

---

SCHOOL OF ENGINEERING - STI  
SIGNAL PROCESSING INSTITUTE  
*Ivana Tasic and Pascal Frossard*

---

CH-1015 LAUSANNE

*Telephone: +4121 6934712*

*Telefax: +4121 6937600*

*e-mail: ivana.tasic@epfl.ch*



ÉCOLE POLYTECHNIQUE  
FÉDÉRALE DE LAUSANNE

# **GEOMETRY-BASED SCENE REPRESENTATION WITH DISTRIBUTED VISION SENSORS.**

**Ivana Tasic and Pascal Frossard**

Swiss Federal Institute of Technology Lausanne (EPFL)

Signal Processing Institute Technical Report

TR-ITS-2007.11

August 20th, 2007

Part of this work has been submitted to the IEEE Transactions on Image Processing.

This work has been partly supported by the Swiss National Science Foundation, under grant 20001-107970/1.

# Geometry-based scene representation with distributed vision sensors.

Ivana Tasic and Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne (EPFL)  
Signal Processing Institute, Lausanne 1015, Switzerland  
{ivana.tasic, pascal.frossard}@epfl.ch  
Fax: +41 21 693 7600, Phone: +41 21 693 4712

**Abstract**— This paper addresses the problem of efficient representation and compression of scenes captured by distributed vision sensors. We propose a novel geometrical model to describe the correlation between different views of a three-dimensional scene. We first approximate the camera images by sparse expansion over a dictionary of geometric atoms, as the most important visual features are likely to be equivalently dominant in images from multiple cameras. The correlation model is then built on local geometrical transformations between corresponding features taken in different views, where correspondences are defined based on shape and epipolar geometry constraints. Based on this geometrical framework, we design a distributed coding scheme with side information, which builds an efficient representation of the scene without communication between cameras. The Wyner-Ziv encoder partitions the dictionary into cosets of dissimilar atoms with respect to shape and position in the image. The joint decoder then determines pairwise correspondences between atoms in the reference image and atoms in the cosets of the Wyner-Ziv image. It selects the most likely correspondence among pairs of atoms that satisfy epipolar geometry constraints. Atom pairing permits to estimate the local transformations between correlated images, which are later used to refine the side information provided by the reference image. Experiments demonstrate that the proposed method leads to reliable estimation of the geometric transformations between views. The distributed coding scheme offers similar rate-distortion performance as joint encoding at low bit rate and outperforms methods based on independent decoding of the different images.

## I. INTRODUCTION

Vision sensor networks have recently been gaining popularity as they find many applications in fields as diverse as 3DTV, surveillance or robotics. These imaging or information processing systems rely on an efficient representation of 3D scenes that includes depth or more generally geometry information. Distributed camera networks actually offer simple and cost effective solutions for scene acquisition, where several views of the scene can be combined to produce a complete representation or to generate new views by interpolation. Bandwidth or power limitations typically impose a distributed processing of the visual information, where rate-distortion effective scene representations take benefit of the correlation from multiple views in order to reproduce depth and visual information.

This work has been partly supported by the Swiss National Science Foundation, under grant 20001-107970/1.

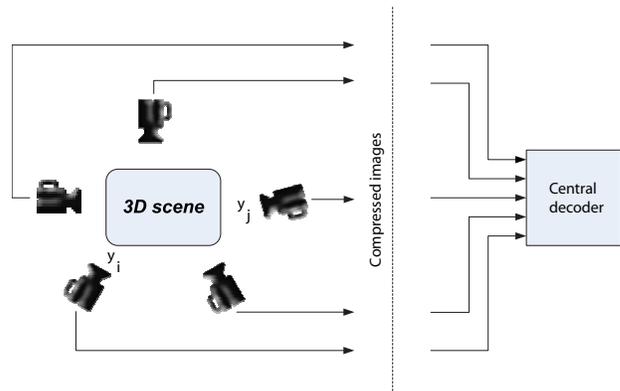


Fig. 1. Distributed coding of 3D scenes. Multiple correlated views  $y_i$  of the scene are encoded independently, and decoded jointly by the central decoder.

In this paper, we consider a framework where a central decoder reconstructs the 3D scene information based on multiples images encoded by distributed cameras (see Figure 1). Distributed coding of the camera images seems a priori suboptimal for a rate-distortion efficient representation of the scene. Interestingly enough, results from information theory have shown that it is possible to exploit the correlation among sources without communication between encoders, as long as the decoding is performed jointly [1], [2]. Distributed coding however relies on the knowledge of a good correlation model between information sources, which is a quite strong assumption in imaging problems. Most DSC schemes that are applied to video coding are based on translational motion estimation at decoder and channel coding at encoder, which assumes a correlation on the level of pixel bit planes modeled by the statistics of a virtual channel. However, the correlation between images in camera networks mostly lies in the motion of the objects in a 3D scene, and translational motion of observed objects cannot cope efficiently with local transforms such as scaling or rotation.

We propose a novel geometry-based correlation model for the design of distributed coding algorithms in camera networks. The main features of a 3D scene are likely to be dominant in the multiple correlated views of the scene, possibly under some transformations due to the geometry of the scene. We propose to capture these features by sparse image expansion with geometrical atoms taken from a redundant

dictionary of functions. The correlation model is then built on local geometrical transformations between corresponding features taken in different different views, where correspondences are defined based on shape and epipolar geometry constraints. Successful pairing of correlated atoms relies on the use of a structured dictionary that is invariant to local transforms like translation, rotation and scaling, or any combination of those. We apply this new correlation model to omnidirectional images that are particularly interesting for scene representation due to their wide field of view and accuracy in capturing the scene geometry. Such images can easily be mapped and processed on spherical manifolds, hence we compute sparse image approximations on the sphere [3] in order to capture the most prominent image components. Local geometrical transformations of atoms then proceed by scaling and rotation on the sphere. It leads to an effective correlation model that can be used to estimate the disparity map between different views for scene rendering or multi-view coding.

The geometrical framework is then used in the design of a distributed coding method with side information for multi-view omnidirectional images. A Wyner-Ziv coder is designed by partitioning the redundant dictionary into cosets based on atom dissimilarity. The joint decoder then selects the best candidate atom within the coset with help of the side information image. The correspondences that are found during decoding between atoms in both image expansions are further used to estimate local transformations and to build a transform field between correlated views. These transformations are used to refine the side information for decoding the following atoms. Experimental results show that the proposed method successfully identifies the local geometric transformations between sparse image components in different views and implicitly provides coarse scene geometry information. Finally, the distributed coding scheme is shown to outperform independent coding strategies and to approach the performance of a joint coding strategy at low bit rate.

The paper is organized as follows. A brief overview of related work on distributed source coding is given in Section II. Section III presents the novel geometrical correlation model in multi-view images, which is further refined for the case of omnidirectional images in Section IV. The Wyner-Ziv coding method that relies on the novel correlation model is described in Section VI and coding results are discussed in Section VII.

## II. RELATED WORK

Distributed source coding (DSC) has been researched for a long time in the information theory community, but its application to imaging problems has been delayed due to the difficulty of finding good models for the correlation between real sources. The first practical DSC schemes for images have been proposed only recently, when the link of DSC with channel coding has been established [4]. Most of the research in the DSC framework till nowadays focused on the application of DSC to low-complexity video coding [5], [6] and error-resilient video coding [5], [7]. However, only few works have addressed the problem of distributed coding in camera networks, mainly due to the difficulty of modeling the

statistical correlation among distributed cameras for 3D scene representation.

The application of DSC principles in camera networks is generally based on the disparity estimation between views under epipolar constraints. Most of the solutions proposed in the literature are built on coding with side information that is a special case of DSC. For example, cameras can be divided into conventional cameras that perform independent image coding and Wyner-Ziv cameras that use DSC coding [8]. The Wyner-Ziv images are decoded with the help of disparity estimation and interpolation from independent views. Shape adaptation is used to enhance the side information with the shape information sent by the encoders. Super-resolution techniques have been also applied to distributed coding in camera networks [9]. Low-resolution images from each camera are combined after registration at the joint decoder into a high-resolution image. The image registration is performed by shape analysis and image warping with respect to the shape transforms that are however limited to only simple translations and rotations. In [10] the authors propose a distributed coding scheme for camera networks where the multi-view correlation is modeled by relating the locations of discontinuities in the polynomial representation of image scanlines. To the best of our knowledge, this scheme has however not been extended to the case of natural 2D images.

Disparity-based solutions have also been proposed for multi-view video compression. In [11], the authors propose a DSC method for highly correlated image sequences that combines distributed video coding applied to motion-compensated temporal wavelet coding and disparity compensation for distributed multi-view compression. Authors in [12] present a transform-based DSC method for multi-view video coding that tracks epipolar correspondences between macroblocks in different views. The Wyner-Ziv encoder has however partial access to the side information (Intra macroblocks and motion vectors), so that this scheme cannot be classified as fully distributed multi-view coding scheme. On the other side, a completely distributed stereo-view video coding method is proposed in [13]. It performs independent coding of I-frames and Wyner-Ziv coding of P frames, where the side information is generated by fusing the disparity map with the motion field. The achieved bit rates are still quite far from the Slepian-Wolf bound, mainly due to independent coding of I-frames and this gap can be reduced by encoding more coarsely the I-frames [14]. Finally, the DSC principles can be also exploited for the error-resilient delivery of multi-view video in wireless camera networks [15].

The common characteristics of all state-of-the-art disparity-based DSC frameworks lie on the need of at least two independently encoded views in order to perform disparity estimation for DSC decoding, which leads to high encoding rates. Moreover, the disparity estimation usually requires high-resolution images, which is quite restricting in practical camera network scenarios. This work contributes to solving these two main problems by efficiently relating the correlated data in multiple views under geometric local transforms. This enables the estimation of scene geometry and a correct decoding of Wyner-Ziv frames, even with a single reference frame that has

been highly compressed.

### III. MULTI-VIEW CORRELATION MODEL

Images of a 3D scene taken by distributed cameras are likely correlated as they capture the same objects in the scene from different viewpoints. The correlation between multi-view images arises from the rigid motion of the objects in the scene due to viewpoint change, and can be simply described by local changes of image components that represent the moving objects. In other words, if we decompose each image into components that capture the objects in the scene, we can assume that the most prominent components are present in all images with high probability, possibly with some local transformations. However, image decompositions by standard transforms like the wavelet or DCT do not describe the image semantics. Due to the orthogonality of the basis vectors, extracted image components rarely capture the scene objects and their geometry. On the other side, sparse image approximations with overcomplete dictionaries of basis vectors (atoms) have shown capable of capturing the image structure and geometry using only few basis vectors [16], while offering excellent approximation performance. Sparse approximations have been also successfully applied to video [17], [18] and 3D object compression [3]. One of the most important advantages of sparse approximations is the flexibility in the design of an overcomplete dictionary. When the dictionary is built on geometrical functions with local support, the sparse image decomposition results in a set of meaningful geometrical features that represent the visual information of the scene. The comparison of these features in different views permits to estimate the geometry of the scene and the correlation between views. The correlation between multi-view images is driven by local transformations of sparse image components in different views that represent the same component in the 3D scene. Interestingly, sparse image approximations with redundant dictionaries of geometrical features are believed to mimic the behavior of the human visual system for encoding visual information [19].

We briefly overview the basics of sparse signal approximation that are used to build our correlation model. Given a certain basis, or a redundant dictionary of atoms  $\mathcal{D} = \{\phi_k\}, k = 1, \dots, N$ , in a Hilbert space, every image  $y$  can be represented as:

$$y = \Phi x = \sum_{k=1}^N x_k \phi_k, \quad (1)$$

where the matrix  $\Phi$  is composed of atoms  $\phi_k$  as columns. When the dictionary is over-complete,  $x$  is not unique. In order to find a compact image approximation one has to search for a sparse vector  $x$  that contains a small number of significant coefficients, while the rest of coefficients are close or equal to zero. In other words, we say that  $y$  has a *sparse* representation in  $\mathcal{D}$  if it can be represented as a linear combination of a small number of atoms in  $\mathcal{D}$ , up to an approximation error  $\eta$ , i.e.:

$$y = \Phi_I c + \eta = \sum_{k \in I} x_k \phi_k + \eta, \quad (2)$$

where  $c$  is the vector of significant elements of  $x$ ,  $I$  labels the set of atoms  $\{\phi_k\}_{k \in I}$  participating in the representation, and  $\Phi_I$  is a sub-matrix of  $\Phi$  with respect to  $I$ . One is generally not interested in finding an exact representation, but rather in finding a sparse expansion with a small approximation error. In order to find the sparsest approximation of  $y$  with a bounded error norm  $\|\eta\| \leq \varepsilon$ , the following minimization problem needs to be solved:

$$\min_c \|c\|_0 \quad \text{subject to} \quad \|y - \Phi_I c\|_2 \leq \varepsilon, \quad (3)$$

where  $\|\cdot\|_0$  denotes the  $l_0$  norm. This minimization problem involves searching for the shortest vector of significant coefficient in  $x$ , which has combinatorial complexity and it is NP-complete. However, there exist algorithms that search for a suboptimal solution for a sparse vector  $x$  with a limited complexity. They can be classified in two main groups: greedy algorithms (Matching Pursuit (MP), Orthogonal MP (OMP), Weak OMP, etc.) that iteratively select locally optimal basis vectors; and algorithms based on convex relaxation methods (Basis Pursuit) that solve however a slightly different problem where the  $l_0$  norm in Eq. (3) is replaced by an  $l_1$  norm. For details on these algorithms we refer the reader to [20].

We are now interested in defining the correlation model between sparse approximations of two correlated multi-view images<sup>1</sup>:

$$\begin{aligned} y_1 &= \Phi_{I_1} c_1 + \eta_1 \\ y_2 &= \Phi_{I_2} c_2 + \eta_2. \end{aligned}$$

Since  $y_1$  and  $y_2$  capture the same 3D scene, there exists a subset of atoms indexed respectively by  $J_1 \in I_1$  and  $J_2 \in I_2$  that represent image projections of the same prominent 3D features in the scene. We assume that these atoms are correlated, possibly under some local geometric transformations. Let  $F(\phi)$  denote the transform of an atom between two image decompositions that results from the motion of an object in the 3D space. Equivalently, it represents the transformation imposed to an atom  $\phi$  in a correlated view due to camera displacement. Therefore, the correlation between the images can be modeled as a set of transforms  $F_i$  between corresponding atoms in sets indexed by  $J_1$  and  $J_2$ . The approximation of the image  $y_2$  can be rewritten as the sum of the contributions of transformed atoms, remaining atoms in  $I_2$ , and noise  $\eta_2$ :

$$y_2 = \sum_{i \in J_1} x_{2,i} F_i(\phi_i) + \sum_{k \in I_2 \setminus J_2} x_{2,k} \phi_k + \eta_2. \quad (4)$$

The hereabove model is independent of the sparse approximation algorithm used for image decomposition, and generic with respect to the overcomplete dictionary selection. However, we choose a dictionary built on locally defined geometric atoms that can approximate multidimensional discontinuities like edges. These represent important information about the scene geometry.

The main challenge in the proposed model is to define the transforms  $F_i$  in the Eq. (4) that relate corresponding atoms in sparse decompositions of omnidirectional multi-view images.

<sup>1</sup>We take two images for the sake of clarity, but the framework can be generalized to any number of images.

Due to motions of objects in the 3D space various types of transforms are introduced in the image projective space. Most of these transforms can be represented by the 2-D similarity group elements, which include 2-D translation, rotation and isotropic scaling of the image features. We also consider anisotropic scaling to further expand the space of possible transforms among image features. In order to efficiently capture transforms between sparse image components, we propose to use a structured redundant dictionary of atoms for image representation. Atoms in the structured dictionary are derived from a single waveform that undergoes rotation, translation and scaling. Hence, the transformation of an atom by any of the 2-D similarity group elements or anisotropic scaling, results in another atom in the same dictionary: the dictionary is invariant with respect to any transform action. More formally, given a generating function  $g$  defined in the Hilbert space, the dictionary  $\mathcal{D} = \{\phi_k\} = \{g_\gamma\}_{\gamma \in \Gamma}$  is constructed by changing the atom index  $\gamma \in \Gamma$  that defines rotation, translation and scaling parameters applied to the generating function  $g$ . This is equivalent to applying a unitary operator  $U(\gamma)$  to the generating function  $g$ , i.e.:  $g_\gamma = U(\gamma)g$ . When the dictionary is defined this way, the transform of one atom  $g_{\gamma_i}$  to another atom  $g_{\gamma_j}$  reduces to a transform of its parameters, i.e.,

$$g_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma')g_{\gamma_i} = U(\gamma' \circ \gamma_i)g. \quad (5)$$

This equality holds for any transform-invariant overcomplete dictionary in the Hilbert space. Note that the size and redundancy of the dictionary is directly driven by the number of distinct atom transformations.

#### IV. TRANSFORMS IN OMNIDIRECTIONAL IMAGES

Omnidirectional imaging represents an interesting and increasingly popular framework for 3D scene representation. It offers a wider field of view and therefore necessitates only a small number of camera sensors for capturing a 3D scene. In addition, it permits to process the visual information without the discrepancies introduced by Euclidian assumptions in planar imaging. Therefore, we address the problem of correlation modeling for multi-view omnidirectional images. As these images can be precisely mapped on a sphere, we further use a dictionary of atoms on the 2-D unit sphere. The generating function  $g$  is hence defined in the space of square-integrable functions on a unit two-sphere  $S^2$ ,  $g(\theta, \varphi) \in L^2(S^2)$ , while the dictionary is built by changing the atom indexes  $\gamma = (\tau, \nu, \psi, \alpha, \beta) \in \Gamma$ . The triplet  $(\tau, \nu, \psi)$  represents Euler angles that respectively describe the motion of the atom on the sphere by angles  $\tau$  and  $\nu$ , and the rotation of the atom around its axis with an angle  $\psi$ , and  $\alpha, \beta$  represent anisotropic scaling factors. As an example, Gaussian atoms on the sphere are illustrated on the Figure 2, for different motion, rotation and anisotropic scaling parameters.

We are interested in finding correspondences between atoms that respectively represent the images  $y_1$  and  $y_2$ , generated by two omnidirectional cameras that capture the same scene. For the sake of clarity, let  $\{g_\gamma\}_{\gamma \in \Gamma}$  and  $\{h_\gamma\}_{\gamma \in \Gamma}$  respectively denote the set of functions used for the expansions of images  $y_1$  and  $y_2$ . The same dictionary is used for both images, so

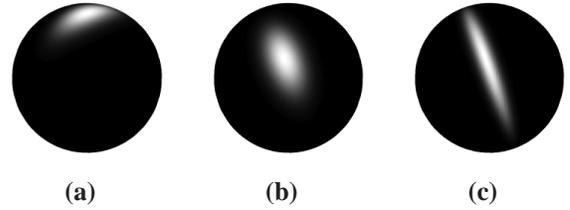


Fig. 2. Gaussian atoms: a) on the North pole ( $\tau = 0, \nu = 0$ ),  $\psi = 0, \alpha = 2, \beta = 4$ ; b)  $\tau = \frac{\pi}{4}, \nu = \frac{\pi}{4}, \psi = \frac{\pi}{8}, \alpha = 2, \beta = 4$ ; c)  $\tau = \frac{\pi}{4}, \nu = \frac{\pi}{4}, \psi = \frac{\pi}{8}, \alpha = 1, \beta = 8$ .

that two corresponding atoms  $g_{\gamma_i}$  and  $h_{\gamma_j}$  in images  $y_1$  and  $y_2$  are linked by a simple transform of the atom parameters, and Eq. (5) can be rewritten as

$$h_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma')g_{\gamma_i} = U(\gamma' \circ \gamma_i)g. \quad (6)$$

The subset of transforms  $V_i^0 = \{\gamma' | h_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma')g_{\gamma_i}\}$  allows to relate  $g_{\gamma_i}$  to the atoms  $h_{\gamma_j}$  in the expansion of  $y_2$ . However, not all these transforms are feasible in multi-view correlated images. The set of possible transforms can be greatly reduced by identifying two constraints between corresponding atoms, namely *shape similarity* constraints and *epipolar* constraint.

First, we assume that the 3D motion of an object results in a limited difference between shapes of corresponding atoms since they represent the same object in the 3D scene. Therefore, we can restrict the set of possible transforms by the shape similarity constraints between candidate atoms. We measure the similarity or coherence of atoms by the inner product  $\mu(i, j) = |\langle g_{\gamma_i}, h_{\gamma_j} \rangle|$ , and we impose a minimal coherence between candidate atoms, i.e.,  $\mu(i, j) > s$ . This defines a set of possible transforms  $V_i^\mu \subseteq V_i^0$  with respect to atom shape, as:

$$V_i^\mu = \{\gamma' | h_{\gamma_j} = U(\gamma')g_{\gamma_i}, \mu(i, j) > s\}. \quad (7)$$

Equivalently, the set of atoms  $h_{\gamma_j}$  in  $y_2$  that are possible transformed versions of the atom  $g_{\gamma_i}$  is denoted as the *shape candidates set*. It is defined by the set of atoms indexes  $\Gamma_i^\mu \subset \Gamma$ , with

$$\Gamma_i^\mu = \{\gamma_j | h_{\gamma_j} = U(\gamma')g_{\gamma_i}, \gamma' \in V_i^\mu\}. \quad (8)$$

Second, pairs of atoms that correspond to the same 3D points have to satisfy epipolar constraints, that represent one of the fundamental relations in multi-view analysis. The epipolar constraint defines the relation between 3D point projections  $(\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^3)$  on two cameras, as:

$$\mathbf{z}_2^T \hat{T} R \mathbf{z}_1 = 0, \quad (9)$$

where  $R$  and  $T$  are the rotation and translation matrices of one camera frame with respect to the other, and  $\hat{T}$  is obtained by representing the cross product of  $T$  with  $R\mathbf{z}_1$  as matrix multiplication, i.e.,  $\hat{T}R\mathbf{z}_1 = T \times R\mathbf{z}_1$ . The set of possible transforms between atoms from different views is therefore further reduced to the transforms that respect epipolar constraints between the atom  $g_{\gamma_i}$  in  $y_1$  and the candidates atoms  $h_{\gamma_j}$  in  $y_2$ . The constraint given in Eq. (9) is rarely exactly satisfied for corresponding pixels or areas

in two multi-view images, and the decision on the epipolar matching of two correspondences is commonly taken when their epipolar distance is smaller than a certain threshold  $\kappa$ .

By imposing the epipolar constraint on atoms in  $V_i^0$ , we define the set  $V_i^E \subseteq V_i^0$  of possible transforms of atom  $g_{\gamma_i}$  as:

$$V_i^E = \{\gamma' | h_{\gamma_j} = U(\gamma')g_{\gamma_i}, d_{EA}(g_{\gamma_i}, h_{\gamma_j}) < \kappa\}, \quad (10)$$

where  $d_{EA}(g_{\gamma_i}, h_{\gamma_j})$  denotes the epipolar distance between atoms  $g_{\gamma_i}$  and  $h_{\gamma_j}$  (see below). Similarly, we define a set of candidate atoms in  $y_2$ , called the *epipolar candidates set*, whose indexes belong to  $\Gamma_i^E \subset \Gamma$ , with:

$$\Gamma_i^E = \{\gamma_j | h_{\gamma_j} = U(\gamma')g_{\gamma_i}, \gamma' \in V_i^E\} \subset \Gamma. \quad (11)$$

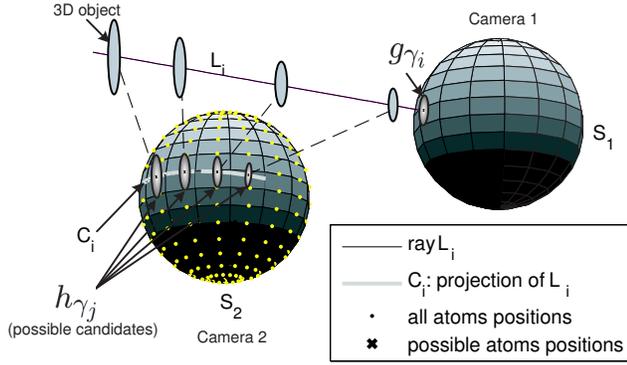


Fig. 3. Selection of positions of atoms that satisfy epipolar constraints.

A graphical interpretation of the epipolar constraint for spherical images is shown on the Figure 3, where we denote as  $S_1$  and  $S_2$  the two unit spheres corresponding to camera projection surfaces. A given atom  $g_{\gamma_i}$  in  $y_1$ , on the sphere  $S_1$ , can be a projection of infinitely many different 3D objects, at different scales and distances from  $S_1$ . We show an example of several different objects whose projection on  $S_1$  is  $g_{\gamma_i}$  and projections on  $S_2$  are  $h_{\gamma_j}$ . Due to epipolar constraints, the atoms  $h_{\gamma_j}$  are positioned on the part of a great circle  $C_i$  obtained by projecting the ray  $L_i$  on the sphere  $S_2$ . This ray originates from the center of camera 1 and passes through the atom  $g_{\gamma_i}$  on the sphere  $S_1$ .

Finally, we combine the epipolar and shape similarity constraints to define the set of possible transforms for atom  $g_{\gamma_i}$ , as  $V_i = V_i^E \cap V_i^\mu$ . Similarly, we denote the set of possible parameters of the transformed atom in  $y_2$  as  $\Gamma_i = \Gamma_i^E \cap \Gamma_i^\mu$ . Given the set  $\Gamma_i$  of possible atom parameters in  $y_2$  corresponding to the atom  $g_{\gamma_i}$  in  $y_1$ , the correspondence  $h_{\gamma_j}$  in  $y_2$  can be defined with high probability under the assumption that the decomposition of  $y_2$  is sparse.

## V. DISPARITY MAP ESTIMATION BY ATOM TRANSFORMS

The local transformations between geometric atoms are now used to estimate the correlation between pixels in multiview images, as represented by a *disparity map*. A disparity map typically permits view interpolation under epipolar constraints. It is defined as the point-wise correlation between multi-view images, which relates a point  $\mathbf{z}_1$  on the image  $y_1$  to a point  $\mathbf{z}_2$

on  $y_2$ , such that the epipolar constraint from Eq. (9) is satisfied. This mapping is most commonly estimated by identifying corresponding feature points like corners in multi-view images and relating their local neighborhoods by a cross-correlation similarity measure [21]. However, cross-correlation measure is not rotationally invariant and it fails to capture rotation of patterns between views. Since our correlation model relates local geometric features by atoms with different scale and rotation parameters in different views, it represents a similarity measure that is invariant with respect to rotation and scaling. Therefore, a pair of corresponding atoms can give a reliable estimate of the disparity map, obtained by the atom transform. We describe here the estimation of the disparity map from the atom transforms, and we define a measure of the estimation error that can be used to refine the atom pairing process.

Let's consider a pair of corresponding atoms  $(g_{\gamma_i}, h_{\gamma_j})$  in two images. We want to find a mapping of each point on  $g_{\gamma_i}$  to its corresponding point on  $h_{\gamma_j}$ . Since this mapping is point-wise, we need to define  $g_{\gamma_i}$  in the discrete space, i.e., on the spherical grid  $\mathcal{G}_1$ . Then, the disparity mapping translates to the grid distortion induced by the local transform between  $g_{\gamma_i}$  and  $h_{\gamma_j}$ , denoted as  $\mathcal{F}\{\mathcal{G}_1\}$ . Let  $P_1$  be a point on  $\mathcal{G}_1$ , given in Euclidean coordinates as  $\mathbf{z}_1$ . Similarly, let  $P_2$  be a point on  $\mathcal{G}_2$ , given in Euclidean coordinates as  $\mathbf{z}_2$ , which is obtained by applying the grid transform  $\mathcal{F}$  to  $P_1$ . Let further  $\gamma_i = (\tau_i, \nu_i, \psi_i, \alpha_i, \beta_i)$  and  $\gamma_j = (\tau_j, \nu_j, \psi_j, \alpha_j, \beta_j)$ . The grid transform  $\mathcal{G}_2 = \mathcal{F}\{\mathcal{G}_1\}$  includes two transforms:

- 1) transform of the motion of atom  $g_{\gamma_i}$ , given by Euler angles  $(\tau_i, \nu_i, \psi_i)$ , into the motion of atom  $h_{\gamma_j}$ , given by Euler angles  $(\tau_j, \nu_j, \psi_j)$
- 2) transform of anisotropic scaling of the atom  $g_{\gamma_i}$ , given by the pair of scales  $(\alpha_i, \beta_i)$ , into the anisotropic scaling of the atom  $h_{\gamma_j}$ , given by the pair of scales  $(\alpha_j, \beta_j)$ .

By combining these two transforms, the point  $\mathbf{z}_2$  can be written as:

$$\mathbf{z}_2 = R_{\gamma_j}^{-1} \cdot u(R_{\gamma_i} \cdot \mathbf{z}_1), \quad (12)$$

where  $R_{\gamma_i}$  and  $R_{\gamma_j}$  are rotation matrices given by Euler angles  $(\tau_i, \nu_i, \psi_i)$  and  $(\tau_j, \nu_j, \psi_j)$ , respectively, and  $u(\cdot)$  defines the grid transform due to anisotropic scaling. Since the anisotropic scaling of atoms on the sphere is performed on the plane tangent to the North pole by projecting the atom with stereographic projection, the grid  $\mathcal{G}_1$  is first rotated such that the North pole is aligned with the center of atom  $g_{\gamma_i}$ , then deformed with respect to anisotropic scaling, and finally rotated back with the rotation matrix of atom  $h_{\gamma_j}$ .

In more details, the stereographic projection [22] at the North pole projects a point  $(\theta, \varphi)$  on the sphere to a point  $(x, y)$  on the plane tangent to the North pole, and it is formally given with:

$$x + jy = \rho e^{j\varphi} = 2 \tan\left(\frac{\theta}{2}\right) e^{j\varphi}. \quad (13)$$

Let now  $(\theta_1, \varphi_1)$  and  $(\theta_2, \varphi_2)$  denote the spherical coordinates of points  $P_1$  and  $P_2$  respectively (the point belongs to the unit sphere and  $r = 1$  is assumed). Under the stereographic projection, the transform of the point  $(\theta_1, \varphi_1)$  on the grid  $\mathcal{G}_1$  to the point  $(\theta_2, \varphi_2)$  on the grid  $\mathcal{G}_2$  due to anisotropic scaling can

be obtained by scaling the stereographic projection of  $(\theta_1, \varphi_1)$  with  $1/\alpha_j$  and  $1/\beta_j$ , in the following way:

$$\begin{aligned} x_2 &= \rho_2 \cos \varphi_2 = \frac{1}{\alpha_j} \alpha_i x_1 = \frac{\alpha_i}{\alpha_j} \rho_1 \cos \varphi_1 \\ y_2 &= \rho_2 \sin \varphi_2 = \frac{1}{\beta_j} \beta_i y_1 = \frac{\beta_i}{\beta_j} \rho_1 \sin \varphi_1, \end{aligned}$$

where  $\rho_2 = 2 \tan \theta_2 / 2$  and  $\rho_1 = 2 \tan \theta_1 / 2$ . By solving the system of Eq. (14) for  $\theta_2$  and  $\varphi_2$ , we get:

$$\varphi_2 = u_p(\varphi_1) = \arctan \left( \frac{\alpha_j \beta_i \sin \varphi_1}{\alpha_i \beta_j \cos \varphi_1} \right) \quad (14)$$

$$\begin{aligned} \theta_2 &= u_t(\theta_1, \varphi_1, \varphi_2) \\ &= 2 \arctan \left[ \tan \frac{\theta_1}{2} \sqrt{\frac{\alpha_i^2 \cos^2 \varphi_1 + \beta_i^2 \sin^2 \varphi_1}{\alpha_j^2 \cos^2 \varphi_2 + \beta_j^2 \sin^2 \varphi_2}} \right] \quad (15) \end{aligned}$$

We can therefore define the function  $u(\cdot)$  as a pair of transforms  $u_p(\varphi_1)$  and  $u_t(\theta_1, \varphi_1, u_p(\varphi_1))$  followed by the transform of spherical coordinates  $(\theta_2, \varphi_2)$  to Euclidean coordinates  $\mathbf{z}_2$ . The relation given in Eq. (12) is now completely defined, based on the parameters of corresponding atoms in two images. When the transformation is applied to all points, it forms the disparity map between the correlated views.

Finally, we define the *Symmetric epipolar atom distance* in order to quantify the mismatch between two corresponding atoms  $g_{\gamma_i}$  and  $h_{\gamma_j}$  related by the disparity map. The symmetric epipolar atom distance actually measures how much the atom pair  $g_{\gamma_i}$  and  $h_{\gamma_j}$  deviates from the perfect epipolar matching given in the correlation model of Eq. (10), when  $d_{EA}(g_{\gamma_i}, h_{\gamma_j}) = 0$ . It is evaluated as the weighted average of the symmetric epipolar distance of all pairs of points given by the disparity map:

$$d_{EA}(g_{\gamma_i}, h_{\gamma_j}) = \sum_{\mathbf{z}_1 \in \mathcal{G}_1} w_{\gamma_i}(\mathbf{z}_1) d_{SE}(\mathbf{z}_1, \mathbf{z}_2). \quad (16)$$

The points  $\mathbf{z}_1$  and  $\mathbf{z}_2$  are related by the disparity map and  $d_{SE}(\mathbf{z}_1, \mathbf{z}_2)$  stands for the symmetric epipolar distance between  $\mathbf{z}_1$  and  $\mathbf{z}_2$  [21]. It is defined as:

$$d_{SE}(\mathbf{z}_1, \mathbf{z}_2) = \sqrt{d(\mathbf{z}_1, \mathcal{C}_{\mathbf{z}_2}) + d(\mathbf{z}_2, \mathcal{C}_{\mathbf{z}_1})} \quad (17)$$

where  $d(\mathbf{z}_1, \mathcal{C}_{\mathbf{z}_2})$  denotes the Euclidean distance of the point  $\mathbf{z}_1$  to the epipolar circle  $\mathcal{C}_{\mathbf{z}_2}$  corresponding to point  $\mathbf{z}_2$ . The weight  $w_{\gamma_i}$  is a normalized weight function that prioritizes the points where the atom  $g_{\gamma_i}$  has higher response. The goal of this function is to give more importance to the disparity mismatch of points that lie closer to the geometrical component captured by the atom (typically edges). One example could be a 2-dimensional Gaussian weight function, anisotropically scaled and oriented, which fits the atom  $g_{\gamma_i}$ . If the overcomplete dictionary is composed of Gaussian atoms, the weight function is equal to the atom itself. We use 2D Gaussian weight function in the rest of this paper.

## VI. DISTRIBUTED SCENE CODING

### A. Encoder and coset design

The correlation model introduced before can be exploited for the design of a distributed algorithm, as it explicitly

relates atom parameters with scene geometry constraints in the compressed domain. We propose here a scheme for coding with side information, as a special case of DSC, where image  $y_1$  is independently encoded at a rate  $R_{y_1} \geq H(y_1)$ , and the image  $y_2$  is encoded with coset coding at the rate  $R_{y_2} \geq H(y_2|y_1)$ . The sparse decomposition of the reference image  $y_1$  is independently encoded, while the decomposition of the Wyner-Ziv image  $y_2$  is encoded by coset coding of atom indexes and entropy coding of their respective coefficients, as shown on the Figure 4. We propose to partition the set of atom indexes  $\Gamma$  into distinct cosets that contain dissimilar atoms with respect to their position and shape. Under the assumption that an atom  $h_{\gamma_j}$  in the image decomposition has its corresponding atom  $g_{\gamma_i}$  in the side information expansion, the Wyner-Ziv encoder does not need to code the entire  $\gamma_j$ . It rather transmits only the information that is necessary to identify the correct atom in the transform candidate set given by  $\Gamma_i = \Gamma_i^E \cap \Gamma_i^\mu$ , as given by Eq. (8) and (11). The side information and the coset index are therefore sufficient to recover the atom  $g_{\gamma_j}$  in the Wyner-Ziv image. The achievable bit rate for encoding the atom index  $\gamma_j$  is reduced therefore from  $R_{y_2} \geq H(\gamma_j|\gamma_j \in \Gamma)$  to  $R_{y_2} \geq H(\gamma_j|\gamma_j \in \Gamma_i)$ .

Due to the independency of epipolar and shape constraints, the cosets can be designed independently for atom shape parameters  $(\psi, \alpha, \beta)$ , and for atom positions  $(\tau, \nu)$  according to epipolar constraints. We therefore construct two types of cosets, respectively the Shape cosets:  $K_l^\mu, l = 1, \dots, N_2$  and the Position cosets  $K_k^E, k = 1, \dots, N_1$ . We design Shape cosets by distributing all atoms whose parameters belong to  $\Gamma_i^\mu$  into different cosets. The encoder eventually sends for each atom only the indexes of the corresponding cosets (i.e.,  $k_n$  and  $l_n$  in Figure 4).

Next, we propose two design methods for constructing the Position cosets, that correspond to scenario where the camera pose  $(R, T)$  is known, or not available respectively. We first design *Epipolar cosets* based on the fact that the centers of two corresponding atoms  $g_{\gamma_i}$  and  $h_{\gamma_j}$ , denoted as  $m_i$  and  $m_j$  respectively, satisfy the epipolar constraint, i.e.,  $m_j^T \hat{T} R m_i = 0$ . This condition is a special case of the general epipolar constraint given in the Eq. (10) when  $\mathcal{G}_1 = m_i$ , which transforms into  $\mathcal{G}_2 = m_j$ . The epipolar candidates set given in (11) reduces to:

$$\tilde{\Gamma}_i^E = \{\gamma_j | h_{\gamma_j} = U(\gamma') g_{\gamma_i}, d_{SE}(m_i, m_j) \leq \delta\}, \quad (18)$$

where  $\delta$  represents a small threshold value on the symmetric epipolar distance. The main design idea is to separate into different cosets the atoms that belong to the same set  $\Gamma_i^E$  for  $\mathcal{G}_1 = m_i$ . The parameter  $\delta$  can be used in the coset design for selecting the number of cosets and for adapting the encoding rate. Given the side information atom  $g_{\gamma_i}$ , the decoder only needs to know the coset index of  $h_{\gamma_j}$  for joint decoding.

As an alternative, we propose to design Position cosets based on *Vector Quantization* of positions in the absence of information about the relative camera poses. The VQ cosets are constructed by 2-dimensional interleaved uniform quantization of atom positions  $(\tau, \nu)$  on a rectangular lattice. This coset design can be formulated analogously to the Epipolar coset design, where the set of position candidates (called the

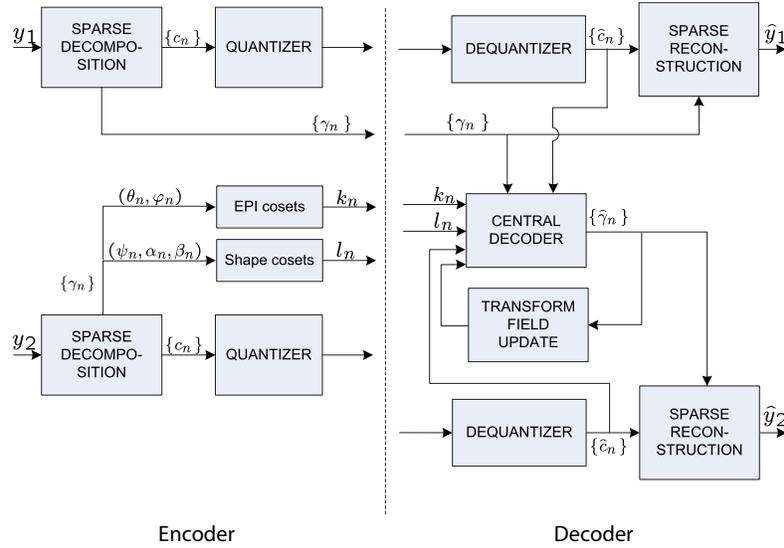


Fig. 4. Block diagram for the Wyner-Ziv codec.

set of epipolar candidates in Eq. (18) gathers the candidates positions  $(\tau_j, \nu_j)$  within the neighborhood of the reference atom position  $(\tau_i, \nu_i)$ , i.e.:

$$\tilde{\Gamma}_i^V = \{\gamma_j | h_{\gamma_j} = U(\gamma')g_{\gamma_i}, |\tau_i - \tau_j| < \Delta\tau, |\nu_i - \nu_j| < \Delta\nu\}. \quad (19)$$

The interleaved vector quantization of  $\tau$  and  $\nu$  will distribute the pairs  $(\tau, \nu)$  that belong to the same  $\tilde{\Gamma}_i^V$  into different cosets, while keeping the distance between coset elements constant and equal to  $(\Delta\tau, \Delta\nu)$ . Note that the constant intra-coset distance can not be however guaranteed in the case of EPI cosets. Both coset design methods are used in the experiments, and their selection depends on the constraints of the camera network application.

### B. Decoder and image reconstruction

The central decoder (CD) builds on the correlation model based on local atom transformations, in order to establish correspondences between atoms in the reference image and atoms within the cosets of the Wyner-Ziv image decomposition (see Figure 4). It also uses the information provided by the quantized coefficients of atoms, in order to improve the atom matching process. In other words, for decoding of the  $n^{\text{th}}$  atom in the Wyner-Ziv frame, the decoder has the following information: the index of the Position coset  $k_n$ , the index of the Shape coset  $l_n$ , and the coefficient  $\hat{c}_n$  after inverse quantization. The goal of the decoder is to select the atom position  $(\tau_n, \nu_n)$  from  $K_{k_n}^E$  and the atom shape  $(\psi_n, \alpha_n, \beta_n)$  from  $K_{l_n}^\mu$ . Let  $A_n$  denote the set of possible candidates for decoding the  $n^{\text{th}}$  atom in  $y_2$ , with  $|A_n| = |K_{k_n}^E| \cdot |K_{l_n}^\mu|$  when  $|\cdot|$  denotes the cardinality of a set. However, only a small subset of atoms in  $A_n$  have corresponding atoms in the reference image  $y_1$ . The decoder has therefore to identify the possible pairs of corresponding atoms between  $A_n$  and  $I_1$ .

Since the atoms coefficients of the Wyner-Ziv image  $\hat{c}_n$  are known at the decoder, the decoder selects a subset of atoms in  $I_1$  whose coefficient values are close to  $\hat{c}_n$ . The relation

between coefficients can be established when the coefficients are obtained as projections of the image to the corresponding atom, i.e. when  $c_n = \langle y_2, h_{\gamma_n} \rangle$ . Under the assumption that the image approximations are sparse enough the projections of two corresponding atoms  $g_{\gamma_i}$  and  $h_{\gamma_j}$  are related as:

$$\frac{\langle y_1, g_{\gamma_i} \rangle}{n_i} = \frac{\langle y_2, h_{\gamma_j} \rangle}{n_j}, \quad (20)$$

where  $n_i$  and  $n_j$  denote the norms of atoms  $g_{\gamma_i}$  and  $h_{\gamma_j}$  prior to atom normalization. Therefore, the decoder can select a subset of atoms  $J_n = \{\gamma_i\}$  in  $I_1$  whose coefficients satisfy:

$$\Delta c = \left| \frac{\hat{c}_i - \hat{c}_n}{\hat{c}_n} \right| \approx \left| \frac{\langle y_1, g_{\gamma_i} \rangle - \langle y_2, h_{\gamma_n} \rangle}{\langle y_2, h_{\gamma_n} \rangle} \right| < \sigma, \quad (21)$$

where  $\sigma$  is a chosen threshold. For each  $g_{\gamma_i} \in J_n$  we have a set of possible transformed atoms given by  $\tilde{\Gamma}_i^E = \tilde{\Gamma}_i^E \cap \Gamma_i^\mu$  or  $\tilde{\Gamma}_i = \tilde{\Gamma}_i^V \cap \Gamma_i^\mu$  respectively for epipolar or VQ cosets. The decoder further looks if any of the candidates in  $A_n$  belongs to  $\bigcup_{\gamma_i \in J_n} \tilde{\Gamma}_i$ . Note that, in the general case, the parameters  $\delta, \Delta\tau, \Delta\nu, s$  that define the correlation sets  $\tilde{\Gamma}_i^E, \tilde{\Gamma}_i^V$  and  $\Gamma_i^\mu$  can have different values for the coset design and for the decoding. This permits to put more strict conditions for the selection of corresponding atom pairs.

The search for atom correspondences then proceeds in two major steps. First, the decoder eliminates the candidates that do not belong to  $\bigcup_{\gamma_i \in J_n} \tilde{\Gamma}_i$ , as well as candidates with a large symmetric epipolar atom distance, i.e., for which  $d_{EA}(g_{\gamma_i}, h_{\gamma_j}) > \kappa$ . If all candidates in  $A_n$  get eliminated, the decoder decides that the  $n^{\text{th}}$  atom in  $y_2$  does not have a corresponding atom in  $y_1$ . Second, the decoder selects as a correspondence the pair of atoms with the smallest symmetric epipolar atom distance  $d_{EA}(g_{\gamma_i}, h_{\gamma_j})$  among the candidates that have not been eliminated in the first step.

Once a correspondence is identified, the decoder updates the transform field that represents the estimates of the disparity maps for each pixel in the Wyner-Ziv image, with respect to the reference image. The transform field is updated by

combining the disparity map induced by the last pair of atoms with the disparity maps from correspondences that have been defined previously. The transform field represents the fusion of disparity maps from multiple correspondences, which is performed by selecting the most confident mapping for each point  $\mathbf{z}_2$  from different mappings  $\mathbf{z}_1^{(i)}, i = 1, \dots, n$ , defined by  $n$  correspondences. The final mapping point  $\mathbf{z}_1^*$  is selected as:

$$\mathbf{z}_1^* = \arg \max_{\mathbf{z}_1^{(i)}, i=1, n} w_{\gamma_i}(\mathbf{z}_1^{(i)}), \quad (22)$$

where we have used the same weight function as for the symmetric epipolar atom distance.

The transformation of the reference image with respect to the transform field provides an approximation of the Wyner-Ziv image that is used as a side information for decoding the following atoms in the Wyner-Ziv image expansion. The atoms that do not have any correspondence in the reference frame are simply decoded based on the maximal projection on the residual image. The residual image is evaluated as a difference between the side information and previously decoded atoms.

Finally, the reconstruction of the Wyner-Ziv image  $\hat{y}_2$  is obtained as a linear combination of the decoded image  $y_d$ , formed of recovered atoms from  $\Phi_{\mathbf{I}_2}$ , and the transformed reference image  $y_{tr}$ , i.e.,:

$$\hat{y}_2 = y_d + \lambda \Psi_d y_{tr}. \quad (23)$$

The matrix  $\Psi_d$  denotes the orthogonal complement to the basis formed by the decoded atoms in  $\Phi_{\mathbf{I}_2}$ , and  $\lambda$  is an optimization parameter. The reconstructed Wyner-Ziv image benefits from both the decoded information and the transformed features that are not present in the decoded data. We estimate the value of  $\lambda$  from the energy conservation principle. Namely, under the assumption that  $\|\Psi_d y_{tr}\| \approx \|\Psi_d y_2\|$ , we get  $\lambda$  from Eq. (23) as  $\lambda \approx \sqrt{1 - \|y_d\|^2 / \|y_2\|^2}$ , where the energy of the original image  $\|y_2\|^2$  is sent to the decoder as side information.

## VII. EXPERIMENTAL RESULTS

We analyze here the performance of the above Wyner-Ziv coding method for two sets of multi-view images: synthetic spherical images of the Room scene (Fig. 5) and natural omnidirectional images of the Lab scene (Fig. 6). Each set of images includes two  $128 \times 128$  spherical images  $y_1$  and  $y_2$  captured from different viewpoints. The natural omnidirectional images are mapped to spherical images as explained in [23]. For the Room scene the relative pose of one camera with respect to the other is given with  $R = I$  and  $T = [0 \ 0.3 \ 0]^T$ . For the Lab scene the camera pose has been estimated using an algorithm based on sparse approximations (see [24]), and it is given by  $R \approx I$  and  $T = [0.8677 \ 0.0957 \ 0.4878]^T$ .

Sparse image expansions have been constructed using a Matching Pursuit (MP) algorithm implemented on the sphere. The dictionary is based on two generating functions in order to capture both low-frequency components and edge-like features in the scene. The first one consists in a 2D Gaussian function, given as :

$$g_{LF}(\theta, \varphi) = \exp \left( -\tan^2 \frac{\theta}{2} (\alpha^2 \cos^2 \varphi + \beta^2 \sin^2 \varphi) \right), \quad (24)$$

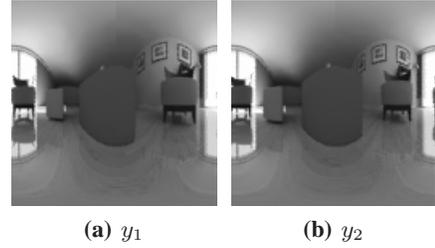


Fig. 5. Original Room images (128x128).

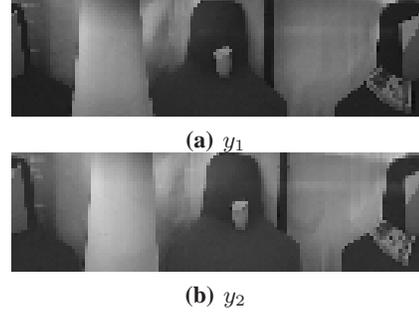


Fig. 6. Original Lab images. The natural omnidirectional images partially cover the sphere due to the boundaries of the mirror in an omnidirectional camera. Here we display a cropped image from the  $128 \times 128$  spherical image, which corresponds to the captured scene.

The second function represents a Gaussian in one direction and the second derivative of a 2D Gaussian in the orthogonal direction (i.e., edge-like atoms similar to the ones presented in [3]). It is written as

$$g_{HF}(\theta, \varphi) = \frac{-1}{K} (16\alpha^2 \tan^2 \frac{\theta}{2} \cos^2 \varphi - 2) \cdot \exp \left( -4 \tan^2 \frac{\theta}{2} (\alpha^2 \cos^2 \varphi + \beta^2 \sin^2 \varphi) \right), \quad (25)$$

where  $K$  is a normalization factor. The position parameters  $\tau$  and  $\nu$  can take 128 different values ( $N_t = N_p = 128$ ), while the rotation parameter uses 16 orientations, between 0 and  $\pi$ . The scales are distributed in a logarithmic scale from 1 to  $N_t/8$  for the Gaussian atoms and from 2 to  $N_p/2$  for edge-like atoms, with 3 scales per octave. The choice of the dictionary is mainly driven by its good approximation properties demonstrated in [3].

The image  $y_1$  is encoded independently, with 100 MP atoms, where the coefficients are quantized by taking benefit of the energy decay properties of Matching Pursuit expansions [25]. The decoded reference images for the Room and Lab scene are shown on the Figures 7(a) and 8(a) respectively. The atom parameters for the expansion of image  $y_2$  are coded with the proposed Wyner-Ziv scheme. The EPI cosets for position coding use a correlation parameter  $\delta = \pi/5$  which gives 1024 Position cosets. Alternatively, Position cosets have also been implemented using VQ in order to generate the same number of cosets. Note that when the center of an atom is close to the epipoles (i.e., degenerative case of epipolar constraints) its parameters have to be encoded independently in the scheme based on EPI cosets. It leads to an overhead in the coding rate for the case of EPI cosets compared to VQ cosets. For the shape cosets, the correlation parameter has been

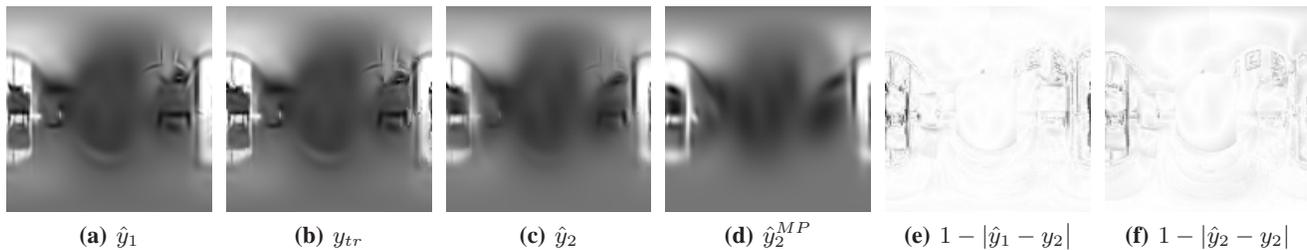


Fig. 7. DSC results for the Room images: (a) decoded reference image  $\hat{y}_1$  (PSNR=30.95dB); (b) transformed reference image  $y_{tr}$ ; (c) decoded Wyner-Ziv image  $\hat{y}_2$  at 0.0534bpp; (d) decoded second image  $\hat{y}_2^{MP}$  when encoded with MP at 0.0534bpp; (e) inverted residue  $1 - |e_1| = 1 - |\hat{y}_1 - y_2|$  without transform compensation (white pixel denotes no error); (f) inverted residue  $1 - |e_2| = 1 - |\hat{y}_2 - y_2|$  after DSC decoding.

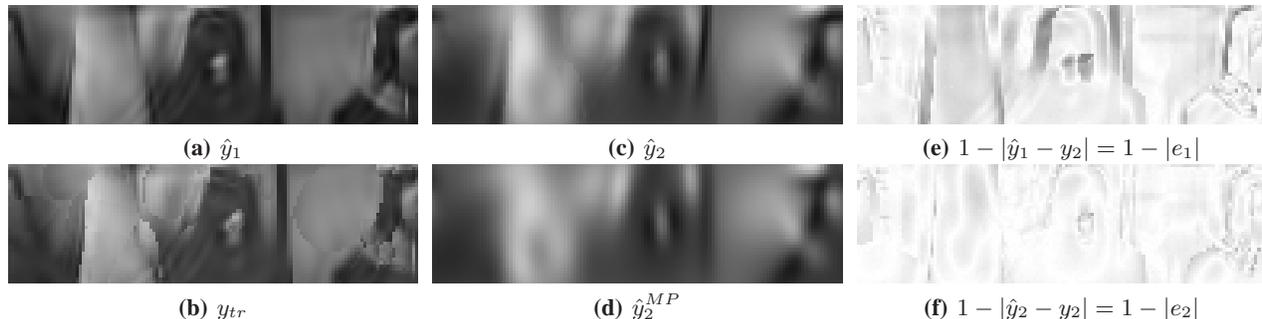


Fig. 8. DSC results for the Lab images: (a) decoded reference image  $\hat{y}_1$  (PSNR=31.82dB); (b) transformed reference image  $y_{tr}$ ; (c) decoded Wyner-Ziv image  $\hat{y}_2$  at 0.038 bpp; (d) decoded second image  $\hat{y}_2^{MP}$  when encoded with MP at 0.039 bpp; (e) inverted residue  $1 - |e_1| = 1 - |\hat{y}_1 - y_2|$  without transform compensation (white pixel denotes no error); (f) inverted residue  $1 - |e_2| = 1 - |\hat{y}_2 - y_2|$  after DSC decoding.

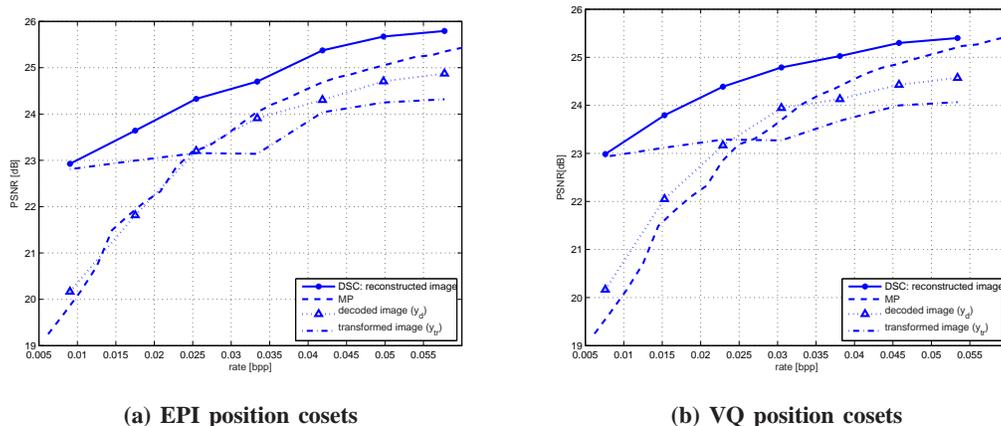


Fig. 9. Rate-distortion performance for the Room image set.

set to  $s_G = 0.85$  (for Gaussian atoms) and  $s_A = 0.51$  (for anisotropic atoms), such that the atoms in the same coset are sufficiently different. These values lead to 128 shape cosets. Finally, the coefficients of the Wyner-Ziv image are obtained by projecting the image  $y_2$  on the atoms selected by MP in order to improve the atom matching process. They are quantized uniformly.

The rate-distortion (RD) performance of the proposed scheme for the Wyner-Ziv image is shown in Figures 9(a) and 9(b) for the Room scene (for EPI and VQ cosets respectively), and in Figure 10 for the Lab scene. The dashed line represents the RD curve of independent coding with Matching Pursuit, while the solid line represents the proposed distributed coding scheme, given by the RD curve of the reconstructed image

$\hat{y}_2$ . The proposed scheme clearly outperforms the independent decoding strategy, especially at low rates. The dash-dotted line represents the RD curve of the side information image, obtained by the application of the transform field on the reference image, showing that the transform field significantly improves the side information. Moreover, it can be noted that the combination of  $y_d$  (dotted line with triangles) and  $y_{tr}$  results in a better overall PSNR of the  $\hat{y}_2$ . The images  $y_{tr}$  and  $\hat{y}_2$  are presented in Figures 7(b)-(c) and 8(b)-(c) for Room and Lab scene respectively. They correspond to the case of coding with VQ cosets at the rate of 0.053 bpp and 0.039 bpp. We can clearly see how the transform field deforms the reference image in order to compensate for different object transforms. Figures 7(d) and 8(d) illustrate the Wyner-Ziv image encoded

independently with MP at the same rate as  $\hat{y}_2$ , resulting in a lower quality than the DSC coded image  $\hat{y}_2$ .

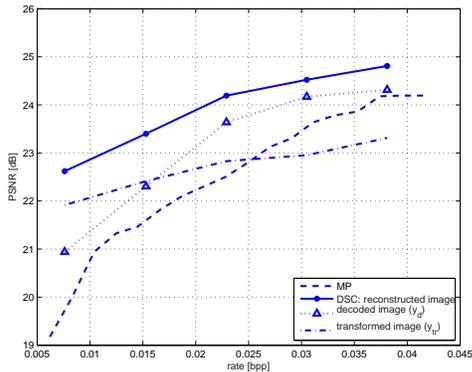


Fig. 10. Rate-distortion performance for the Room image set (VQ Position cosets).

Figure 11 compares the RD performance of our DSC scheme using EPI and VQ cosets. When the decoder finds the same number of correspondences (the curves at lower rate), EPI cosets give worse performance due to the rate overhead for independently coded atoms. However, since EPI cosets offer better matching of atoms, the decoder is able to find more correspondences and the coding with EPI cosets outperforms coding with VQ cosets.

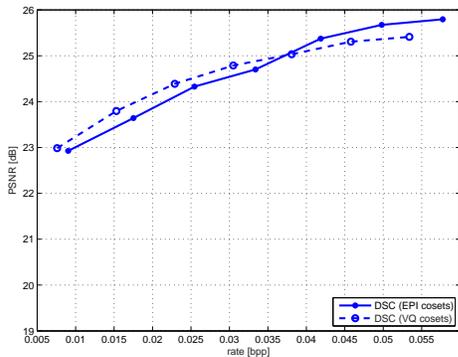


Fig. 11. Rate-distortion coding performance with different position coset design methods (Room image).

Figure 12 compares the proposed DSC method with joint encoding, where the joint encoder finds the atom correspondences and encodes only the parameter differences for the Wyner-Ziv image, while the atoms without correspondences were encoded independently. The reference image is encoded independently at the same rate as in the DSC scheme, where the coefficients are quantized in the same manner. This joint encoding strategy is analogous to our DSC scheme, with the difference that the encoder has access to the side-information. For the sake of fair comparison, the reconstructed image with joint encoding  $\hat{y}_2^J$  is also obtained as a combination of the transformed image  $y_{tr}^J$  and the decoded image  $y_d^J$ , giving a better overall performance. The new DSC scheme performs very

close to the joint encoding at lower rates, where the number of correspondences between views is higher due to the greediness of MP. However, when the number of correspondences drops, the RD performance of DSC saturates. Therefore, the proposed method should be seen as scene geometry estimation and prediction technique that could constitute a first predictive step in a hybrid DSC coding scheme, similar to motion estimation in the hybrid video coding methods. Our correlation model is certainly more advantageous than the block-based motion model since it is able to compensate rotation and scale transforms in addition to translations captured by motion estimation.

Finally, we analyze the efficiency of the geometry-based correlation model. We analyze the residue after DSC coding, denoted with  $e_2 = \hat{y}_2 - y_2$ , and compare it with the difference between the reference image and the original Wyner-Ziv image  $e_1 = \hat{y}_1 - y_1$  (residue without transform compensation). Figures 7(c) and (e) and 8(c) and (e) show the inverted residues  $1 - |e_1|$  and  $1 - |e_2|$  for Room and Lab scene respectively, such that the white pixels correspond to no error. The energy of the error  $e_1$  is respectively 82.65 and 82.85 for the Room and Lab image sets, where the energy is given by the norm of the inner product computed on the sphere. The energy of the error  $e_2$  is respectively 47.12 and 15.65 respectively, which confirms the efficiency of the model based on local geometrical transformations. Unlike  $1 - |e_1|$  where displacements of objects result in high error areas (dark parts), the residue after DSC decoding ( $e_2$ ) is almost exclusively composed of high frequencies since the object transforms have been captured efficiently. The distribution of the residue after transform compensation and decoding can be modeled with the Laplace distribution (see Figure 13). It greatly facilitates the correlation modeling towards the potential DSC encoding of the residual texture information in hybrid coding approaches.

## VIII. CONCLUSIONS

We have presented a geometry-based framework for the efficient representation of 3D scenes, where camera images are approximated by a sparse expansion of prominent geometrical features. A novel correlation model has been proposed based on local geometrical transformations that permit to pair atoms in different images under shape and epipolar geometry constraints. It provides an implicit estimation of the scene geometry that permits to design distributed processing algorithms in camera networks. We have built on this novel framework and designed a distributed coding scheme with side information that offers an efficient rate-distortion representation of 3D scenes. It can lead to effective solutions for distributed sensing and processing of 3D scenes, or high resolution distributed coding when combined with hybrid methods for the representation of texture or unstructured information.

## REFERENCES

- [1] D. Slepian and J. K. Wolf, "Noiseless coding of correlated information sources," *IEEE Transactions on Information Theory*, vol. 19, no. 4, pp. 471–480, July 1973.

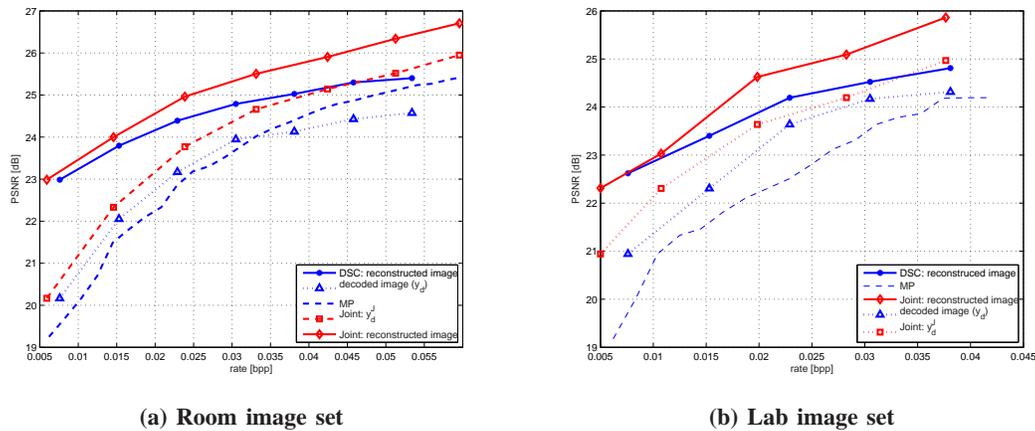


Fig. 12. Comparison of rate-distortion performance for distributed coding and joint encoding (VQ coset design).

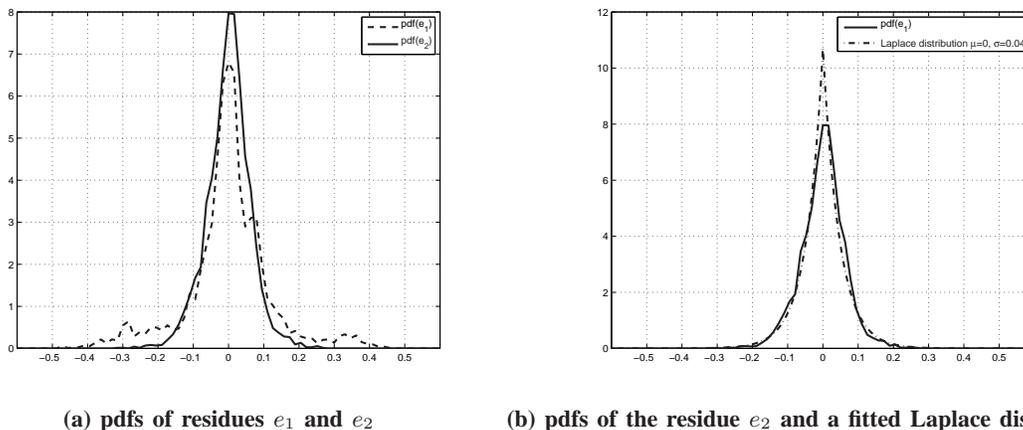


Fig. 13. (a) Comparison of probability density functions of residues  $e_1$  and  $e_2$  for the Lab scene (b) Laplacian distribution fitted to the pdf of the residue  $e_2$  (fitting is performed with the Matlab statistics toolbox).

- [2] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side-information at the decoder," *IEEE Transactions on Information Theory*, vol. 22, no. 1, pp. 1–10, January 1976.
- [3] I. Tasic, P. Frossard and P. Vandergheynst, "Progressive Coding of 3-D Objects Based on Overcomplete Decompositions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 11, pp. 1338–1349, November 2006.
- [4] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS)," *IEEE Transactions on Information Theory*, vol. 49, no. 3, pp. 626–643, March 2003.
- [5] B. Girod, A. Aaron, S. Rane and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71 – 83, January 2005.
- [6] R. Puri and K. Ramchandran, "PRISM: A "reversed" multimedia coding paradigm," in *Proceedings of IEEE ICIP*, 2003.
- [7] A. Sehgal, A. Jagmohan and N. Ahuja, "Wyner-Ziv coding of video: An error-resilient compression framework," *IEEE Transactions on Multimedia*, vol. 6, no. 2, pp. 249–258, April 2004.
- [8] X. Zhu, A. Aaron and B. Girod, "Distributed compression for large camera arrays," in *Proceedings of IEEE SSP*, St Louis, Missouri, September 2003.
- [9] R. Wagner, R. Nowak and R. Baraniuk, "Distributed image compression for sensor networks using correspondence analysis and super-resolution," in *Proceedings of IEEE ICIP*, vol. 1, September 2003, pp. 597–600.
- [10] N. Gehrig, P. L. Dragotti, "DIFFERENT - distributed and fully flexible image encoders for camera sensor networks," in *Proceedings of IEEE ICIP*, vol. 1, September 2005, pp. 690–693.
- [11] M. Flierl and P. Vandergheynst, "Distributed Coding of Highly Correlated Image Sequences with Motion-Compensated Temporal Wavelets," *EURASIP Journal on Applied Signal Processing*, vol. 2006, no. Article ID 46747, p. 10 pages, 2006.
- [12] 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*, vol. 2, no. 3, pp. 9–16, June 2007.
- [13] Y. Yang, V. Stankovic, W. Zhao and Z. Xiong, "Multiterminal video coding," in *Proceedings of IEEE UCSD Workshop on Information Theory and its Applications*, January 2007.
- [14] —, "Multiterminal video coding," in *Proceedings of IEEE ICIP*, September 2007.
- [15] C. Yeo and K. Ramchandran, "Distributed video compression for wireless camera networks," in *Proceedings of SPIE VCIP*, vol. 6508, January 2007, p. 9 pages.
- [16] R.M. Figueras i Ventura, P. Vandergheynst and P. Frossard, "Low rate and flexible image coding with redundant representations," *IEEE Transactions on Image Processing*, vol. 15, no. 3, pp. 726–739, March 2006.
- [17] R. Neff and A. Zakhor, "Very Low Bit-Rate Video Coding based on Matching Pursuits," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 1, pp. 158–171, February 1997.
- [18] A. Rahmoune, P. Vandergheynst and P. Frossard, "Flexible Motion-Adaptive Video Coding with Redundant Expansions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 2, pp. 178–190, February 2006.
- [19] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research*, vol. 37, no. 23, pp. 3311–25, December 1997.
- [20] J. Tropp, "Greed is good: Algorithmic results for sparse approximation."

- IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.
- [21] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
  - [22] J.-P. Antoine and P. Vandergheynst, “Wavelets on the 2-sphere : a group theoretical approach,” *Applied and Computational Harmonic Analysis*, vol. 7, no. 3, pp. 1–30, November 1999.
  - [23] I. Tomic, I. Bogdanova, P. Frossard and P. Vandergheynst, “Multiresolution Motion Estimation for Omnidirectional Images,” in *Proceedings of EUSIPCO*, September 2005.
  - [24] I. Tomic and P. Frossard, “Coarse scene geometry estimation from sparse approximations of multi-view omnidirectional images,” in *Proceedings of EUSIPCO*, September 2007.
  - [25] P. Frossard, P. Vandergheynst, R.M. Figueras i Ventura and M. Kunt, “A posteriori quantization of progressive matching pursuit streams,” *IEEE Transactions on Signal Processing*, vol. 52, no. 2, pp. 525–535, February 2004.