COARSE SCENE GEOMETRY ESTIMATION FROM SPARSE APPROXIMATIONS OF MULTI-VIEW OMNIDIRECTIONAL IMAGES

Ivana Tosic and Pascal Frossard

Signal Processing Institute ITS, Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland
phone: + (41) 21 693 4712, fax: + (41) 21 693 7600
email: {ivana.tosic,pascal.frossard}@epfl.ch, web: http://lts4www.epfl.ch

ABSTRACT

This paper presents a framework for coarse scene geometry estimation, based on sparse representations of omnidirectional images with geometrical basis functions. We introduce a correlation model that relates sparse components in different views with local geometrical transforms, under epipolar constraints. By combining selected pairs of features represented by sparse components, we estimate the disparity map between images, evaluate coarse depth information, and recover the relative camera pose. The proposed framework allows to estimate the geometry of the scene, hence disparity between images, using only coarse approximations of multi-view images. The experimental results demonstrate that only a few components are sufficient to estimate the disparity map and the camera pose. This is certainly beneficial for predictive multi-view compression schemes, where the scene reconstruction relies on the disparity mapping from low-resolution images in order to progressively decode the higher image resolutions.

1. INTRODUCTION

One of the main objectives of 3DTV is to compactly represent a 3D scene with a set of multi-view images and transmit it in an efficient way to the final user, offering thus a real three-dimensional perception of the received information. The 3D experience generally relies on depth estimation algorithms, which identify and match series of features points in multi-view images in order to offer a pleasant perception of depth [1]. Depth estimation in 3DTV generally requires decoding of high-resolution multi-view images, where dense depth fields can be estimated by hybrid recursive matching algorithms for example, as proposed in [2]. 3D rendering of scene information is typically performed independently from coding and transmission, while design problems in 3DTV framework should rather be approached from an inter-disciplinary perspective. In particular, image-based rendering and scene geometry estimation should be considered jointly with image coding for increased overall performances.

We propose a flexible framework for the representation of visual information, which allows for coarse estimation of scene geometry and opens interesting perspectives for efficient coding of multi-view images. We introduce a correlation model between omnidirectional images that captures local geometrical transformations in the 3D scene. Omnidirectional images are particularly interesting for 3D scene representation, since they can be appropriately mapped to spherical images which capture the light field in the radial form [3]. It allows to avoid discrepancies that exist in planar images, and leads to effective algorithms for camera pose estimation [4] or depth estimation from multiple views.

In order to provide a rich representation of the visual information, we propose to approximate multi-view omnidirectional images by sparse expansions with geometrical basis functions. Although finding the sparsest representation of a signal is still an ongoing research problem, there exist some methods such as Matching Pursuit [5], which offer a suboptimal solution with a tractable complexity. One of the important characteristics of the Matching Pursuit is that it captures the most prominent signal features in only few iterations. Hence it provides very good image compression performance at low rates, while producing a progressive and scalable representation. In the same time, the components present in the sparse signal representation carry important information for the estimation of the scene geometry. In particular, we show that correspondences can be identified between geometrical features in different images, under epipolar geometry constraints. Such correspondences allow to construct a disparity map between images and to estimate coarse depth information. Interestingly, since the main scene features are captured with only few basis vectors, the disparity map can be obtained from very coarse descriptions of the multi-view images. Moreover, the proposed correlation model inherently matches the corresponding features between views giving enough information for the recovery of the relative camera pose. The experimental results show that very coarse image approximations can lead to quite precise disparity maps between multiple views, which can be further efficiently exploited for the depth estimation, camera pose estimation and compression of the light field. Such benefits are quite interesting in the design of efficient multi-view coding strategies, and the novel framework presented in this paper certainly opens interesting perspectives for the joint design of coding and rendering of 3D scenes.

The paper is structured as follows. We introduce in Section 2 the view correlation model based on sparse image approximations. In Section 3, we define the transform model for the case of omnidirectional images, while in Section 4, we explain how the transform model is used for disparity map and camera pose estimation. Experimental results are presented in the Section 5.

2. CORRELATION IN SPARSE DECOMPOSITIONS

Modeling the correlation between multi-view images in camera networks is one of the most important problems in both multi-view rendering and coding. A good correlation model should relate corresponding features in different views and hence lead to an efficient scene geometry estimation. In camera sensor networks, the correlation between images is mainly driven by the 3D motion of the objects in the scene, which results in local changes of image components that represent these objects. If we decompose each image into sparse components that capture the objects in the scene, we can assume with high probability that the most prominent components are present in all images, possibly with some local transformations. Since these local transforms capture the motion due to changes of viewpoint, they carry the disparity information that can be advantageously used for depth estimation. We therefore propose to model the correlation between views by local geometrical transformations, which are estimated by pairing components in sparse image decompositions.

More formally, given a certain basis, or a possibly redundant dictionary of atoms $\mathscr{D} = \{\phi_k\}, k = 1,...,N$, in the Hilbert space H,

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

every image y can be represented as:

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

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 representation one has to search for a sparse vector x. We say that y has a *sparse* representation in $\mathscr{D}$ if the number of non-zero components in x is much smaller than the dimension of x. Therefore, the sparse representation of y is:

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

where c is the vector of non-zero elements of x, I labels the set of atoms $\{\phi_k\}_{k \in I}$ participating to the representation, and $\mathbf{\Phi_I}$ is a submatrix of $\mathbf{\Phi}$ with respect to I.

In the case of two correlated images $y_1 = \mathbf{\Phi_{I_1}} c_1$ and $y_2 = \mathbf{\Phi_{I_2}} c_2$, 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 3D features in the scene. We assume that these atoms are correlated, possibly under some local geometric transformation. Denote $F(\phi)$ the transform of an atom in the image decomposition that results from the motion of an object in the 3D space, or equivalently, the transformation imposed to atom ϕ in different views 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, and remaining atoms in I_2 :

$$y_2 = \sum_{i \in J_1} x_{2,i} F_i(\phi_i) + \sum_{k \in I_2 \setminus J_2} x_{2,k} \phi_k. \quad (3)$$

This work addresses the disparity map estimation and camera pose recovery inferred from the proposed correlation model, but this model could also be used for joint or distributed coding of multi-view images. The depth estimation and encoding process are tightly interlaced since geometry is key to effective representation of 3D scenes.

3. TRANSFORMS IN OMNIDIRECTIONAL MULTI-VIEW IMAGES

In this section we apply the generalized correlation model introduced in section 2 to the case of omnidirectional camera network. More precisely, we define the local transform model among sparse components in omnidirectional images that are appropriately mapped to spherical images.

Motions of objects in the 3D space introduce various types of transforms 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, transform 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. In particular, we use a dictionary of atoms on the 2-D unit sphere. Given a generating function g defined in the space of square-integrable functions on a unit two-sphere S^2, $g(\theta,\varphi) \in L^2(S^2)$, the dictionary $\mathscr{D} = \{\phi_k\} = \{g_\gamma\}_{\gamma \in \Gamma}$ is constructed by changing the atom indexes $\gamma = (\theta, \varphi, \psi, \alpha, \beta) \in \Gamma$, i.e., by applying a unitary operator $U(\gamma)$: $g_\gamma = U(\gamma) g$. The triplet (θ, φ, ψ) represents Euler angles that respectively describe the motion of the atom on the sphere by angles θ and φ, and the rotation of the atom around its axis with an angle ψ, and α, β represent anisotropic scaling factors. In such a structured dictionary, the transform of one atom to another interestingly reduces to a transform of its parameters γ, i.e., $g_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma') g_{\gamma_i} = U(\gamma' \circ \gamma_i) g$. Note finally that the size and redundancy of the dictionary is directly driven by the number of distinct atom transformations.

In order to relate the corresponding features in different camera views, we need to find the local transforms between sparse components that describe these features. In other words, we are interested in finding correspondences between atoms that respectively represent the images y_1 and y_2, generated by two spherical cameras that capture the same scene. Consider an atom $g_{\gamma_i}, \gamma_i \in J_1$ from the sparse decomposition of image y_1. The subset of transforms $V_i^0 = \{\gamma' | g_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma') g_{\gamma_i}\}$ allows to relate g_{γ_i} to the atoms g_{γ_j} in the expansion of y_2. However, not all these transforms are feasible under epipolar constraints. These constraints represent one of the fundamental relations in multi-view analysis, and define the relation between 3D point projections ($\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{R}^3$) on two cameras, as:

$$\mathbf{z}_2^T \hat{T} R \mathbf{z}_1 = 0, \quad (4)$$

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 is therefore reduced to the transforms that respect epipolar constraints between the atom g_{γ_i} in y_1 and the candidates atoms g_{γ_j} in y_2. We evaluate the constraints given in Eq. (4) on atom centers denoted $m_l = [sin\theta_l cos\varphi_l \;\; sin\theta_l sin\varphi_l \;\; cos\theta_l]^T$ with $l \in \{i, j\}$, and define the set $V_i^E \subseteq V_i^0$ of possible transforms of atom g_{γ_i} as:

$$V_i^E = \{\gamma' | g_{\gamma_j} = U(\gamma') g_{\gamma_i}, m_j^T \hat{T} R m_i = 0\}. \quad (5)$$

Equivalently, the set of atoms g_{γ_j} in y_2 that are possible transformed versions of the atom g_{γ_i} is denoted as the *epipolar candidates set*. It is defined by the set of atoms indexes Γ_i^E, with

$$\Gamma_i^E = \{\gamma_j | g_{\gamma_j} = U(\gamma') g_{\gamma_i}, \gamma' \in V_i^E\} \subset \Gamma. \quad (6)$$

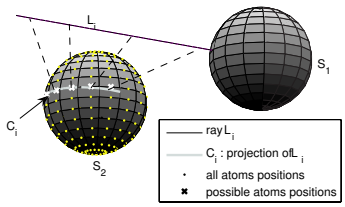


Figure 1: Selection of positions of atoms that satisfy the epipolar constraint.

A graphical interpretation of the epipolar constraint for spherical images is shown on the Figure 1, where we denote as S_1 and S_2 the two unit spheres corresponding to camera projection surfaces. The center m_j of the atom g_{γ_j}, lies on the part of a great circle $\mathscr{C}_i$ 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 center of atom g_{γ_i} on the sphere S_1.

Recall that corresponding atoms represent the same object in the 3D scene. Hence, we assume that the 3D motion of an object results in a limited difference between shapes of corresponding atoms, and we further restrict the set of possible transforms by constraints on the similarity of candidate atoms. We measure the similarity or coherence of atoms by the inner product $\mu(i,j) = |\langle g_{\gamma_i}, g_{\gamma_j} \rangle|$, 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:

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

and a set of candidate atoms in y_2, denoted the *shape candidates set*, whose indexes belong to $\Gamma_i^\mu \subset \Gamma$, with:

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

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 parameters of the transformed atom in y_2 as $\Gamma_i = \Gamma_i^E \cap \Gamma_i^\mu$. Therefore, given an atom g_{γ_i} in y_1, finding its corresponding atom in y_2 reduces to a search for a g_{γ_j} such that: $\gamma_j \in I_2$ and $\gamma_j \in \Gamma_i$. Since the image decomposition of y_2 is sparse, the coherence between two distinct atoms in I_2 is very small, whereas the coherence between two atoms in Γ_i is high due to its definition. Hence, if the corresponding feature is present in y_2, the intersection of the two constraint sets $\gamma_j \in I_2$ and $\gamma_j \in \Gamma_i$ contains only the correct corresponding atom with very high probability. In the case where the correspondence does not exist, this intersection is empty. Correspondences allow to define a set of pairs of atoms from two views y_1 and y_2, which can be exploited for estimating the geometry of the scene, as explained in the next section.

4. SCENE GEOMETRY ESTIMATION FROM ATOM TRANSFORMS

In comparison to feature points matching, the pairing of atoms with local transforms also offers a geometrical description of the local neighborhood of the observed features. Therefore, pairing two atoms in different views provides an estimation of the disparity map in the local neighborhoods where the corresponding atoms have high energy.

Disparity map relates a point $\mathbf{z}_1$ on the image y_1 with a point $\mathbf{z}_2$ on y_2, such that the epipolar constraint from Eq. (4) is satisfied. Consider now a pair of corresponding atoms $(g_{\gamma_i}, g_{\gamma_j})$ in two images. We want to find a mapping of each point on g_{γ_i} to its corresponding point on g_{γ_j}. When g_{γ_i} is defined in the discrete space, i.e., on the spherical grid $\mathscr{G}_1$, disparity mapping translates to the grid distortion induced by the local transform between g_{γ_i} and g_{γ_j}. Let us denote with P_1 the matrix of size $q \times 3$ that consists of Euclidean coordinates of q points on $\mathscr{G}_1$. Let further $\gamma_i = (\theta_i, \varphi_i, \psi_i, \alpha_i, \beta_i)$ and $\gamma_j = (\theta_j, \varphi_j, \psi_j, \alpha_j, \beta_j)$. The transform of the grid $\mathscr{G}_1$, given with P_1, to a grid $\mathscr{G}_2$, given with a matrix P_2, includes two transforms:

1. transform of rotation of the atom g_{γ_i}, given by Euler angles $(\theta_i, \varphi_i, \psi_i)$, into the rotation of the atom g_{γ_j}, given by Euler angles $(\theta_j, \varphi_j, \psi_j)$
2. transform of anisotropic scaling of the atom g_{γ_i}, given by the pair of scales (α_1, β_1), into the anisotropic scaling of the atom g_{γ_j}, given by the pair of scales (α_2, β_2).

By combining these two transforms, P_2 can be evaluated as:

$$P_2 = R_{\gamma_j}^{-1} \cdot u(R_{\gamma_i} \cdot P_1), \quad (9)$$

where R_{γ_i} and R_{γ_j} are rotation matrices given by Euler angles $(\theta_i, \varphi_i, \psi_i)$ and $(\theta_j, \varphi_j, \psi_j)$, 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 $\mathscr{G}_1$ 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 g_{γ_j}.

The stereographic projection [6] 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: $x + jy = \rho e^{j\varphi} = 2tan(\theta/2)e^{j\varphi}$. Under this projection, the transform of the point (θ, φ) on the grid $\mathscr{G}_1$ to the point (θ', φ') on the grid $\mathscr{G}_2$ due to anisotropic scaling can be obtained by scaling the stereographic projection of (θ, φ) with $1/\alpha_2$ and $1/\beta_2$, in the following way:

$$\begin{aligned} x' &= \rho' \cos\varphi' = \frac{1}{\alpha_2} x = \frac{\alpha_1}{\alpha_2} \rho \cos\varphi \\ y' &= \rho' \sin\varphi' = \frac{1}{\beta_2} y = \frac{\beta_1}{\beta_2} \rho \sin\varphi, \end{aligned} \quad (10)$$

where $\rho' = 2\tan\theta'/2$ and $\rho = 2\tan\theta/2$. By solving the system (10) for θ' and φ', we get:

$$\varphi' = u_p(\varphi) = \arctan\left(\frac{\alpha_2 \beta_1 \sin\varphi}{\alpha_1 \beta_2 \cos\varphi}\right) \quad (11)$$

$$\theta' = u_t(\theta, \varphi, \varphi') = 2\arctan\left[\tan\frac{\theta}{2}\sqrt{\frac{\alpha_1^2\cos^2\varphi + \beta_1^2\sin^2\varphi}{\alpha_2^2\cos^2\varphi' + \beta_2^2\sin^2\varphi'}}\right] \quad (12)$$

Finally, we can define the function $u(\cdot)$ as a pair of transforms $u_p(\varphi)$ and $u_t(\theta, \varphi, u_p(\varphi))$ followed by the transform of spherical coordinates (θ', φ') to Euclidean coordinates.

Each row in the matrix P_2 therefore results in a point $\mathbf{z}_2$ that is a disparity mapping of the point $\mathbf{z}_1$ given by the corresponding row in P_1. Finally, the disparity maps obtained from the correspondences in all views are combined into a single disparity map by selecting the most confident mapping for each point $\mathbf{z}_2$ from different mappings $\mathbf{z}_1^{(i)}, i = 1, ..., n$, defined by n correspondences. The final mapping point $\mathbf{z}_1^*$ is selected as:

$$\mathbf{z}_1^* = \arg \max_{\mathbf{z}_1^{(i)}, i=\overline{1,n}} w_{\gamma_i}(\mathbf{z}_1^{(i)}), \quad (13)$$

where the confidence w_{γ_i} is a normalized weighting function that prioritizes the points with highest response of the correspondence g_{γ_i}. The goal of this function is to put more importance onto the disparity mapping 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_{γ_i}.

The disparity map can be further used for recovering the relative pose between cameras and for depth estimation.

4.1 Camera pose recovery

Finding the corresponding atoms in two multi-view images opens the door to an efficient and robust method for estimation of the relative pose between two cameras, denoted as R, T for the rotation and translation. Relying only on the shape similarity constraint between corresponding atoms, we can find for the atom $g_{\gamma_i} \in I_1$ its corresponding atom g_{γ_j} such that $\gamma_j \in I_2$ and $\gamma_j \in \Gamma_i^\mu$. Moreover, since the coefficients of corresponding atoms are correlated, they also help in the pairing process. The pairs of atoms can be used to estimate the disparity map and identify the matching feature points on the whole atoms. Each pair of atoms therefore defines a set of feature points that describe the constraints given by the local transform. Using the eight point algorithm [7] we can recover the camera pose R, T from a set of corresponding feature points $(\mathbf{z}_1^k, \mathbf{z}_2^k), k = 1, ..., n (n \geq 8)$ obtained by gathering feature points from all pairs of atoms. The eight point algorithm consists of three basic steps:

1. **Find an approximation of the essential matrix** $E = \hat{T}R$
   Compute the singular value decomposition of the matrix $\chi = [\mathbf{a}^1, \mathbf{a}^2, ..., \mathbf{a}^n]^T \in \mathbb{R}^9$, where $\mathbf{a}^k = \mathbf{z}_1^k \otimes \mathbf{z}_2^k$ ($\otimes$ denotes the Kronecker product). From the singular value decomposition $\chi = U_\chi \Sigma_\chi V_\chi^T$ take E^s to be the ninth column of V_χ^T, and form the approximation of the essential matrix by unstacking nine elements of E^s into a square 3×3 matrix E.

2. **Project onto the essential space**
   Compute the singular value decomposition of the approximated matrix E to be: $E = Udiag\{\sigma_1,\sigma_2,\sigma_3\}V^T$, where $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq 0$ and $U,V \in SO(3)$. In general, the matrix E is not in the essential space ($\sigma_1 \neq \sigma_2$ and $\sigma_3 \neq 0$), so the final approximation of the essential matrix is the projection of matrix E onto the normalized essential space, evaluated as $E = Udiag\{1,1,0\}V^T$.
3. **Recover R,T from E**
   R and T are extracted from the essential matrix as:

$$R = UR_Z^T(\pm\frac{\pi}{2})V^T, \hat{T} = UR_Z(\pm\frac{\pi}{2})\Sigma U^T, \qquad (14)$$

where

$$R_Z^T(\pm\frac{\pi}{2}) = \begin{pmatrix} 0 & \pm 1 & 0 \\ \mp 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$\hat{T} = \begin{pmatrix} 0 & -T(3) & T(2) \\ T(3) & 0 & -T(1) \\ -T(2) & T(1) & 0 \end{pmatrix}.$$

This algorithm gives four solutions for R,T, but three of them can be eliminated by imposing the positive depth constraint.

The advantage of matching the whole atoms is that we have a much bigger number of epipolar constraints coming from the local neighborhood covered by the matching atoms. This makes the proposed atom matching method very robust to low precision feature points estimates.

4.2 Depth estimation

When the relative camera pose R,T is known, pairing atoms from two images can be done by using both the epipolar and shape constraints. In other words, for the atom $g_{\gamma_i} \in I_1$ we can find its corresponding atom g_{γ_j} such that $\gamma_j \in I_2$ and $\gamma_j \in \Gamma_i$. From atom pairs, the disparity map between P_1 and P_2 is estimated as described in 4. Therefore, the depth of the 3D point whose projections to two spherical images are given with $\mathbf{z}_1$ and $\mathbf{z}_2$, related by the disparity map, can be evaluated as:

$$\rho = \frac{|T \times R\mathbf{z}_2|}{|\mathbf{z}_1 \times R\mathbf{z}_2|}, \qquad (15)$$

where $\times$ denotes the cross product. When R,T is given as a relative pose of camera 2 with respect to camera 1, the depth ρ is evaluated also with respect to camera 1. By computing ρ for each pair of points in P_1 and P_2 we get a depth field for the observed atom correspondence pair.

5. EXPERIMENTAL RESULTS

The performance of the proposed scene geometry estimation method is evaluated on a simple synthetic scene and on the more complex natural scene captured with omnidirectional cameras. The sparse image decomposition has been obtained using the Matching Pursuit (MP) algorithm [5] on the sphere with two dictionaries, based on generating functions that respectively consist in a 2D Gaussian function:

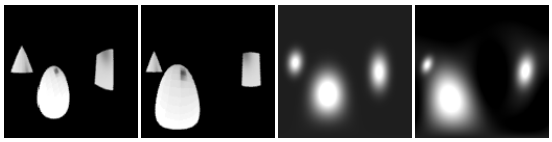
$$g_G(\theta,\varphi) = \exp\left(-\tan^2\frac{\theta}{2}\left(\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi\right)\right), \qquad (16)$$

and a 2D function built on a Gaussian and the second derivative of a 2D Gaussian in the orthogonal direction (i.e., edge-like atoms):

$$g_{EL}(\theta,\varphi) = -\frac{1}{K_2}\left(16\alpha^2\tan^2\frac{\theta}{2}\cos^2\varphi - 2\right) \cdot \exp\left(-\tan^2\frac{\theta}{2}\left(\alpha^2\cos^2\varphi + \beta^2\sin^2\varphi\right)\right). \qquad (17)$$

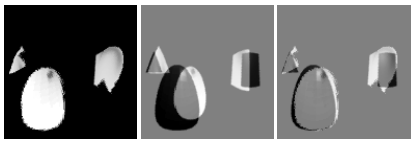
The position parameters θ and φ can take 128 different values ($N_t = N_p = 128$), while the rotation parameter uses 16 orientations, between 0 and π. The scales are distributed in a logarithmic scale from 1 to $N_t/8$ for the Gaussian atoms and from 2 to $N_p/2$ for edge-like atoms, with 3 scales per octave.

For a simple and comprehensive illustration of the proposed scene geometry estimation method based on sparse image decomposition, we have constructed a simple synthetic scene that consists of three objects: a sphere, a cube and a cone. The scene is captured with two spherical cameras, and the original images are shown on Figure 2a) and b) (unwrapped). These images have been decomposed using the MP on the sphere. Three pairs of atoms in two decompositions have been recognized as correspondences, where each of the atoms corresponds to one of the objects. Using the transforms of only these three atoms, the disparity map has been estimated and applied to the image y_1 to reconstruct an approximation of the image y_2, as shown on the Figure 3a). Disparity maps from different atom pairs have been combined by taking a weighted average of the map at each pixel, where the weights have been assigned proportionally to the atom value taken on that particular pixel. This simple example shows that we can obtain a very good estimate of the disparity between two images of the same scene, in using only very coarse approximations of the original images (shown in Figure 2c) and d)). It makes this method suitable in the context of low bitrate compression of light fields. Figure 3b) displays the difference between the two original images, while the Figure 3c) shows the difference between the original and reconstructed image y_2. The range of values for both images is [-1,1]. We can observe that many high valued errors (white and black pixels) are removed or reduced by applying the disparity map as shown in Figure 3a). Therefore, the disparity map efficiently captures the transforms of the three objects.



(a) **(b)** **(c)** **(d)**

Figure 2: Original synthetic images: a) y_1, b) y_2. Approximated synthetic images: c) $\tilde{y}_1$, with 3 atoms, d) $\tilde{y}_2$ with 6 atoms.



(a) **(b)** **(c)**

Figure 3: a) Estimation of y_2 from y_1, using the disparity map: $\hat{y}_2$; b) $y_1 - y_2$; c) $\hat{y}_2 - y_2$

In order to show that we can also extract object depths for the observed scene only from the three obtained transforms, we have evaluated the mean distance to each of the objects to camera 1 based on the obtained disparity map, when the relative camera pose R,T is known. The results are shown in the Table 1, where the estimated distances are compared with the original distances of objects measured at pixels on the spherical image that correspond to centers of the three MP atoms. We can see that the depths are well estimated using only the obtained coarse image descriptions.

The natural scene images shown on the Figure 4 have been decomposed with the MP algorithm. The omnidirectional cameras have been placed in a line (no rotation, same altitude) where successive images are denoted y_1, y_3 and y_2. The proposed method is used to recover camera pose from edge camera images y_1 and y_2. The threshold on the coherence is set to $s = 0.4$ to capture as many correspondences as possible and thus get more accurate results. For each atom, the correspondences are sought in the neighborhood of

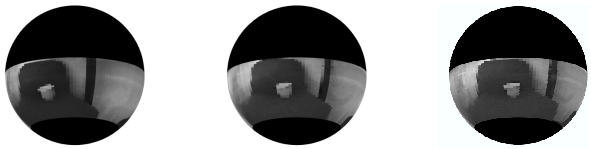
Table 1: Comparison of estimated depths with the ground truth for the simple scene

Object	Ground truth distance	Mean measured distance
Sphere	1.0	1.54
Cube	2.0	2.29
Cone	1.7	1.82

the solid angle $\pi/4$. The camera pose R,T is recovered from 18 correspondences, as described in the Sec. 4. Due to the camera placement in a line, the rotation matrix between them should be equal to the identity matrix and the translation vector to $[1 \;\; 0 \;\; 0]$, in the ideal case. With the proposed method, we obtain the following result for the rotation matrix:

$$R = \begin{pmatrix} 1.0000 & 0.0000 & -0.0015 \\ -0.0002 & 0.9914 & -0.1307 \\ 0.0015 & 0.1307 & 0.9914 \end{pmatrix}$$

and translation vector: $T = [0.9909 \;\; -0.1322 \;\; -0.0252]$. Therefore, R and T have been correctly recovered with a high precision.



y_1 y_3 y_2

Figure 4: Natural images

Figure 5 represents the unwrapped original natural images. Figure 6 shows the reconstructed image y_3 obtained by pixel mapping from images y_1 and y_2, using the combined disparity map obtained from y_1 and y_2 (as explained in the Sec. 4). For the disparity mapping, we have used $s = 0.9$ to have a consistent and smooth disparity map. The differences between the images y_3 and reference images y_1 and y_2 are shown on the Figure 7 a) and b) respectively, with the range of [0, 1], where 1 (white) means no error. Figure 7c) illustrates the difference between the original image y_3 and its reconstruction with the proposed disparity mapping. For a quantitative comparison of the three residual images, we evaluated their energies. The first residual $y_1 - y_3$ carries the highest energy of 89.2, followed by the $y_2 - y_3$ with the energy 77, while the residue after disparity mapping with the proposed scheme resulted in the lowest energy of 49.7, confirming the benefits of our method. Once again, we can see that the obtained disparity map succeeds to compensate the movements and transforms of objects in the scene, which can be advantageously used for the depth estimation and the predictive coding within light-field streaming applications.

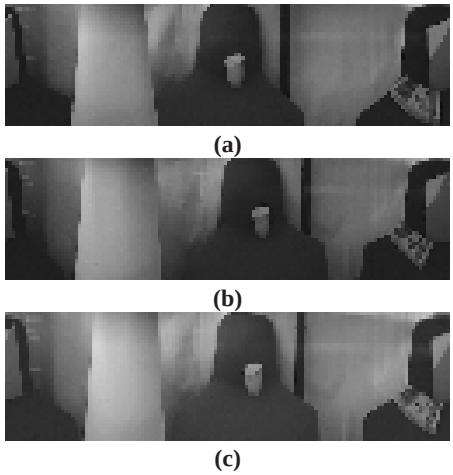
6. CONCLUSIONS

We have presented a method for estimating the disparity map and camera pose from very coarse multi-view image descriptions. This method is based on a novel correlation model between sparse approximations of multiple views. We show that the use of structured dictionaries for sparse decompositions enables pairing corresponding features among views, under local transform constraints. Experimental results demonstrate that this method leads to good estimation of the disparity map and the camera pose.

REFERENCES

[1] J.-H. Kim and T. Sikora, "Robust Anisotropic Disparity Estimation with Perceptual Maximum Variation Modeling", Proc. of IEEE ICIP 2006, October 2006.

[2] O. Schreer, C. Fehn, N. Atzpadin, M. Müller, A. Smolic, R. Tanger and P. Kauff, "A Flexible 3D TV System for Different Multi-Baseline Geometries", Proc. of IEEE ICME 2006, July 2006.



(a) **(b)** **(c)**

Figure 5: Unwrapped natural images (128x128): a) y_1; b) y_2; c) y_3

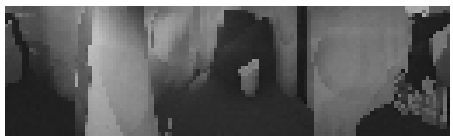
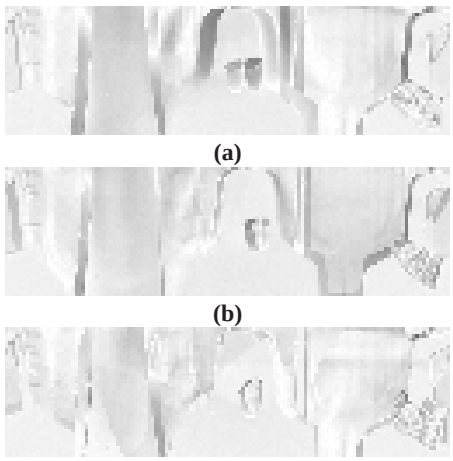


Figure 6: $\hat{y}_3$: reconstructed y_3 using the disparity map from y_1 and y_2.



(a) **(b)** **(c)**

Figure 7: Differences between: a) y_1 and y_3, energy: 89.2; b) y_2 and y_3, energy: 77.0; c) $\hat{y}_3$ and y_3, energy: 49.7.

[3] C. Geyer and K. Daniilidis, "Catadioptric projective geometry", International Journal of Computer Vision, vol. 45(3), pp. 223-243, December 2001.

[4] A. Makadia and K. Daniilidis, "Rotation Recovery from Spherical Images without Correspondences", IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 28(7), pp. 1170- 1175 July 2006.

[5] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries", IEEE Trans. on Signal Processing, vol. 41(12), pp. 3397-3415, December 1993.

[6] J. Antoine and P. Vandergheynst, "Wavelets on the 2-sphere: a group theoretical approach", Applied and Computational Harmonic Analysis, vol. 7(3), pp. 1-30, November 1999.

[7] Y. Ma, S. Soatto, J. Košecka and S.S. Sastry, "An Invitation to 3-D Vision: From Images to Geometric Models", chapter 5, pages 117-130, Springer, 2004.