Geometry-Based Distributed Scene Representation With Omnidirectional Vision Sensors

Ivana Tosic, *Student Member, IEEE*, and Pascal Frossard, *Senior Member, IEEE*

***Abstract*—This paper addresses the problem of efficient representation of scenes captured by distributed omnidirectional vision sensors. We propose a novel geometric model to describe the correlation between different views of a 3-D scene. We first approximate the camera images by sparse expansions over a dictionary of geometric atoms. Since the most important visual features are likely to be equivalently dominant in images from multiple cameras, we model the correlation between corresponding features in different views by local geometric transforms. For the particular case of omnidirectional images, we define the multiview transforms between corresponding features based on shape and epipolar geometry constraints. We apply this geometric framework in the design of 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 in order to identify the most likely atoms to decode under epipolar geometry constraints. Experiments demonstrate that the proposed method leads to reliable estimation of the geometric transforms between views. In particular, 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.**

***Index Terms*—Distributed source coding, multiview geometry, omnidirectional vision, sparse approximations, 3-D scene representation.**

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 3-D 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 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.

Manuscript received August 20, 2007; revised March 7, 2008. This work was supported in part by the Swiss National Science Foundation under Grant 20001-107970/1. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Antonio Ortega.

The authors are with the Ecole Polytechnique Fédérale de Lausanne (EPFL), Signal Processing Laboratory (LTS4), Lausanne, 1015-Switzerland (e-mail: ivana.tosic@epfl.ch; pascal.frossard@epfl.ch).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.924288

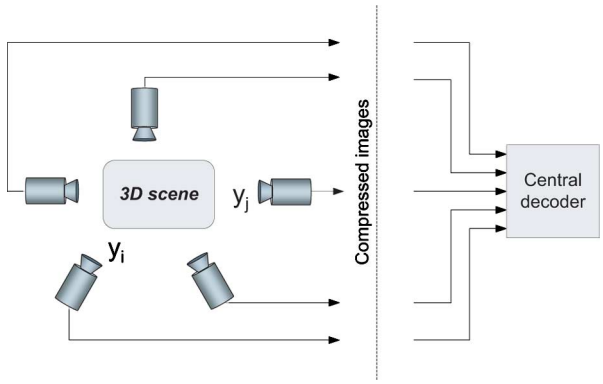


Fig. 1. Distributed coding of 3-D scenes. Multiple correlated views $\{y_i\}$ of the scene are encoded independently and decoded jointly by the central decoder.

In this paper, we consider a framework where a central decoder reconstructs the static 3-D scene information based on multiples images encoded by distributed cameras (see Fig. 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, for example, 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 [3], [4]. In the case of the representation of static scenes, images from different cameras are correlated by geometric constraints defined by 3-D objects in the scene. These arise from the viewpoint change that makes the image projections of the 3-D objects in different views correlated by local transforms such as translation, scaling, or rotation. Hence, the block-translational correlation model is not sufficient to cope efficiently with all types of local transforms that exist in multiview images.

We propose a novel geometry-based correlation model for representation of scenes with distributed cameras, and we apply it to the design of a distributed coding algorithm in camera networks. The main features of a 3-D scene are likely to be dominant in the multiple correlated views of the scene, possibly under some transforms due to the geometry of the scene. We propose to capture these features by sparse image expansions with geometric atoms taken from a redundant dictionary of functions.

The correlation model is then built on local geometric transforms between corresponding features taken in 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. These are particulary interesting for scene representation due to their wide field of view and the single center of projection that permits to accurately capture the scene geometry. Omnidirectional images can be mapped and processed on spherical manifolds; hence, we compute sparse image approximations on the sphere [5] in order to capture the most prominent image components. Local geometric 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 multiview coding.

We then exploit the geometric framework in the design of a distributed coding method with side information for multiview 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 transforms and to build a transform field between correlated views. These transforms are used to refine the side information for decoding the following atoms. Experimental results show that the proposed method successfully identifies the local geometric transforms 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 and generic geometric correlation model, which is further refined in 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 [6]. Most of the research in the DSC framework till nowadays focused on the application of DSC to low-complexity video coding [3], [7] and error-resilient video coding [3], [4]. 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 3-D 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, which 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 multiview correlation is modeled by relating the locations of discontinuities in the polynomial representation of image scanlines. This scheme has been extended to the case of natural 2-D images in [11]. Among state-of-the-art methods, the geometric approach for distributed coding in [11] is the closest to the work proposed in this paper in the sense that it exploits the epipolar constraint for the design of Slepian–Wolf code and for the joint decoding. However, authors in [11] consider only translations as correlation in multiple views (shifts of the discontinuities of the piecewise polynomials), while the correlation model in this paper includes translations, rotations and anisotropic scaling in a single framework.

Disparity-based solutions have also been proposed for distributed multiview video compression. Several works build on the advantages of distributed video coding for low complexity encoding and distributed multiview coding for exploiting both temporal and interview correlation [12]–[14]. They take different approaches for modeling the correlation among views, like the disparity-based model [12], affine model [14], or homography-based model [13]. Another direction for the distributed multiview video compression is based on classical motion compensated video encoding at each camera, while the interview correlation is exploited in a distributed manner [15]–[17]. Authors in [15] present a transform-based DSC method for multiview video coding that tracks epipolar correspondences between macroblocks in different views. The geometric information given by the epipolar constraint is exploited for joint multiview video decoding, but not for the design of the Slepian–Wolf code. Moreover, the Wyner–Ziv encoder has partial access to the side information (Intra macroblocks and motion vectors), so that this scheme cannot be classified as fully distributed multiview coding scheme. On the other side, a completely distributed stereo-view video coding method is proposed in [16]. 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. However, it does not exploit the correlation among I-frames, and, thus, the achieved bit rates are still quite far from the Slepian–Wolf bound. This gap can be reduced by encoding more coarsely the I-frames [17], but a lot of geometric

correlation between I-frames is still left unexploited. The work presented in this paper is substantially different from [15] and [17] since it uses epipolar geometry information for the design of the Slepian–Wolf code and, therefore, exploits the correlation between multiview images in a completely distributed manner.

The common characteristics of most state-of-the-art disparity-based distributed coding frameworks (except [11]) 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. The work that we propose in this paper 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 the correct decoding of Wyner–Ziv frames, even with a single reference frame that has been highly compressed.

III. Multiview Correlation Model

A. Geometric Image Representation

Images of a 3-D scene taken by distributed cameras are likely correlated as they capture the same objects in the scene from different viewpoints. The correlation between multiview images arises from the geometric constraints on the objects in the scene due to viewpoint change, and can be simply described by local changes of image components that represent the captured 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 transforms. However, image decompositions by common orthogonal transforms like the wavelet or DCT do not describe the image semantics. Due to the nonlocalized support and shift-variance 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) are capable of capturing the image structure and geometry using only few basis vectors [18], while offering excellent approximation performance. Sparse approximations have been also successfully applied to video [19], [20] and 3-D object compression [5]. One of the most important advantages of sparse approximations is the flexibility in the design of the overcomplete dictionary. When the dictionary is built on geometric functions with local support, the sparse image decomposition results in a set of meaningful geometric 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. Finally, the correlation between multiview images is driven by local transforms of sparse image components in different views that represent the same component in the 3-D scene. It is interesting to note here that sparse image approximations with redundant dictionaries of geometric features probably mimic the behavior of the human visual system for encoding visual information [21].

B. Sparse Approximations

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 $\mathcal{D} = \{\phi_k\}, k = 1, \ldots, N$, in a Hilbert space, every image y can be represented as

$$y = \mathbf{\Phi} x = \sum_{k=1}^{N} x_k \phi_k \tag{1}$$

where the matrix $\mathbf{\Phi}$ is composed of atoms ϕ_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 η, i.e.,

$$y = \mathbf{\Phi_I} c + \eta = \sum_{k \in I} x_k \phi_k + \eta \tag{2}$$

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 $\mathbf{\Phi_I}$ is a sub-matrix of $\mathbf{\Phi}$ 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:

$$\min_{c} \|c\|_0 \text{ subject to } \|y - \Phi_I c\|_2 \leq \varepsilon \tag{3}$$

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 (3) is replaced by an l_1 norm. For details on these algorithms, we refer the reader to [22].

C. Geometric Correlation Model

We are now interested in defining the correlation model between sparse approximations of two correlated multiview images[1]

$$y_1 = \mathbf{\Phi_{I_1}} c_1 + \eta_1$$
$$y_2 = \mathbf{\Phi_{I_2}} c_2 + \eta_2.$$

Since y_1 and y_2 capture the same 3-D 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 3-D features

[1]We take two images for the sake of clarity, but the correlation model that we develop can be generalized to any number of images.

in the scene. We assume that these atoms are correlated, possibly under some local geometric transforms. Let $F(\phi)$ denote the transform of an atom ϕ between two image decompositions that results from the change of camera viewpoint to the 3-D scene. 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 η_2

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

The above 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 (4) that relate corresponding atoms in sparse decompositions of multiview images. Due to the change of viewpoint on the 3-D scene 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_{γ_i} to another atom g_{γ_j} reduces to a transform of its parameters, i.e.,

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

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 transforms.

The correlation model given in (4) does not put any assumption on the type of cameras used for multiview image acquisition. It can be applied to planar or omnidirectional multiview images by introducing the epipolar geometry constraints that are defined for that particular image projection geometry. In the next section, we define the multiview transforms that satisfy the epipolar geometry for omnidirectional images, in particular, due to their advantages for compact 3-D scene representation.

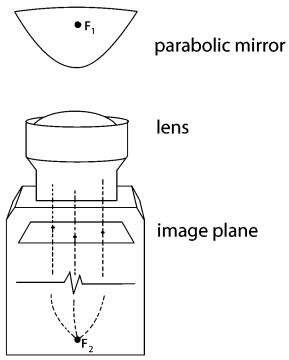


Fig. 2. Omnidirectional system with parabolic mirror.

IV. Geometric Transforms in Omnidirectional Images

A. Spherical Image Expansions

We focus now more specifically on omnidirectional images and describe in more detail the specific correlation model for images that can be mapped on spheres. The first research interests for omnidirectional imaging appeared with applications such as video surveillance [23], autonomous robot navigation [24], telepresence, etc. Recently, omnidirectional imaging became an interesting and increasingly popular framework for 3-D scene representation, as it offers a wider field of view and, therefore, necessitates only a small number of camera sensors for capturing a 3-D scene. With a single point of projection, omnidirectional cameras record the light field in its radial form. This permits to process the visual information without the discrepancies introduced by Euclidian assumptions in planar imaging. Moreover, any perspective image on any designated image plane or any panoramic image can be generated from the captured omnidirectional image. Therefore, we address the problem of correlation modeling for multiview omnidirectional images. We use omnidirectional cameras which are constructed by placing a parabolic mirror in front of a camera with an orthographically projecting lens as depicted in Fig. 2. Such parabolic catadioptric sensor outputs an omnidirectional image.

As these images can be precisely mapped on a sphere through inverse stereographic projection [25], [26], we further use a dictionary of atoms on the 2-D unit sphere [5]. 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 (τ, ν, ψ) represents Euler angles that respectively describe the motion of the atom on the sphere by angles τ and ν, and the rotation of the atom around its axis with an angle ψ, and α, β represent anisotropic scaling factors. As an example, Gaussian atoms on the sphere are illustrated in Fig. 3, 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

Fig. 3. Gaussian atoms: (a) on the North pole $(\tau = 0, \nu = 0), \psi = 0, \alpha = 2, \beta = 4$; (b) $\tau = (\pi/4), \nu = (\pi/4), \psi = (\pi/8), \alpha = 2, \beta = 4$; (c) $\tau = (\pi/4), \nu = (\pi/4), \psi = (\pi/8), \alpha = 1, \beta = 8$.

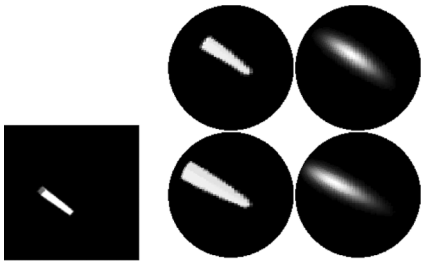


Fig. 4. Example of atoms transform in approximation of two views of a 3-D scene. (a) Simple 3-D scene. (b) First column: two captured spherical images of the scene; second column: sparse approximations of the two views with one Gaussian atom per view. The atom in the approximation of the second view is related to the atom in the first view by the following transform: $(\tau_2 - \tau_1 = \pi/128, \nu_2 - \nu_1 = -12\pi/128, \psi_2 - \psi_1 = -\pi/16, \alpha_2/\alpha_1 = 1, \beta_2/\beta_1 = 0.8)$.

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 that two corresponding atoms g_{γ_i} and h_{γ_j} in images y_1 and y_2 are linked by a simple transform of the atom parameters, and (5) can be rewritten as

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

The subset of transforms $V_i^0 = \{\gamma' \,|\, h_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma')g_{\gamma_i}\}$ allows to relate g_{γ_i} to the atoms h_{γ_j} in the expansion of y_2. Fig. 4 depicts an example of a simple 3-D scene captured by two omnidirectional cameras and shows how their sparse approximations can be linked with a transform of atom parameters. However, not all transforms in V_i^0 are feasible in multiview correlated images. The set of possible transforms can be greatly reduced by identifying two constraints between corresponding atoms, namely *shape similarity* constraint and *epipolar* constraint.

B. Geometric Constraints

First, we assume that the change of viewpoint on a 3-D object results in a limited difference between shapes of corresponding atoms since they represent the same object in the scene. Therefore, we can restrict the set of possible transforms by the shape similarity constraints between candidate atoms. From the set of atom parameters γ, the last three parameters (ψ, α, β) describe the atom shape (its rotation and scaling), and, therefore, they are taken into account for the shape similarity constraint. We measure the similarity or coherence of atoms by the inner product $\mu(i,j) = |\langle g_{\gamma_i}, h_{\gamma_j}\rangle|$ between centered atoms (at the same position (τ,ν)), 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

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

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

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

Second, pairs of atoms that correspond to the same 3-D points have to satisfy epipolar constraints, that represent one of the fundamental relations in multiview analysis [27]. The epipolar constraint defines the relation between 3-D point projections $(\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^3)$ on two cameras, as

$$\mathbf{z}_2^T \hat{T} R \mathbf{z}_1 = 0 \tag{9}$$

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_{γ_i} in y_1 and the candidates atoms h_{γ_j} in y_2. The constraint given in (9) is rarely exactly satisfied for corresponding pixels or areas in two multiview images, and the decision on the epipolar matching of two correspondences is commonly taken when their epipolar distance is smaller than a certain threshold κ.

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_{γ_i} as

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

where $d_{\mathrm{EA}}(g_{\gamma_i}, h_{\gamma_j})$ denotes the epipolar distance between atoms g_{γ_i} and h_{γ_j}. This distance measures how much atoms g_{γ_i} and h_{γ_j} deviate from the perfect epipolar matching defined by the (9) and it is defined in the Section V. 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

$$\Gamma_i^E = \left\{\gamma_j \,|\, h_{\gamma_j} = U(\gamma')g_{\gamma_i}, \gamma' \in V_i^E\right\}. \tag{11}$$

A graphical interpretation of the epipolar constraint for spherical images is shown in Fig. 5, where we denote as S_1 and S_2 the two unit spheres corresponding to camera projection surfaces. A given atom g_{γ_i} in y_1, on the sphere S_1, can be a projection of infinitely many different 3-D objects, at different scales and distances from S_1. We show an example of several different objects whose projection on S_1 is g_{γ_i} and projections on S_2 are h_{γ_j}. Due to epipolar constraints, the atoms h_{γ_j} are positioned on the part of a great circle $\mathcal{C}_i$ 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_{γ_i} on the sphere S_1. For more details on epipolar geometry for spherical images we refer the interested reader to [28].

Finally, we combine the epipolar and shape similarity constraints to define the set of possible transforms for atom g_{γ_i}, as $V_i = V_i^E \cap V_i^\mu$. Similarly, we denote the set of possible param-

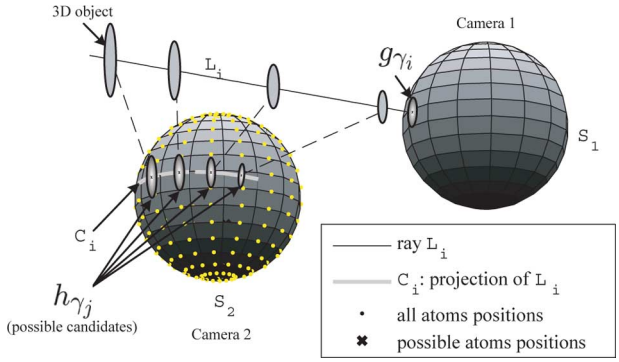


Fig. 5. Selection of positions of atoms that satisfy epipolar constraints.

eters of the transformed atom in y_2 as $\Gamma_i = \Gamma_i^E \cap \Gamma_i^\mu$. Given the set Γ_i of possible atom parameters in y_2 corresponding to the atom g_{γ_i} in y_1, the correspondence h_{γ_j} in y_2 can be uniquely defined with high probability under the assumption that the decomposition of y_2 is sparse.

V. Disparity Map Estimation by Atom Transforms

The local transforms 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 allows for view interpolation under epipolar constraints. It is defined as the point-wise correlation between multiview 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 (9) is satisfied. The dense disparity mapping is most commonly estimated based on pixel-wise correlation between rectified stereo images, or by block-based matching. The performance of these approaches unfortunately deteriorates with the decrease of the image quality, like for images compressed at very low bit rates. The disparity map can also be estimated by identifying corresponding feature points in multiview images and relating them with a cross-correlation similarity measure, like proposed in [29]. However, cross-correlation measure [30] 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 feature 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. Moreover, this estimation can be performed with images that are encoded at very low bit rates, regardless of the image quality. 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 us 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_{γ_i} to its corresponding point on h_{γ_j}. Since this mapping is point-wise, we need to define g_{γ_i} in the discrete space, i.e., on the spherical grid $\mathcal{G}_1$. Then, the disparity mapping translates to the grid deformation induced by the local transform between g_{γ_i} and h_{γ_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_{γ_i}, given by Euler angles (τ_i, ν_i, ψ_i), into the motion of atom h_{γ_j}, given by Euler angles (τ_j, ν_j, ψ_j);
2) transform of anisotropic scaling of the atom g_{γ_i}, given by the pair of scales (α_i, β_i), into the anisotropic scaling of the atom h_{γ_j}, given by the pair of scales (α_j, β_j).

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

$$\mathbf{z}_2 = R_{\gamma_j}^{-1} \cdot \zeta(R_{\gamma_i} \cdot \mathbf{z}_1) \tag{12}$$

where R_{γ_i} and R_{γ_j} are rotation matrices given by Euler angles (τ_i, ν_i, ψ_i) and (τ_j, ν_j, ψ_j), respectively, and $\zeta(\cdot)$ 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_{γ_i}, then deformed with respect to anisotropic scaling, and finally rotated back with the rotation matrix of atom h_{γ_j}.

In more detail, the stereographic projection [31] at the North pole projects a point (θ, φ) on the sphere to a point (x, y) on the plane tangent to the North pole, and it is formally given with

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

Now let (θ_1, φ_1) and (θ_2, φ_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 (θ_1, φ_1) on the grid $\mathcal{G}_1$ to the point (θ_2, φ_2) on the grid $\mathcal{G}_2$ due to anisotropic scaling can be obtained by scaling the stereographic projection of (θ_1, φ_1) with $1/\alpha_j$ and $1/\beta_j$, in the following way:

$$\begin{aligned} x_2 = \rho_2 \cos\varphi_2 &= \frac{1}{\alpha_j}\alpha_i x_1 = \frac{\alpha_i}{\alpha_j}\rho_1 \cos\varphi_1 \\ y_2 = \rho_2 \sin\varphi_2 &= \frac{1}{\beta_j}\beta_i y_1 = \frac{\beta_i}{\beta_j}\rho_1 \sin\varphi_1 \end{aligned} \tag{14}$$

where $\rho_2 = 2\tan\theta_2/2$ and $\rho_1 = 2\tan\theta_1/2$. By solving the system of (14) for θ_2 and φ_2, we get

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

We can, therefore, define the function $\zeta(\cdot)$ as a pair of transforms $\zeta_p(\varphi_1)$ and $\zeta_t(\theta_1, \varphi_1, \zeta_p(\varphi_1))$ followed by the transform of spherical coordinates (θ_2, φ_2) to Euclidean coordinates $\mathbf{z}_2$. The relation given in (12) is now completely defined, based on the parameters of corresponding atoms in two images. When the transform 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_{γ_i} and h_{γ_j} related by the disparity map. The symmetric epipolar atom distance actually measures how much the atom pair g_{γ_i} and h_{γ_j} deviates from the perfect epipolar matching given in the correlation model of (10), when $d_{\mathrm{EA}}(g_{\gamma_i}, h_{\gamma_j}) = 0$. It is evaluated as the weighted average of the symmetric epipolar distance of all pairs of points given by the disparity map

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

The points $\mathbf{z}_1$ and $\mathbf{z}_2$ are related by the disparity map and $d_{\mathrm{SE}}(\mathbf{z}_1, \mathbf{z}_2)$ stands for the symmetric epipolar distance between $\mathbf{z}_1$ and $\mathbf{z}_2$ [30]. It is defined as

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

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 $\mathbf{z}_2$ (see Fig. 8 for an illustration of this distance). The weight w_{γ_i} is a normalized weight function that prioritizes the points where the atom g_{γ_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 geometric component captured by the atom (typically edges). One example could be a 2-D Gaussian weight function, anisotropically scaled and oriented, which fits the atom g_{γ_i}. If the overcomplete dictionary is composed of Gaussian atoms, the weight function is equal to the atom itself. We use 2-D Gaussian weight function in the rest of this paper.

VI. Distributed Coding of 3-D Scenes

A. Encoder and Coset Design

The correlation model introduced above 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 $R_{y_2} \geq H(y_2|y_1)$. This corresponds to an asymmetric scheme, where the rate is not balanced between the encoders. 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 quantization of their respective coefficients, as shown in Fig. 6.

We propose to partition the set of atom indexes Γ into distinct cosets that contain dissimilar atoms with respect to their position and shape. Under the assumption that an atom h_{γ_j} in the image decomposition has its corresponding atom g_{γ_i} in the side information expansion, the Wyner–Ziv encoder does not need to code the entire γ_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 (8) and (11). The side information and the coset index are, therefore, sufficient to recover the atom g_{γ_j} in the Wyner–Ziv image. The achievable

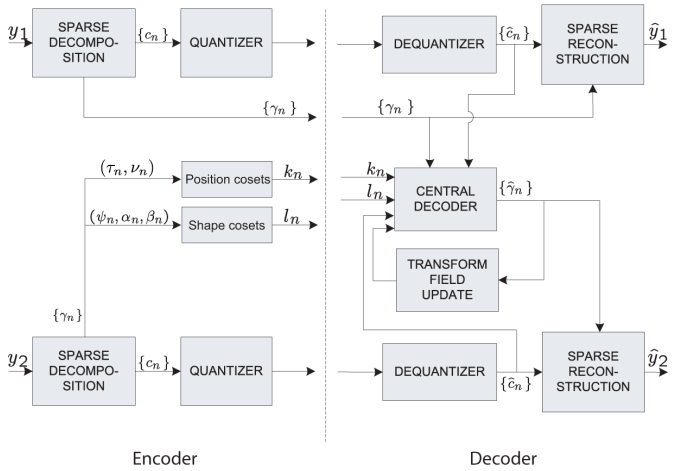


Fig. 6. Block diagram for the Wyner–Ziv codec.

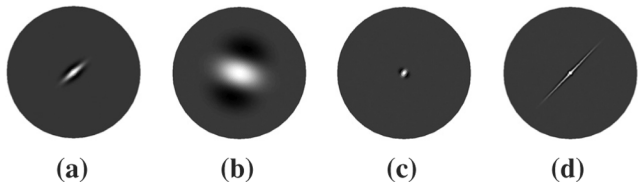


Fig. 7. Samples of atoms in the same Shape coset.

bit rate for encoding the atom index γ_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 (ψ, α, β), and for atom positions (τ, ν) according to epipolar constraints. We, therefore, construct two types of cosets, respectively the Shape cosets: $K_l^\mu, l = 1, \ldots, N_2$ and the Position cosets $K_k^E, k = 1, \ldots, N_1$. The encoder eventually sends for each atom only the indexes of the corresponding cosets (i.e., k_n and l_n in Fig. 6). We design Shape cosets by distributing all atoms whose parameters belong to Γ_i^μ into different cosets. Samples of atoms that belong to the same Shape coset are illustrated in Fig. 7, showing the significant difference in their shapes.

Next, we propose two design methods for constructing the Position cosets, that correspond to the scenarios where the camera pose (R, T) is known, or not available, respectively. We first design *Epipolar (EPI) cosets* based on the fact that the centers of two corresponding atoms g_{γ_i} and h_{γ_j}, determined by the coordinates of their positions (τ_i, ν_i) and (τ_j, ν_j) and 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 (10) when $\mathcal{G}_1 = m_i$, which transforms into $\mathcal{G}_2 = m_j$. The epipolar candidates set given in (11) reduces to

$$\Gamma_i^M = \{\gamma_j | h_{\gamma_j} = U(\gamma') g_{\gamma_i}, d_{\mathrm{SE}}(m_i, m_j) \leq \delta\} \tag{18}$$

where δ 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 Γ_i^M for $\mathcal{G}_1 = m_i$. This is illustrated in Fig. 8 for an exemplary epipolar line.

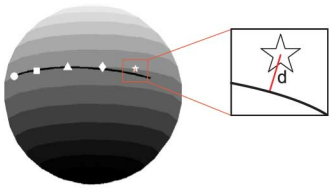


Fig. 8. Illustration of the epipolar coset design. The atoms with positions marked with different shapes are put in different epipolar cosets since they belong to the same epipolar line. Zoomed: Example of the distance $d(\mathbf{z}_1, \mathcal{C}_{\mathbf{z}_2})$ that is equal to the Euclidean distance of a point shown by a star to the epipolar line, and denoted as d.

The parameter δ 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_{γ_i}, the decoder only needs to know the coset index of h_{γ_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-D interleaved uniform quantization of atom positions (τ, ν) on a rectangular lattice. This coset design can be formulated similarly to the Epipolar coset design, where the set of position candidates (called the set of epipolar candidates in (18)) gathers the candidates positions (τ_j, ν_j) within the neighborhood of the reference atom position (τ_i, ν_i), i.e.,

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

The interleaved vector quantization of τ and ν will distribute the pairs (τ, ν) that belong to the same Γ_i^V into different cosets, while keeping the distance between coset elements constant and equal to $(\Delta\tau, \Delta\nu)$. Note that the constant intracoset 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 transforms, in order to establish correspondences between atoms in the reference image and atoms within the cosets of the Wyner–Ziv image decomposition (see Fig. 6). 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 nth 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 (τ_n, ν_n) from $K_{k_n}^E$ and the atom shape $(\psi_n, \alpha_n, \beta_n)$ from $K_{l_n}^\mu$. The pseudo-code for the decoder is given in the Algorithm 1. Let A_n denote the set of possible candidates for decoding the nth 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_{γ_i} and h_{γ_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_{γ_i} and h_{γ_j} prior to atom normalization. Therefore, the decoder can select a subset of atoms $J_n = \{\gamma_i\}$ in I_1 whose coefficients satisfy

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

where σ is a chosen threshold. For each $g_{\gamma_i} \in J_n$ we have a set of possible transformed atoms given by $\tilde{\Gamma}_i = \Gamma_i^M \bigcap \Gamma_i^\mu$ or $\tilde{\Gamma}_i = \Gamma_i^V \bigcap \Gamma_i^\mu$ 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 Γ_i^M, Γ_i^V, and Γ_i^μ can have different values for the coset design and for the decoding. This permits to put stricter conditions for the selection of corresponding atom pairs.

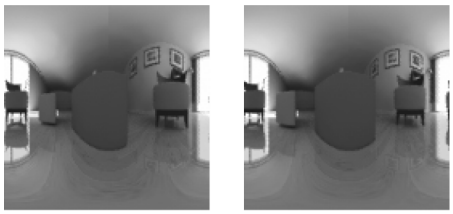
The search for atom correspondences then proceeds in two major steps. First, the decoder eliminates the candidates that do not belong to $\bigcup_{\gamma_i \in J_n} \tilde{\Gamma}_i$, as well as candidates with a large symmetric epipolar atom distance, i.e., for which $d_{\mathrm{EA}}(g_{\gamma_i}, h_{\gamma_j}) > \kappa$. If all candidates in A_n get eliminated, the decoder decides that the nth 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 $d_{\mathrm{EA}}(g_{\gamma_i}, h_{\gamma_j})$ 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 $\mathbf{z}_1^{(i)}, i = 1, \ldots, n$, defined by n correspondences. The final mapping point $\mathbf{z}_1^*$ is selected as

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

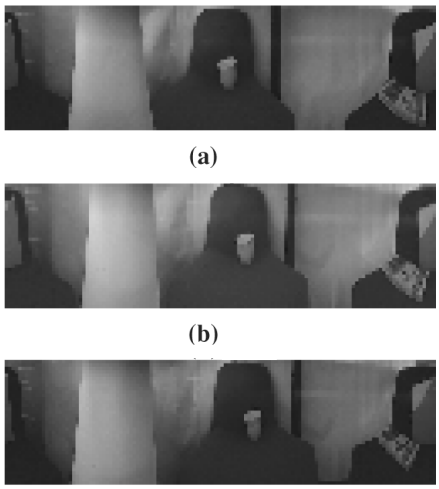
where we have used the same weight function as for the symmetric epipolar atom distance.

The transform 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



(a) (b)

Fig. 9. Original Room images (128 × 128). (a) y_1; (b) y_2.



(a) (b) (c)

Fig. 10. 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 cropped images from the 128 × 128 spherical images, which correspond to the captured scene. (a) y_3; (b) y_1; (c) y_2.

following atoms in the Wyner–Ziv image expansion. The atoms that do not have any correspondence in the reference frame are decoded based on the mean square error between the side information $(y_{\rm tr})$ and the Wyner–Ziv image reconstructed from previously decoded atoms and the current decoding candidate from A_n. The candidate from A_n with the minimal mean square error is selected as the decoded atom.

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 $\mathbf{\Phi_{I_2}}$, and the transformed reference image $y_{\rm tr}$, i.e.,

$$\hat{y}_2 = y_d + \lambda \Psi_d \, y_{\rm tr}. \tag{23}$$

The matrix Ψ_d denotes the orthogonal complement to the basis formed by the decoded atoms in $\mathbf{\Phi_{I_2}}$, and λ 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 λ from the energy conservation principle. Namely, under the assumption that $||\Psi_d \, y_{\rm tr}|| \approx ||\Psi_d \, y_2||$, we get λ from (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.

Algorithm 1 Decoder

initialization: $TF = \mathcal{G}_1, \Phi = [\,], C = [\,]$;
input: $\tilde{c}_i, g_{\gamma_i}, i = 1, \ldots, N_r$, evaluate $\hat{y}_1 = \sum_{i=1}^{N_r} \tilde{c}_i g_{\gamma_i}$;
for $\mathrm{n} = 1, \ldots, \mathrm{N}$ **do**
  input: $k_n, l_n, \hat{c}_n$
  initialization: $W_c = \varnothing$ (set of paired candidates)
  $A_n = \{\gamma \,|\, (\tau, \nu) \in K^E_{k_n}, (\psi, \alpha, \beta) \in K^\mu_{l_n}\}$
  for all $\gamma_p \in A_n$ **do**
    for all $\tilde{c}_i, g_{\gamma_i}, i = 1, \ldots, N_r$ **do**
      if $(\Delta c = |(\tilde{c}_i - \hat{c}_n/\hat{c}_n)| < \sigma) \& (\gamma_p \in \tilde{\Gamma}_i) \& (d_{\rm EA}(g_{\gamma_i}, h_{\gamma_p}) < \kappa)$ **then**
        add γ_p to W_c
      end if
    end for
  end for
  if W_c not empty **then**
    $\hat{\gamma}_n = \arg\min_{\gamma_p} |d_{\rm EA}(g_{\gamma_i}, h_{\gamma_p})|$;
    announce $\hat{\gamma}_n$ decoded, update TF
    $\Phi = [\Phi; h_{\hat{\gamma}_n}], C = [C; \hat{c}_n]$
  else
    γ_n not decoded.
  end if
end for
$y_{\rm tr} = TF(\hat{y}_1)$
for $\mathrm{n} = 1, \ldots, \mathrm{N}$ **do**
  if γ_n not decoded **then**
    for all $\gamma_p \in A_n$ **do**
      $\Phi_p = [\Phi; h_{\hat{\gamma}_p}], C_p = [C; \hat{c}_n], e_p = ||y_{\rm tr} - \Phi_p^\dagger C_p||^2$
    end for
    $\hat{\gamma}_n = \arg\min_{\gamma_p} |e_p|$, $\Phi = [\Phi; h_{\hat{\gamma}_n}], C = [C; \hat{c}_n]$
  end if
end for
$y_d = \Phi^\dagger C, \lambda \approx \sqrt{1 - ||y_d||^2/||y_2||^2}$, $\hat{y}_2 = y_d + \lambda \Psi_d \, y_{\rm tr}$

VII. Experimental Coding Results

A. Experimental Settings

We analyze here the performance of the above Wyner–Ziv coding method for two sets of multiview images: synthetic

Fig. 11. DSC results for the Room images: (a) decoded reference image $\hat{y}_1$ ($\mathrm{PSNR} = 30.95$ dB); (b) transformed reference image y_{tr}; (c) decoded Wyner–Ziv image $\hat{y}_2$ at 0.0534 bpp; (d) decoded second image $\hat{y}_2^{\mathrm{MP}}$ 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. (a) $\hat{y}_1$; (b) y_{tr}; (c) $\hat{y}_2$; (d) $\hat{y}_2^{\mathrm{MP}}$; (e) $1-|\hat{y}_1 - y_2|$; (f) $1-|\hat{y}_2 - y_2|$.

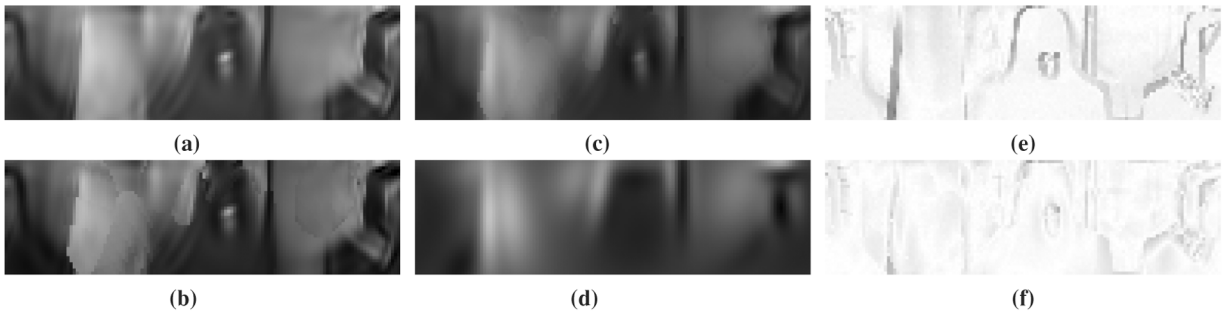


Fig. 12. DSC results for the Lab images: (a) decoded reference image $\hat{y}_1$ ($\mathrm{PSNR} = 29.4$ dB); (b) transformed reference image y_{tr}; (c) decoded Wyner–Ziv image $\hat{y}_2$ at 0.035 bpp; (d) decoded second image $\hat{y}_2^{\mathrm{MP}}$ when encoded with MP at 0.035 bpp; (e) inverted 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. (a) $\hat{y}_1$; (b) y_{tr}; (c) $\hat{y}_2$; (d) $\hat{y}_2^{\mathrm{MP}}$; (e) $1-|\hat{y}_1 - y_2| = 1-|e_1|$; (f) $1-|\hat{y}_2 - y_2| = 1-|e_2|$.

spherical images of the Room scene (Fig. 9) and natural omnidirectional images of the Lab scene (Fig. 10). Room scene images include two 128×128 spherical images y_1 and y_2 captured from different viewpoints, with the relative pose of one camera with respect to the other given as $R = I$ and $T = [0\ 0.3\ 0]^T$. The Lab scene images include three natural omnidirectional images y_3, y_1 and y_2, taken from omnidirectional cameras placed in a straight line in order 3-1-2 (camera number corresponds to the index of the image). Captured images are mapped to 128×128 spherical images as explained in [32]. We have used the catadioptric sensor from the Remote Reality Corporation,[2] which has 360° view in the azimuth angle and the elevation view which ranges from 35° to 92.5°. For the Lab scene we have used the approximate camera pose for the specific linear camera arrangement: $R \approx I$ and $T \approx [1\ 0\ 0]$.[3]

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 2-D Gaussian function, given as

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

The second function represents a Gaussian in one direction and the second derivative of a 2-D Gaussian in the orthogonal direction (i.e., edge-like atoms similar to the ones presented in [5]). It is written as

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

where K is a normalization factor. The position parameters τ and ν can take 128 different values ($N_t = N_p = 128$), while the rotation parameter uses 16 orientations, between 0 and π. The scales are distributed in a logarithmic scale of power of 2, from 1 to $N_t/8$ for the Gaussian atoms and from 2 to $N_t/2$ for edge-like atoms, with three scales per octave. The choice of the dictionary is mainly driven by its good approximation properties demonstrated in [5].

B. Rate-Distortion Analysis of the Wyner–Ziv Image

We first evaluate the performance of the proposed Wyner–Ziv coding method by taking the image y_1 as a reference image and image y_2 as Wyner–Ziv image, for both image sets. 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 [33]. The decoded reference images for the Room and Lab scene are shown in Figs. 11(a) and 12(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

[2]http://www.remotereality.com/

[3]Since we do not perform depth estimation, we can use the camera translation vector which is correct up to a scale factor, hence the value 1 for the translation on x axis.

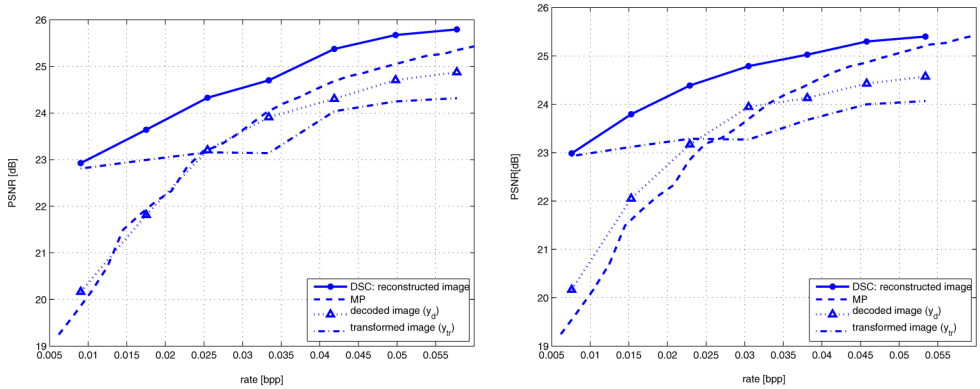


Fig. 13. Rate-distortion performance for the Room image set. (a) EPI position cosets; (b) VQ position cosets.

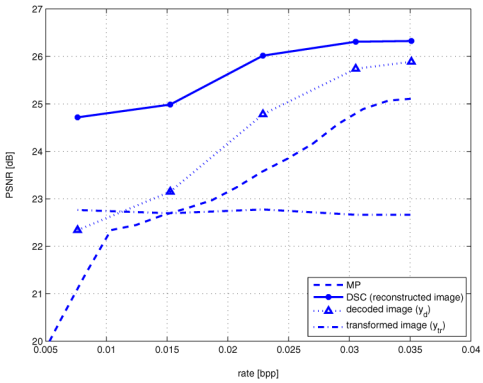


Fig. 14. Rate-distortion performance for the Lab image set (VQ Position cosets). Image y_2 was used as Wyner–Ziv image and image y_1 as reference image.

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 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 Fig. 13(a) and (b) for the Room scene (for EPI and VQ cosets respectively), and in Fig. 14 for the Lab scene. The bit rate is changed with the number of received atoms, while the quantization of coefficients is kept constant. 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. It shows that the transform field significantly improves the side information for the Room images, while for the Lab images the improvement is smaller. 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 performance of the JPEG2000 has also been evaluated for the independent compression of the Wyner–Ziv image. Since JPEG2000 cannot perform at such low rates, we have evaluated the rate at which JPEG2000 reaches the PSNR performance of the proposed DSC coder. For the Room image, JPEG2000 gives the PSNR of 25.5 dB at 0.12 bpp, which is ≈ 2.4 times greater than the rate DSC coder needs to obtain the same image quality. For the Lab image JPEG2000 gives PSNR of 25.7 dB at 0.095 bpp, where the compression was done on cropped omnidirectional images. Again, the proposed DSC coder achieves this PSNR at only 1/4 of the rate that JPEG2000 needs.

The images y_{tr} and $\hat{y}_2$ are presented in Figs. 11(b) and (c) and 12(b) and (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 0.035 bpp. We can clearly see how the transform field deforms the reference image in order to compensate for different object transforms. Figs. 11(d) and 12(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 $\hat{y}_2$. However, the quality of the image $\hat{y}_2$ is still lower than the quality of the encoded reference image $\hat{y}_1$, showing that the method results in unbalanced image qualities. On the other side, independent coding of two images at the same overall rate could give balanced qualities, but at the price of smaller PSNR for the image considered as reference image in the DSC scheme. In order to achieve more balanced PSNR of multiview images in the distributed coding settings, the proposed scheme could be complemented with a more efficient method for coding the texture difference between multiple views, or by substituting the coding with side-information with a balanced distributed coding scheme. Such approaches can be envisaged in the future work as extensions or applications of the correlation model proposed in this paper.

C. Overall Rate-Distortion Performance

Fig. 15 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

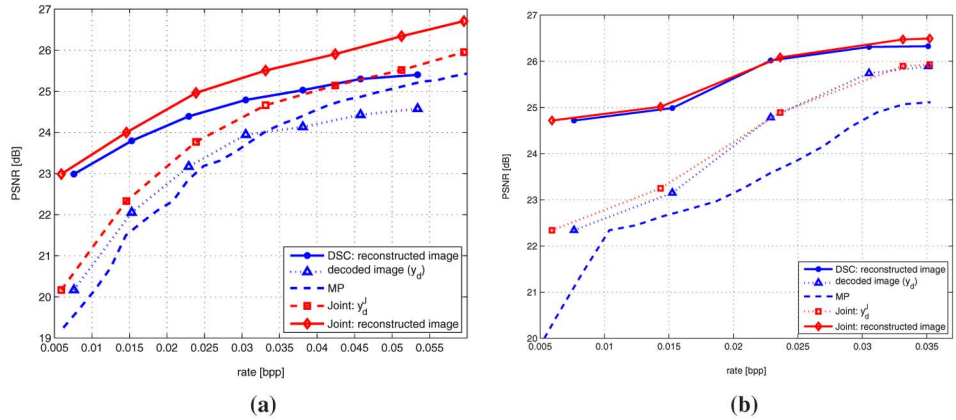


Fig. 15. Comparison of rate-distortion performance for distributed coding and joint encoding (VQ coset design). (a) Room image set; (b) lab image set.

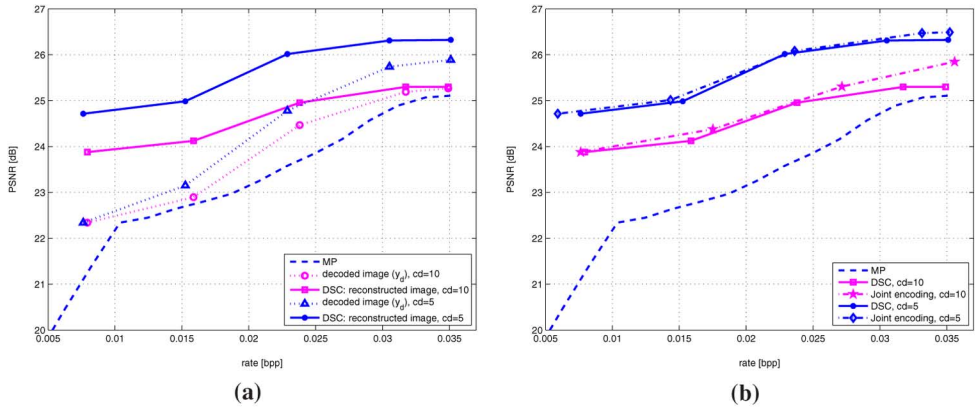


Fig. 16. Influence of the camera distance on the DSC performance for the Lab image set. (a) Comparison of the rate-distortion performance for the reference camera distance $\mathrm{cd} = 5$ and $\mathrm{cd} = 10$. (b) Comparison of rate-distortion performance for distributed coding and joint encoding for $\mathrm{cd} = 5$ and $\mathrm{cd} = 10$.

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} 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.

The influence of the camera distance on the performance of the proposed DSC scheme has been evaluated for the Lab image set. The Wyner–Ziv image y_2 is decoded using the image y_3 as a reference image and compared to the decoding of the same Wyner–Ziv image y_2 when using the image y_1 as a reference. The distance between cameras 2 and 3 is set to $\mathrm{cd} = 10$, while the distance between cameras 1 and 2 is equal to $\mathrm{cd} = 5$. Therefore, the disparity between images y_2 and y_3 is certainly greater than the disparity between images y_2 and y_1. For decoding y_2 using y_3 as reference the number of Shape cosets is increased to 256 cosets, while the number of Position cosets stays the same, due to smaller correlation between the Wyner–Ziv image and the reference image y_3. Fig. 16(a) compares the RD performance of the DSC coding of image y_2 with different reference images: y_1 and y_3 corresponding to camera distances $\mathrm{cd} = 5$ and $\mathrm{cd} = 10$ respectively. We can see that the RD curves for the decoded images y_d in these two cases are close, showing that the atom decoding process based on the proposed correlation model is not much influenced by the increase of disparity between images. Certainly, the RD curve of the reconstructed image $\hat{y}_2$ for $\mathrm{cd} = 5$ outperforms for 1 dB the same curve for $\mathrm{cd} = 10$ due to the higher correlation of the Wyner–Ziv image with the reference image and, hence, with the transformed image, as well. Fig. 16(b) presents the comparisons of the DSC method with the joint encoding method for $\mathrm{cd} = 5$ and $\mathrm{cd} = 10$, showing that the DSC scheme performs close to the joint encoding strategy for both cases of camera distances. Fig. 17(a) and (b) shows, respectively, the decoded reference image $\hat{y}_3$ and its transformed version y_{tr}, for the case of $\mathrm{cd} = 10$. Similarly as for the smaller

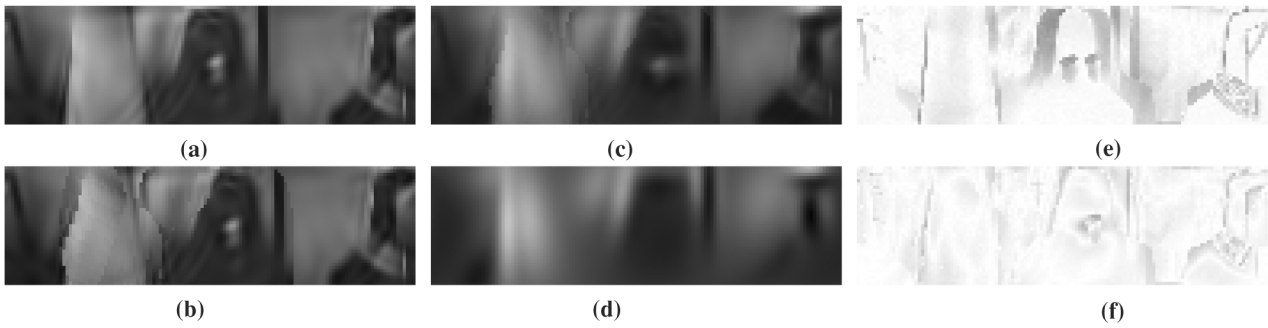


Fig. 17. DSC results for the Lab images: (a) decoded reference image $\hat{y}_3$ ($\mathrm{PSNR} = 31.82$ dB); (b) transformed reference image y_{tr}; (c) decoded Wyner–Ziv image $\hat{y}_2$ at 0.035 bpp; (d) decoded second image $\hat{y}_2^{\mathrm{MP}}$ when encoded with MP at 0.035 bpp; (e) inverted residue $1 - |e_3| = 1 - |\hat{y}_3 - y_2|$ without transform compensation (white pixel denotes no error); (f) inverted residue $1 - |e_2| = 1 - |\hat{y}_2 - y_2|$ after DSC decoding. (a) $\hat{y}_3$; (b) y_{tr}; (c) $\hat{y}_2$; (d) $\hat{y}_2^{\mathrm{MP}}$; (e) $1 - |\hat{y}_3 - y_2| = 1 - |e_3|$; (f) $1 - |\hat{y}_2 - y_2| = 1 - |e_2|$.

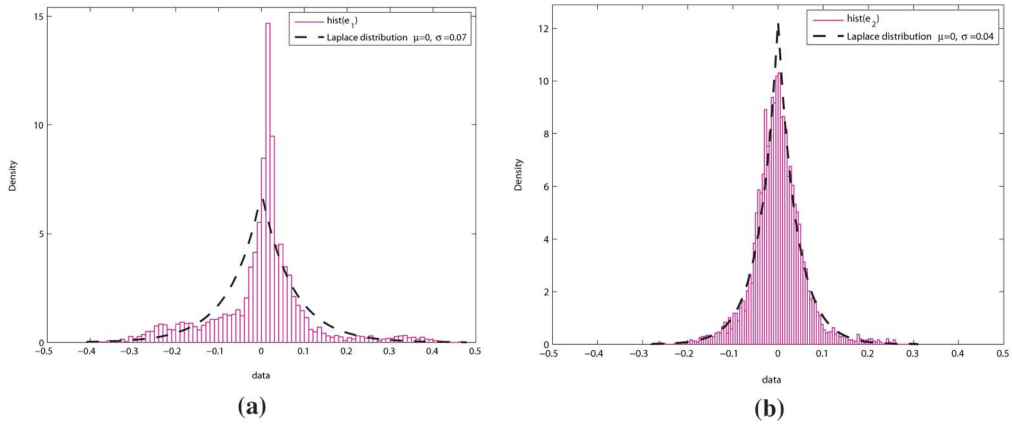


Fig. 18. (a) Laplacian distribution fitted to the histogram of the residue e_3 for the Lab scene with $\mathrm{cd} = 10$. (b) Laplacian distribution fitted to the histogram of the residue e_2 with $\mathrm{cd} = 10$ (fitting is performed with the Matlab statistics toolbox, using the maximum likelihood estimator).

camera distance, in the case of $\mathrm{cd} = 10$ the quality of the DSC coded Wyner–Ziv image $\hat{y}_2$, shown in Fig. 17(c), is higher than the quality of the same image encoded at the equal rate with MP, shown on the Fig. 17(d).

Finally, we discuss 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 $e_1 = \hat{y}_1 - y_2$ (residue without transform compensation), or $e_3 = \hat{y}_3 - y_2$ for the Lab scene with y_3 as a reference image $(cd = 10)$. Figs. 11(c) and (e) and 12(c) and (e) show, respectively, the inverted residues $1 - |e_1|$ and $1 - |e_2|$ for Room and Lab scene with $\mathrm{cd} = 5$, while Fig. 17(c) and (e) shows and the inverted residues $1 - |e_3|$ and $1 - |e_2|$ for the Lab scene with $\mathrm{cd} = 10$. The white pixels correspond to no error. The energy of the residue without transform compensation is respectively 82.65, 37.89, and 52.9661 for the Room images and Lab images with $\mathrm{cd} = 5$ and $\mathrm{cd} = 10$, where the energy is given by the square of the norm with the inner product computed on the sphere. The energy of the error e_2 is 47.12, 11.0380, and 13.9376 respectively, which confirms the efficiency of the model based on local geometric transforms. 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 pixel values in the residue image after transform compensation and decoding can be modeled with the Laplace distribution (see Fig. 18). It greatly facilitates the correlation modeling towards the potential DSC encoding of the residual texture information in hybrid coding approaches.

VIII. Conclusion

We have presented a geometry-based framework for the efficient representation of 3-D scenes, where camera images are approximated by sparse expansions of prominent geometric features. A novel correlation model has been proposed based on local geometric transforms that permit to pair atoms in different images under shape and epipolar geometry constraints. It provides an implicit estimation of the scene geometry that proves to be useful in the design of 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 3-D scenes at low bit rate. It can lead to effective solutions for distributed sensing and processing of 3-D 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 Trans. Inf. Theory*, vol. 19, no. 4, pp. 471–480, Jul. 1973.

[2] A. D. Wyner and J. Ziv, "The rate-distortion function for source coding with side-information at the decoder," *IEEE Trans. Inf. Theory*, vol. 22, no. 1, pp. 1–10, Jan. 1976.

[3] B. Girod, A. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proc. IEEE*, vol. 93, no. 1, pp. 71–83, Jan. 2005.

[4] A. Sehgal, A. Jagmohan, and N. Ahuja, "Wyner–Ziv coding of video: An error-resilient compression framework," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 249–258, Apr. 2004.

[5] I. Tosic, P. Frossard, and P. Vandergheynst, "Progressive coding of 3-D objects based on overcomplete decompositions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 11, pp. 1338–1349, Nov. 2006.

[6] S. S. Pradhan and K. Ramchandran, "Distributed source coding using syndromes (DISCUS)," *IEEE Trans. Inf. Theory*, vol. 49, no. 3, pp. 626–643, Mar. 2003.

[7] R. Puri and K. Ramchandran, "PRISM: A video coding paradigm with motion estimation at the decoder," *IEEE Trans. Image Process.*, vol. 16, no. 10, pp. 2436–2448, Oct. 2007.

[8] X. Zhu, A. Aaron, and B. Girod, "Distributed compression for large camera arrays," presented at the IEEE SSP, St. Louis, MO, Sep. 2003.

[9] R. Wagner, R. Nowak, and R. Baraniuk, "Distributed image compression for sensor networks using correspondence analysis and super-resolution," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2003, vol. 1, pp. 597–600.

[10] N. Gehrig and P. L. Dragotti, "DIFFERENT—Distributed and fully flexible image encoders for camera sensor networks," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2005, vol. 1, pp. 690–693.

[11] N. Gehrig and P. L. Dragotti, "Distributed compression of multi-view images using a geometrical coding approach," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2007, vol. 6, pp. VI-421–VI-424.

[12] M. Flierl and P. Vandergheynst, "Distributed coding of highly correlated image sequences with motion-compensated temporal wavelets," *EURASIP J. Appl. Signal Process.*, vol. 2006, p. 10, 2006, Article ID 46747.

[13] M. Ouaret, F. Dufaux, and T. Ebrahimi, "Fusion-based multiview distributed video coding," in *Proc. ACM Int. Workshop on Video Surveillance and Sensor Networks*, Oct. 2006, pp. 139–144.

[14] X. Guo, Y. Lu, F. Wu, W. Gao, and S. Li, "Distributed multi-view video coding," in *Proc. SPIE VCIP*, Jan. 2006, vol. 6077, p. 8.

[15] B. Song, E. Tuncel, and A. K. Roy-Chowdhury, "Towards a multi-terminal video compression algorithm by integrating distributed source coding with geometrical constraints," *J. Multimedia*, vol. 2, no. 3, pp. 9–16, June 2007.

[16] Y. Yang, V. Stankovic, W. Zhao, and Z. Xiong, "Multiterminal video coding," in *Proceedings of IEEE UCSD Workshop on Information Theory and its Applications*, Jan. 2007.

[17] Y. Yang, V. Stankovic, W. Zhao, and Z. Xiong, "Multiterminal video coding," presented at the Proceedings of IEEE Int. Conf. Image Processing, Sep. 2007.

[18] R. M. Figueras i Ventura, P. Vandergheynst, and P. Frossard, "Low rate and flexible image coding with redundant representations," *IEEE Trans. Image Process.*, vol. 15, no. 3, pp. 726–739, Mar. 2006.

[19] R. Neff and A. Zakhor, "Very low bit-rate video coding based on matching pursuits," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 1, pp. 158–171, Feb. 1997.

[20] A. Rahmoune, P. Vandergheynst, and P. Frossard, "Flexible motion-adaptive video coding with redundant expansions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 178–190, Feb. 2006.

[21] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, Dec. 1997.

[22] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[23] T. E. Boult, X. Gao, R. Micheals, and M. Eckmann, "Omni-directional visual surveillance," *Image Vis. Comput.*, vol. 22, no. 7, pp. 515–534, Jul. 2004.

[24] Y. Yagi, Y. Nishizawa, and M. Yachida, "Map-based navigation for a mobile robot with omnidirectional image sensor COPIS," *IEEE Trans. Robot. Autom.*, vol. 11, no. 5, pp. 634–648, Oct. 1995.

[25] C. Geyer and K. Daniilidis, "Catadioptric projective geometry," *Int. J. Comput. Vis.*, vol. 45, no. 3, pp. 223–243, Dec. 2001.

[26] C. Geyer and K. Daniilidis, "Paracatadioptric camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 5, pp. 687–695, May 2002.

[27] Y. Ma, S. Soatto, J. Košeckà, and S. S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*. New York: Springer, 2004.

[28] A. Torii, A. Imiya, and N. Ohnishi, "Two- and three- view geometry for spherical cameras," presented at the OMNIVIS, Oct. 2005.

[29] J. Konrad and Z.-D. Lan, "Dense disparity estimation from feature correspondences," presented at the SPIE Symp. Electronic Imaging, Jan. 2000.

[30] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[31] J.-P. Antoine and P. Vandergheynst, "Wavelets on the 2-sphere: A group theoretical approach," *Appl. Comput. Harmon. Anal.*, vol. 7, no. 3, pp. 1–30, Nov. 1999.

[32] I. Tosic, I. Bogdanova, P. Frossard, and P. Vandergheynst, "Multiresolution motion estimation for omnidirectional images," presented at the EUSIPCO, Sep. 2005.

[33] P. Frossard, P. Vandergheynst, R. M. Figueras i Ventura, and M. Kunt, "*A posteriori* quantization of progressive matching pursuit streams," *IEEE Trans. Signal Process.*, vol. 52, no. 2, pp. 525–535, Feb. 2004.

Ivana Tosic (S'04) received the Dipl.Ing. degree in telecommunications from the University of Nis, Serbia, in 2003, and graduated from the Doctoral School in Information and Communication Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 2004. She is currently pursuing the Ph.D. degree at the Signal Processing Laboratory, EPFL.

She joined the Signal Processing Laboratory, EPFL, in 2004, as a Research and Teaching Assistant. Her research interests include representation and coding of visual information, 3-D objects compression, distributed source coding, and plenoptic sampling.

Pascal Frossard (S'96–M'01–SM'04) received the M.S. and Ph.D. degrees in electrical engineering from the Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1997 and 2000, respectively.

Between 2001 and 2003, he was a member of the research staff at the IBM T. J. Watson Research Center, Yorktown Heights, NY, where he worked on media coding and streaming technologies. Since 2003, he has been an Assistant Professor at EPFL, where he heads the Signal Processing Laboratory (LTS4). His research interests include image representation and coding, nonlinear representations, visual information analysis, joint source and channel coding, multimedia communications, and multimedia content distribution.

Dr. Frossard was the General Chair of IEEE ICME 2002 and Packet Video 2007 and a member of the organizing or technical program committees of numerous conferences. He is an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA (since 2004) and the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (since 2006). He is an elected member of the IEEE Image and Multidimensional Signal Processing Technical Committee (since 2007), the IEEE Visual Signal Processing and Communications Technical Committee (since 2006), and the IEEE Multimedia Systems and Applications Technical Committee (since 2005). He has served as Vice Chair of the IEEE Multimedia Communications Technical Committee (2004–2006) and as a member of the IEEE Multimedia Signal Processing Technical Committee (2004–2007). He received the Swiss NSF Professorship Award in 2003 and the IBM Faculty Award in 2005.