum_{n=0}^{N-1} \langle g_{\gamma_n}, R^n f \rangle g_{\gamma_n} + R^N f \tag{3}$$

where $R^n f$ is the residual at the nth step and $R^0 f = f$.

The convergence of MP depends on both the dictionary and the (sub)optimal search strategy, as expressed hereafter in Eq. (4) that links the norm, therefore the energy decay, of the residual at the step m and $m+1$. The demonstration of Eq. (4) can be found in [3]; it can be shown that there are two real numbers $\alpha, \beta \in]0,1]$ such that for all $m \geq 0$ the following relation is valid:

$$\left\|R^{m+1} f\right\| \leq (1 - \alpha^2 \beta^2)^{1/2} \cdot \left\|R^m f\right\| . \tag{4}$$

α is an optimality factor related to the strategy adopted to determine the best atom in the dictionary, while β strictly depends on the dictionary capability of capturing the features of the input function f.

Being a redundant basis, well-designed MP dictionaries may be able to accurately catch, with very few atoms, most of the energy of the patterns that appear in the dfd signals; namely it is possible a fast convergence, at the very first iterations as it was shown in [4].

Another advantage of the MP is that atoms can be placed on a dfd image wherever needed. In fact, as the coder architecture is not based anymore on a block based DCT structure, atoms can indeed straddle an area crossed by the boundaries of the ideal 8x8 pixel grid.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

3 APPROXIMATION AND DICTIONARY

A fundamental issue in a hybrid video coder architecture is how to code the dfd signal, namely the texture information. Being the latter the result of the mismatch of the motion compensation and prediction, it can contain a considerable amount of edges. Now, let us assume to have a bi-dimensional signal φ supported in $[0,1]^2$ that has a discontinuity along a C^2 curve Ψ and that it is otherwise smooth. Then using a standard Fourier representation, and approximating with $\tilde{\varphi}_m^F$ built from the best *m* nonzero Fourier terms we have:

$$\left\|\varphi - \tilde{\varphi}_m^F\right\|_2^2 \approx m^{-1/2}, m \to \infty \qquad (5)$$

Now this rate is improved by wavelets [5] and the approximate $\tilde{\varphi}_m^W$ obtained from the best *m* nonzero wavelet terms satisfies:

$$\left\|\varphi - \tilde{\varphi}_m^W\right\|_2^2 \approx m^{-1}, m \to \infty \qquad (6)$$

In [6] Candes and Donoho demonstrated, for a class of star shaped objects and using an adaptive dictionary $\tilde{\varphi}_m^A$ of overcomplete polygonal wedges ideally fitted to approximate the shape of the discontinuities, that the following convergence is achievable:

$$\left\|\varphi - \tilde{\varphi}_m^A\right\|_2^2 \approx m^{-2}, m \to \infty \qquad (7)$$

This theoretical demonstration is valid for a class of objects and even though it cannot be simply generalized to any class of bidimensional signals (specifically to the dfd images with which we are dealing) it has the great importance of providing an upper bound for the efficiency of the convergence rate of appropriately designed overcomplete dictionaries. This demonstrates the potential of a tool such as MP, when dictionaries are appropriately designed.

If the main structures of the signal to be represented are included in the dictionary, these same structures will be easily detected and represented at the first steps of the approximation. Therefore the main advantage of an MP based approach can be obtained at very low bitrates: this is what we would expect in the final results.

As for previous attempts with MP based video coders, Neff and Zakhor pioneered an MP coder using a dictionary of separable gabor functions in [7]; their architecture outperformed, at very low bitrates, the state of the art block based DCT coder, at that time MPEG-4[8]. Frossard and Vandergheynst investigated a dictionary based on anisotropic atoms in [9]. They demonstrated the theoretical improvement of such a dictionary over the separable gabor functions by measuring the better representation capability of that dictionary in terms of fewer coefficients needed to reach the same distortion for real images; a complete coder implementation to prove the overall efficiency in a coding system was though missing.

3.1 The proposed dictionary

The most important characteristic to be taken into account for the design of a dictionary is its capacity to identify the main structures of the signal, especially for the sake of the sparsity of the representation, thus for compression purposes. From the observation of the dfd signal we can notice the presence of many elongated and curved edges and patterns, therefore we propose a new dictionary of rotated anisotropic atoms that include a curvature factor that enables a better match with the dfd structure of the 2D signal. The dictionary of bidimensional curved and rotated anisotropic atoms is created by acting on a generating function of unit L^2 norm by means of a set of unitary operators Uγ:

$$D = \{U_\gamma, \gamma \in \Gamma\}, \qquad (8)$$

for a given set of indexes Γ. This set contains four types of operations:

- translation of the coordinates
- anisotropic scaling along the horizontal and vertical direction
- rotation, for the orientation of the atom
- curvature.

The generating function *g* is the following, the combination of a gaussian and a parabola :

$$g(x,y) = \left(2x'^2 - 1\right) \cdot e^{-\left(x'^2 + y^2\right)} \qquad (9)$$

where $x' = x + K\, y^{\lambda}$, (with $K,\lambda \in R$). This last curvature function acts on the coordinate system as an isomorphism of R_+^2, giving to the functions of the anisotropic dictionary a characteristic skew form.

3.2 The β factor: a measurement of dictionary efficiency

Stepping backward for a moment to Eq (4), it expresses the relationship linking the energy decay of the residual at the step *m* and the property of the dictionary. In particular, exploiting this relationship, we have the possibility of measuring the efficiency of the dictionary itself. Namely, for α=1 we can measure β. Using α=1 means using a complete search technique relative to both the position where to place the atom and the exhaustive search of the functions in the dictionary. Therefore, measuring the value of β leads to the proper evaluation of the efficiency of the generating function g_{γ_0} and therefore of the ability of the dictionary to match the most important features present in the 2-D signal. In tab. 1 we can find the values of β for various dictionaries: gabor based (G) with 4096 functions, anisotropic atoms (A) with 4096 and 9216 functions as in [10] and the proposed dictionary (P) with 3888 and 6912 functions). The gains indicated in the graph are relative to the Gabor based dictionary, since the adoption of gabor functions were first suggested in [3]. To perform this test we have used functions with a compact domain of 32x32 pixels. We have used an exhaustive search for the various parameters of the functions and we have measured the ratio in the decrease of the Mean

Square Error of the Residual at the step *m* and *m+1*. We have therefore compared the efficiency of the various dictionaries and the results lead to two interesting observations:

- having a non-separable generating function improves the efficiency of the dictionary when compared to a gabor based one
- each dictionary has a certain efficiency according to the properties of the generating function; increasing the number of functions that form the dictionary marginally alter the value of β.

Indeed the second observation comes from the fact that even though the proposed dictionary P6912 has less functions than that of the A9216 (see Tab. 1), its efficiency in terms of signal representation is higher. And for both category of dictionaries A and P for which we have considered a different generating function, the efficiency in terms of β is more linked to the characteristic of the generating function itself rather than to the number of atoms that compose the dictionary.

4 CODER ARCHITECTURE

The main structure of the proposed coder is shown in Fig. 1. We can easily recognize the traditional hybrid coder architecture where though some significant changes are noticeable. In particular, MP is the tool used to represent the texture of the dfd image and it replaces the DCT. As for the entropy coding we adopted an arithmetic coder for both the texture information (therefore the atoms) and the Motion Vectors (MVs). An adaptive grid is used to code the atoms. In the following sub-section more details will be given on the coding methodology, hereafter we describe the criterion adopted to position the atoms in a frame.

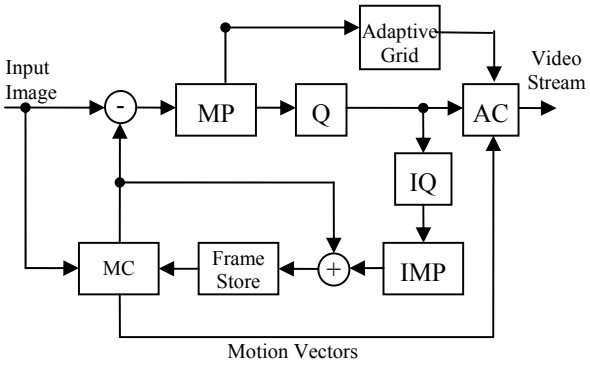

Fig. 1 Video coder architecture

4.1 Atom search criterion

The proper application of the MP algorithm would require to look for all the possible positions in the dfd frame for every function of the dictionary. This would however cause an explosion of complexity, as already observed in [7]. Therefore a pre-scan of the dfd frame is necessary. In particular the MSE between the original and motion compensated frame is computed. The pocket of highest energy (pockets of 16x16 pixels area) is selected and later the 2D functions of the base are accordingly correlated with the selected dfd area. The atom giving the higher correlation is selected and the correlation coefficient, together with other parameters such as position and index of the function are transmitted to the decoder.

4.2 Atom coding technique

Atom coding is realized employing a grid that divides the frame in equally sized units. Our approach adopts a context-based adaptive arithmetic coder. Each of the grid unit comprises the following information:

- flag identifying the presence or not of an atom AF
- number of atoms: AN
- atom modulus after quantization: AM
- position of each atom: AP
- index of the basis for each atom: IB

AF and AN are coded according to a context adopted in the arithmetic coder.

Because of the orthogonality between Rf and g_{γ_0} in (1), the coding order of atoms can be opportunely chosen. Atoms in the same grid unit are sorted according to their amplitude in a descending order; the sign of AtomMod and the difference between two successive absolute values of AtomMod are then coded.

Grid size has no impact on IB whose coding efficiency depends merely on the dictionary.

APs are coded with fixed length code, as their positions appear to be random values in a grid unit. As APs have to be transmitted for the horizontal and vertical coordinate, they have a noticeable impact on the overall amount of bits needed to code the atoms. Obviously, the smaller the size of the grid, the fewer bits needed to code the APs; in this case though more bits to code AF and AN will be needed. A tradeoff is necessary, as AP and AN may substantially vary from frame to frame in a sequence, depending on the motion and on the target coding quality. As a consequence, employing a fixed grid has the drawback of a lack of adaptability to the evolution of the sequence. For this reason we adopted an adaptive grid approach, namely the size of the grid unit can be changed from frame to frame and is selected on the basis of rate optimization. The size of grid unit is signaled in the bitrate with a flag.

5. RESULTS

We have compared the proposed coder with the H.264 baseline profile, which is the profile conceived for mobile applications. Moreover, in the comparisons CABAC (the context adaptive arithmetic coder) has been selected in H.264 as the proposed coder also employs an arithmetic coder. The size of the dictionary used in the tests was 1024 functions selected from the P6912 dictionary with a vector quantization approach similar to that described in [11].

The 30Hz QCIF sequences for the tests have been selected from the H.264 official test set, among them the typical head and shoulder sequences (eg. Foreman and

News) are used as test sequences as typical applications of video in the mobile communication environment. The search window for both coders was fixed to +/- 15 pixels.

Tests have been carried out at very low bitrates (20-40 kbs), which in terms of quality translates into a PSNR of up to 31 dB. The graphs in Fig. 2 and Fig. 3 show that the proposed coder can outperform H.264 at bitrates of around 20-30kbs up to 20% in terms of compression efficiency.

Results reported here are typical for videocommunication sequences. Graphs also show a clear pattern of convergence of the two coders towards higher quality. This is also a reasonable behavior and in line with the results obtained by a previous MP coder, namely [7]; the main advantage of overcomplete basis lies in fact in the property that at the very first approximations very few coefficients can extract most of the energy, providing a powerful and efficient approximation [4, 7].

For completeness we underline that comparisons have been made against H.264 widely recognized by the community as the best video coder available at the moment. Comparisons against the gabor based dictionary MP coder are not possible as the coder is not publicly available. Though, the MP based coder proposed in [7] was around 10-20% better in terms of coding efficiency than MPEG4 at very low bitrates (10kbs-40kbs). H.264 is around 30-40% better (sometimes more) in terms of coding efficiency than MPEG4 at these same rates.

6. CONCLUSIONS

In this paper we have introduced a new bidimensional dictionary and a coding scheme for an MP based video coder. The dictionary has some new features, like the orientable curved contour, that provide considerable advantages for the representation of the dfd frames in natural sequences. This was proved by measuring the theoretical efficiency. The coding scheme differs from the previous MP coder proposed in [7] and it employs an arithmetic coder.

When compared to the state of the art it shows a better compression performance (up to 20% better) for typical mobile communication sequences. The main advantage is obtained in the very low rates (the ones used for wireless video communications), that are also the ranges of application where MP can provide noticeable advantages over an orthogonal transform [4], given its ability to extract big pockets of energy with few coefficients.

7. REFERENCES

[1].CCITT, *Recommendation H.261 Video Codec for Audiovisual Services at px64 kbit/s*. 1990 Aug., Geneva.

[2].Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, *Draft ISO/IEC 14496-10:2002*. 2002.

[3].Mallat S. and Zhang Z., *Matching Pursuit with Time-Frequency Dictionary*. IEEE Trans. on Signal Processing, December 1993. **vol. 41**(n. 12): p. 3397-3415.

[4].Mark R. Banham and J.C. Brailean, *A selective update approach to matching pursuit video coding*. IEEE Trans. on Circuits and Systems for Video Technology, 1997. **vol. 7**(no. 1): p. 119-129.

[5].Vetterli M., *Wavelets, Approximation and Compression*. IEEE Signal Processing, 2001. **vol. 18**(5): p. 59-73.

[6].Candes J. and Donoho D., *A Surprisingly Effective Nonadaptive Representation for Objects with Edges, Curves and Surfaces*. Vanderbilt University Press, Nashville,, 1999.

[7].Neff R. and Zakhor A., *Very Low Bit-Rate Video Coding Based on Matching Pursuit*. IEEE Trans. Circuits Syst. Video Technol., February 1997. **vol. 7**(no. 1): p. 158-171.

[8].JTC1/SC29/WG11/MPEG97/N1902, I.I., *Committee Draft of ISO/IEC 14496-2 (MPEG-4 Visual)*. November 1997.

[9].Frossard P. and Vandergheynst P. *Redundancy in Non-Orthogonal Transforms*. in *ISIT*. 2001. Washington DC.

[10].Vandergheynst P. and Frossard P. *Efficient Image Representation by Anisotropic Refinement in Matching Pursuit*. in *ICASSP 2001*. 2001. Salt Lake City.

[11].Philippe Schmid-Saugeon and A. Zakhor. *Learning dictionaries for matching pursuits based video coders*. in *ICIP 2001*. 2001. Thessaloniki, Greece.

	Mobile	Foreman	News	Stefan
G 4096	0.333	0.42	0.44	0.3441
A 4096	0.372	0.496	0.51	0.379
β improv.	11.70%	18.00%	15.90%	10.10%
A 9216	0.373	0.497	0.513	0.38
β improv.	12%	18.33%	16.60%	10.40%
P 3888	0.3932	0.516	0.529	0.397
β improv.	18%	22.90%	20.20%	15.40%
P 6912	0.3939	0.518	0.542	0.4
β improv.	18.30%	23.30%	23.20%	16.30%

Tab. 1 Efficiency of the various dictionaries measured as β of Eq. 4 (for α=0)

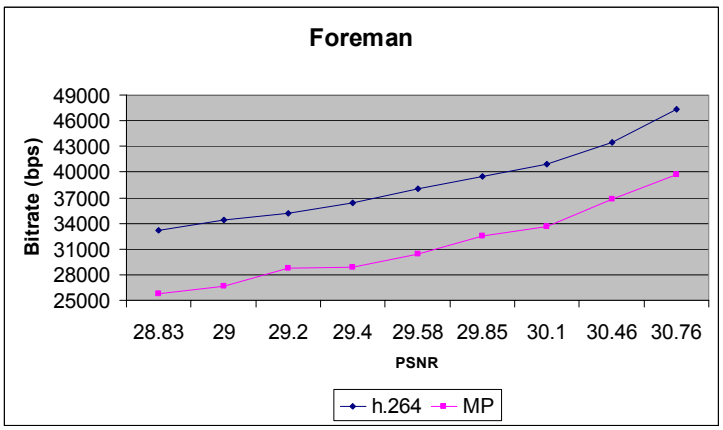

Fig. 2 Rate versus PSNR for the sequence Foreman

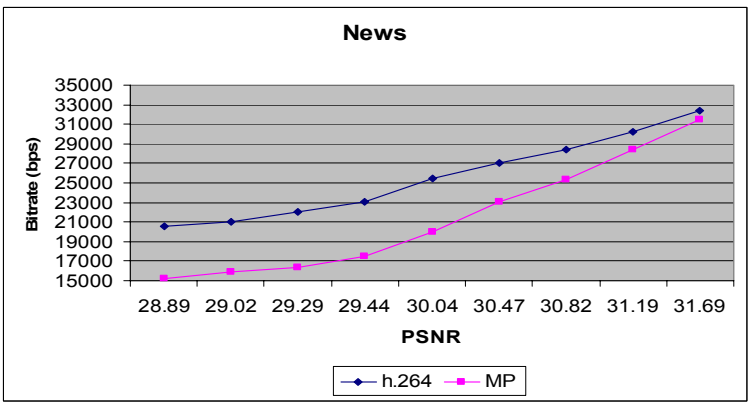

Fig. 3 Rate versus PSNR for the sequence News