SPARSE STEREO IMAGE CODING WITH LEARNED DICTIONARIES

Dimitri Palaz[†]*, Ivana Tošić*[‡] *and Pascal Frossard*[†]

[†]Signal Processing Laboratory (LTS4), Ecole Polytechnique Fédérale de Lausanne (EPFL)
Lausanne 1015, Switzerland
[‡] Helen Wills Neuroscience Institute, University of California, Berkeley
Berkeley, CA 94720-3810, USA

ABSTRACT

This paper proposes a framework for stereo image coding with effective representation of geometry in 3D scenes. We propose a joint sparse approximation framework for pairs of perspective images that are represented as linear expansions of atoms selected from a dictionary of geometric functions learned on a database of stereo perspective images. We then present a coding solution where atoms are selected iteratively as a trade-off between distortion and consistency of the geometry information. Experimental results on stereo images from the Middlebury database show that the new coder achieves better rate-distortion performance compared to the MPEG4-part10 scheme, at all rates. In addition to good rate-distortion performance, our flexible framework permits to build consistent image representations that capture the geometry of the scene. It certainly represents a promising solution towards the design of multi-view coding algorithms where the compressed stream inherently contains rich information about 3D geometry.

Index Terms— Stereo image coding, sparse approximation, 3D geometry representation.

1. INTRODUCTION

Imaging applications built on stereo and multiview streams are becoming very popular with the recent advent of services that offer increased interactivity such as free viewpoint TV (FTV) or richer content like 3DTV. In general, multiple views are compressed by encoding algorithms that capture the inter-stream redundancy with block-based motion compensation or disparity estimation [1]. Multiview compression using disparity information has also been considered in distributed video coding solutions such as [2]. However, since these schemes are based on the classical compression approaches that involve orthogonal transforms and block-based matching, they are not ideal in representing 3D geometry, especially at low coding rates.

Sparse approximations, on the other hand, offer increased flexibility in image representation by using overcomplete dictionaries. Studies of sparse models for correlated signals (such as multi-view or stereo images) that are based on joint sparsity models are usually limited to decompositions with the same support, where the difference between signals is noise [3]. These models are not appropriate for the representation of stereo images since they cannot capture the geometry information. Recently, we have proposed a geometry-based sparse stereo image model for representation of multivew/stereo omnidirectional images, and developed an algorithm that trades-off approximation and geometric consistency in the model [4].

This paper addresses the problem of stereo image coding for perspective cameras. We first show that the geometry-based correlation model is also valid for pairs of perspective images, by deriving geometric constraints adapted to the camera geometry. We further exploit this model in the design of a full stereo image coder, which achieves efficient compression and allows implicit geometry representation in the encoded stream. The coder uses the Multi-View Matching Pursuit (MVMP) algorithm, which finds pairs of atoms in stereo images that correspond to the same 3D features in a scene [4]. Atoms are selected from dictionaries learned under geometric constraints. The proposed coder includes a coefficient quantization step and an entropy coding step. We show that the rate-distortion performance of the proposed coder is better than the baseline MPEG4-part10 coder. Moreover, our coder outperforms independent coding of images using Matching Pursuit [5] at lower rates, for up to 1dB. It thus represents a flexible framework for stereo image coding, that offers good coding performance and allows implicit representation of 3D geometric atom correspondences in the compressed stream.

2. SPARSE STEREO IMAGE REPRESENTATION

Efficient image coding significantly relies on the image representation methods. Although single image and video representations have been widely studied in literature, stereo and multiview image representation models have just started to gain interest. Recently, we have proposed a sparse stereo image representation method that exploits 3D geometry implicitly contained in stereo images [4]. Although initially applied to omnidirectional images, this method can be adapted to perspective images taken by two pinhole cameras in a 3D scene, by appro-

I. Tošić is supported by the Swiss National Science Foundation under the fellowship no:PBELP2-127847.

priately incorporating the planar geometric constraints. This section first overviews the method of [4], and then introduces the specific changes inherent to pinhole camera geometry.

2.1. Joint geometry-based sparsity model

In the sparse stereo model proposed in [4], the images $\mathbf{y}_L$ and $\mathbf{y}_R$ have m-sparse approximations in dictionaries $\mathbf{\Phi}$, resp. $\mathbf{\Psi}$, of size M, up to an approximation error $\mathbf{e}_L$, resp. $\mathbf{e}_R$:

$$\begin{aligned} \mathbf{y}_L &= \mathbf{\Phi a} + \mathbf{e}_L = \sum_{k=1}^{m} a_{l_k} \phi_{l_k} + \mathbf{e}_L \\ \mathbf{y}_R &= \mathbf{\Psi b} + \mathbf{e}_R = \sum_{k=1}^{m} b_{r_k} \psi_{r_k} + \mathbf{e}_R, \end{aligned} \quad (1)$$

where the vectors $\mathbf{a}$ and $\mathbf{b}$ represent the coefficients for the left and right image, respectively. The index sets $\mathcal{L} = \{l_k\}, \mathcal{R} = \{r_k\}, k = 1, ..., m$ label the atoms that participate in the sparse decompositions of $\mathbf{y}_L$ and $\mathbf{y}_R$, respectively. In other words, $\{l_k\}, \{r_k\}, k = 1, ..., m$ denote the atoms with non-zero coefficients, i.e., $a_{l_k} \neq 0$ and $b_{r_k} \neq 0$. Since images $\mathbf{y_L}$ and $\mathbf{y_R}$ capture the same 3D scene from different viewpoints, there is geometric correlation between them. The model further assumes that $\mathbf{y}_L$ and $\mathbf{y}_R$ are correlated in the following way:

$$\mathbf{y}_R = \sum_{k=1}^{m} b_{r_k} \psi_{r_k} + \mathbf{e}_R = \sum_{k=1}^{m} b_{r_k} F_{l_k r_k}(\phi_{l_k}) + \mathbf{e}_R, \quad (2)$$

where $F_{l_k r_k}(\cdot)$ denotes a local geometric transform of an atom ϕ_{l_k} in $\mathbf{y}_L$ to an atom ψ_{r_k} in $\mathbf{y}_R$, and it differs for each $k = 1, ..., m$. This correlation can be nicely captured by using parametric dictionaries, built by applying geometric transformations to a generative function $g(x, y)$. In this case, an atom ϕ is given by g_γ, where $\gamma = [s_x, s_y, \theta, t_x, t_y]$ is the set of parameters that include scaling (s_x, s_y), rotation (θ) and translation (t_x, t_y). Due to such dictionary construction, applying these geometric transforms (translation, rotation, scaling, or their combination) on an atom in the sparse image representation becomes equivalent to a transform of its parameters.

Since transformations between pairs of atoms in stereo image expansions are due to the common 3D geometry of the scene, they have to satisfy epipolar geometry constraints [6]. The formulation of these constraints obviously depends on the camera geometry. Therefore, we derive in the next section explicit geometric constraints for perspective images, which will then be incorporated in the described stereo image model.

2.2. Geometric constraints in the perspective image model

Let two points $\mathbf{v}$ and $\mathbf{u}$ represent image projections of the same 3D point p on the left and right camera, respectively. We denote the essential matrix between cameras as $\mathbf{E}$, which depends on the rotation and translation between cameras [6]. These points satisfy the epipolar geometry constraint when $\mathbf{uEv} = 0$. Let $\mathbf{v}$ lie on the atom ϕ_l and $\mathbf{u}$ lie on the atom $\psi_r = F_{lr}(\phi_l)$. Since we consider parametric dictionaries built on the same generating function, transforming the atom ϕ_l with F_{lr} reduces to a linear transform of the coordinate system $Q_{lr}(\cdot)$, i.e., $\mathbf{u} = Q_{lr}(\mathbf{v})$. This transform immediately follows from translation $\mathbf{t}$, rotation $\mathbf{R}$ and anisotropic scaling s_x, s_y applied on the x-y coordinates $\mathbf{u} = [x \;\; y]^T$, i.e.,: $\tilde{\mathbf{u}} = \mathbf{S} \cdot \mathbf{R}(\mathbf{u} + \mathbf{t})$, where

$$\mathbf{S} = \begin{bmatrix} 1/s_x & 0 \\ 0 & 1/s_y \end{bmatrix}, \mathbf{R} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \mathbf{t} = \begin{bmatrix} t_x \\ t_y \end{bmatrix}, \quad (3)$$

and it has the form:

$$\mathbf{u} = Q_{rl}(\mathbf{v}) = \mathbf{S}_l \mathbf{R}_l \mathbf{R}_r^{-1} \mathbf{S}_r^{-1} \cdot \mathbf{v} - \mathbf{S}_l \mathbf{R}_l (\mathbf{t}_r - \mathbf{t}_l), \quad (4)$$

where $(\mathbf{S}_l, \mathbf{R}_l, \mathbf{t}_l)$ and $(\mathbf{S}_r, \mathbf{R}_r, \mathbf{t}_r)$ denote the transformation matrices corresponding to the parameters of the left and right atoms, respectively. In general, $\mathbf{u}$ and $\mathbf{v}$ will not satisfy the epipolar constraint exactly, but up to an error $d_l = [Q_{rl}(\mathbf{v})]^T \mathbf{E} \mathbf{v}$ evaluated on the left image coordinate system, or error $d_r = [Q_{rl}^{-1}(\mathbf{u})]^T \mathbf{E} \mathbf{u}$ evaluated on the right image coordinate system. Taking the average of these two errors and summing them over all pixels, we obtain the epipolar distance between two atoms ϕ_l and ψ_r in the planar geometry:

$$W_{lr} = \sum_{i=1}^{q} \left(w_l^{[i]} (d_l^{[i]})^2 + w_r^{[i]} (d_r^{[i]})^2 \right), \quad (5)$$

where $w_l^{[i]} = w_l(x_i, y_i)$ is a weighting function that favors points that are closer to a geometric discontinuity (e.g., a Gaussian envelope). Due to the geometric correlation between stereo images, transforms F_{lr} between corresponding atoms in the model (2) will have small epipolar distance W_{lr}.

3. SPARSE STEREO IMAGE CODING

Besides achieving redundancy reduction using sparse representation, the described stereo image model for perspective cameras also carries information about the 3D geometry, implicitly contained in the transform of parameters. We therefore propose to use this representation in a joint stereo image encoding scheme, which at the same time reduces the number of bits required for image transmission and transmits the 3D geometry information encoded in the transforms. The block scheme of the proposed joint encoder is shown in Figure 1.

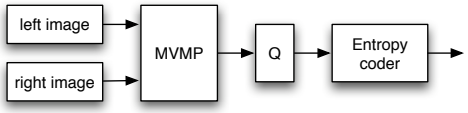


Fig. 1. Joint stereo image encoder.

Stereo images are first processed by the MVMP algorithm (Multi-View Matching Pursuit) [4], which decomposes them into linear combinations of atoms that follow the described sparse stereo image model (Eqs. 1 and 2). MVMP is a greedy

algorithm that selects at each iteration a stereo pair of atoms that gives the minimal value of the following energy:

$$E = \frac{1}{2\sigma_I^2}(||\mathbf{y}_L - \mathbf{\Phi a}||_2^2 + ||\mathbf{y}_R - \mathbf{\Psi b}||_2^2) + \rho \left[\sum_{l,r=1}^{M} \mathcal{I}(a_l)\mathcal{I}(b_r)W_{rl} + \kappa \sum_{l,r=1}^{M} (b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 \right], \quad (6)$$

where W_{rl} is given by Eq. 5, κ is a normalization parameter, and $\mathcal{I}$ is the activity indicator function: $\mathcal{I}(x) = 0$ if $x = 0$ and $\mathcal{I}(x) = 1$ otherwise. J_{lr} is the Jacobian determinant of the transform Q_{lr}, which in the case of perspective images is:

$$J_{rl} = \left| \frac{\partial Q_{lr}(\mathbf{v})}{\partial \mathbf{v}} \right| = \left| \begin{matrix} s_{x,r} & s_{y,r} \\ s_{x,l} & s_{y,l} \end{matrix} \right|.$$

Finally, ρ represents a trade-off parameter between the approximation error and the geometric penalty given by the epipolar matching of atoms and the correlation of their coefficients. After finding the atoms, MVMP removes their contributions from the images, and repeats the selection process on the residues.

The dictionary used in the MVMP can be any parametric dictionary. However, to achieve the best coding performance, we propose to use dictionaries whose sets of parameters are learned from a database of stereo images. We adapt the maximum likelihood dictionary learning method proposed in [4] to the case of perspective images. The dictionary parameters are optimized by iterating between two steps: 1) sparse approximation: where sparse coefficients are computed for a large set of images using MVMP and a fixed set of dictionary parameters; and 2) dictionary update, where the dictionary parameters are updated while keeping the coefficients constant. The second step is done using the multivariate gradient descent.

Coefficients obtained by the MVMP are then quantized, as shown in Figure 1. Besides the simplest uniform quantization approach, we also propose to use vector quantization, whose goal is to exploit the correlation between the coefficients of corresponding atoms in two views. The optimal 2D bins and centroids are evaluated by the Lloyd-Max (i.e., K-means) algorithm. Quantized coefficients are then entropy coded.

Atom indexes can be represented as a combination of parameters, corresponding to the geometric transformation of the generative function. For a given atom, four parameters have to be encoded : the scale pair index (each pair of scales gets one index), the rotation index and the two shift indexes. Since the parameters of two corresponding atoms in the left and right image are correlated (through a local transform), we compute their differences and perform entropy coding on those differences.

Finally, the decoding scheme involves only simple linear summations over the decoded atoms, weighted by dequantized coefficients. Note that the encoding of parameters is lossless, hence the transforms Q_{lr} between corresponding stereo atoms are available at the decoder in the original form. This is important, as those transforms carry geometric information that can be used for camera pose or depth estimation [7].

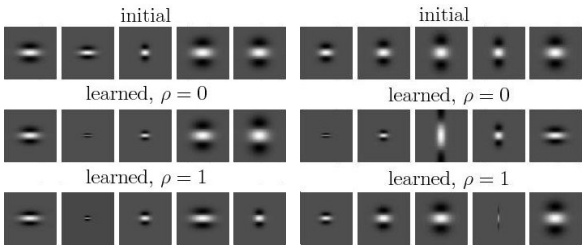


Fig. 2. Learned dictionaries (10x10): a) Left, $\mathbf{\Phi}$; b) Right, $\mathbf{\Psi}$

4. EXPERIMENTAL RESULTS

We first show the results obtained by the dictionary learning algorithm and then show the rate-distortion performance of the proposed stereo image coder.

4.1. Learned dictionary

Dictionary learning is performed using 21 planar stereo image pairs from the Middlebury 2006 dataset [8]. As the images are rectified and without radial distortion, the essential matrix is given by [6] : $E = [0\ 0\ 0;\ 0\ 0\ -1;\ 0\ 1\ 0]$.

We learn the parameters of the parametric dictionary built on the generative function that is a Gaussian in one direction and its second derivative in the orthogonal direction:

$$g(x, y) = -\frac{1}{K}(4x^2 - 2)e^{-(x^2+y^2)}. \quad (7)$$

This function has been shown to be well suited for representing edges in images. We only learn scaling parameters, since other parameters depend on the relative pose between cameras. Five samples of the scale parameters are used, initialized by a random value in the interval $[5, 15]$. We use four orientations uniformly distributed from 0 to π, and translations that cover all pixel shifts. In each iteration of learning, 50 pairs of patches of size 10×10 are randomly selected from the image database. The preprocessing of patches includes whitening and variance normalization [4]. Patches are then decomposed by MVMP (using 12 atoms/patch) in order to find coefficients used for the dictionary update step. The process is stopped when the stable solution is found.

We train dictionaries for two cases: 1) $\rho = 0$, i.e., learning only under the approximation constraint; and 2) $\rho = 1$, i.e., learning under both approximation and geometry constrains. Learned atoms are shown in Figure 2 for the left and right dictionary, displayed at the center and with orientation 0. We can see that the atoms learned under the geometric constraints tend to be more spatially compact than the atoms learned in the unconstrained case ($\rho = 0$). However, it is hard to tell which atoms are better using qualitative assessment. Hence, we evaluate in the next section the coding performance of learned dictionaries in the stereo image coding scheme proposed in Sec. 3.

4.2. Sparse stereo image coding results

Coding performance is evaluated on the Moebius stereo pair from the Middlebury database [8], which is outside the training set. Coefficients are quantized using uniform quantization (UQ) or vector quantization (VQ). VQ was performed using the K-means algorithm on 500 training samples. Quantized coefficients and indexes are encoded with the Huffman algorithm, using probability models based on 500 randomly chosen patches.

Figure 3 shows the rate-distortion curves for the proposed joint sparse stereo coding scheme using UQ (red curve) and VQ (black curve). The rate is the total rate for encoding both left and right images, while the PSNR is calculated on the average mean square error over the two images. Both curves correspond to the performance of the dictionary evaluated for $\rho = 1$. We can see that UQ and VQ perform comparably. VQ performs slightly better at low rates, when the most of the geometric correlation that leads to the correlation of coefficients is exploited.

For comparison, the stereo pair is encoded using MPEG4-part10 (high profile, level 1.2), without deblocking filter (green curve), where the right image is encoded with respect to the first one, using block-based motion estimation. We also evaluate the performance of the independent coding scheme, which is based on the independent Matching Pursuit (MP) encoding of each image, UQ, and Huffman coding (blue curve). The coder uses the dictionary optimized for minimal approximation error, since MP does not use geometric constraints.

The proposed sparse stereo coding scheme outperforms MPEG4 at all rates in terms of PSNR, and provides competitive performance in terms of visual quality. Moreover, it outperforms independent MP at low rates, showing improvements of up to 1dB. At higher rates, MP has better performance, which is due to the fact that the geometric correlation between atoms is prominent at the beginning of the MVMP, i.e., at lower rates.

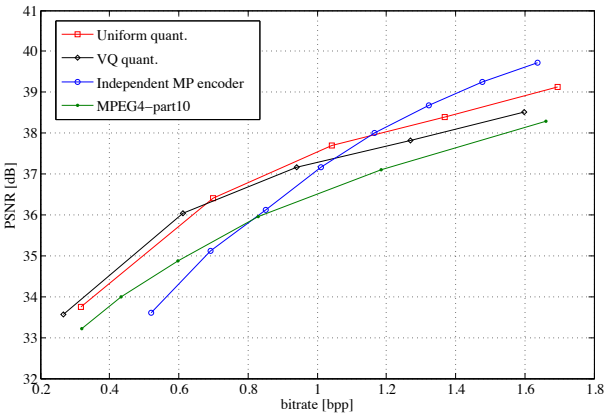
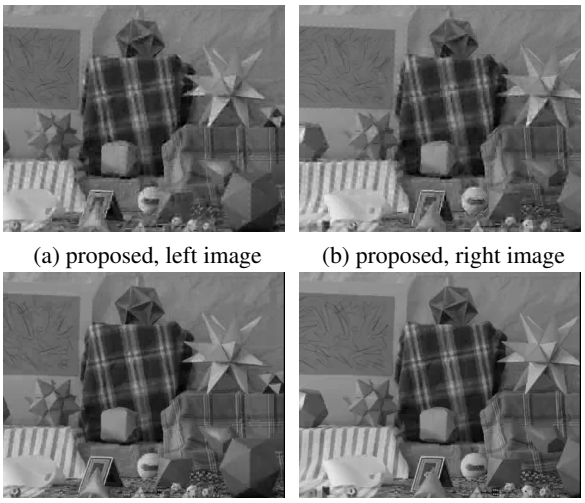


Fig. 3. Rate-distortion performance (Moebius).



(a) proposed, left image (b) proposed, right image

(c) MPEG4, left image (d) MPEG4, right image

Fig. 4. Visual comparisons at 0.5 bpp: (a-b) proposed coder, mean PSNR 35.2dB, (c-d) MPEG4, mean PSNR 34.3dB.

5. CONCLUSIONS

We have presented a novel stereo image coder based on joint sparse approximation under geometry constraints. Our first contribution is the adaptation of the MVMP and the dictionary learning algorithm developed for omnidirectional cameras [4] to the case of perspective cameras, by deriving geometric constraints adequate for planar epipolar geometry. Our second contribution is the development of the entire joint sparse coding scheme for stereo images, which leads to improved rate-distortion performance compared to the independent MP coder at low rates, and to the MPEG4 coder at all rates.

6. REFERENCES

[1] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. B. Akar, G. Triantafyllidis and A. Koz, "Coding Algorithms for 3DTV - A Survey", *IEEE Trans. on Circuits and Systems for Video Technology*, 17(11):1606–1621, 2007.

[2] B. Song, E. Tuncel and A.K. Roy-Chowdhury, "Towards A Multi-Terminal Video Compression Algorithm By Integrating Distributed Source Coding With Geometrical Constraints". *Journal Of Multimedia*, 2(3):9, 2007.

[3] M. B. Wakin, M. F. Duarte, S. Sarvhotam, D. Baron, and R. G. Baraniuk, "Recovery of Jointly Sparse Signals from Few Random Projections", *Neural Information Processing Systems*, 2005.

[4] I. Tošić and P. Frossard, "Dictionary learning for stereo image representation", *IEEE Trans. on Image Processing*, 2010, in press.

[5] S. G. Mallat and Z. Zhang, "Matching Pursuits With Time-Frequency Dictionaries", *IEEE Trans. on Signal Processing*, 41(12):3397-3415, 1993.

[6] R. Hartley and A. Zissermann, "Multiple View Geometry in Computer Vision", *Cambridge University Press*, 2003.

[7] I. Tošić and P. Frossard, "Coarse scene geometry estimation from sparse approximations of multi-view omnidirectional images", *Proceedings of the European Signal Processing Conference*, 2007.

[8] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, 47(1/2/3): 7–42, 2002.