MVMP: MULTI-VIEW MATCHING PURSUIT WITH GEOMETRY CONSTRAINTS

Ivana Tošić[‡], Antonio Ortega[†] and Pascal Frossard[*]

[‡]Redwood Center for Theoretical Neuroscience, University of California at Berkeley, USA

[†] Department of Electrical Engineering, Signal and Image Processing Institute, University of Southern California, USA

[*] Signal Processing Laboratory (LTS4), Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Sets of multi-view images that capture plenoptic information from different viewpoints are typically related by geometric constraints. The proper analysis of these constraints is key to the definition of consistent compact representations of such images. We propose an algorithm for joint sparse approximation of multi-view images driven by epipolar geometry considerations. We extend greedy pursuit algorithms, such that the representation of multi-view images into linear combination of geometric atoms is able to balance approximation error and geometric consistency. We further add a rate penalty constraint that favors representations with small entropy towards efficient coding applications. Experimental results illustrate the trade-off between approximation, geometry and rate constraints in the representation of stereo omnidirectional images. In particular, we show that geometry constraints lead to a consistent description of the correlation among views, which is particularly beneficial for scene analysis or view interpolation applications. At the same time, we show that the rate constraint leads to compact representations, possibly to the detriment of geometry consistency.

1. INTRODUCTION

Multi-view images obtained by camera networks represent a realistic way of conveying 3D information, which is implicitly contained in the geometric relations between correlated features in different views. Because they capture the same scene, multi-view images contain a lot of redundancy that should be exploited for building compact representations. Lossy multi-view compression should rely on the implicit geometry constraints and exploit the correlation between corresponding features. Since the underlying scene geometry drives the redundancy within multiple views, geometry-based correlation models lead to the construction of consistent representations and efficient coding. Many 3D applications require the interpolation of intermediate views, which relies on this geometric information. In this case, preserving the geometric relations and finding true disparity is crucial, even if it comes at the cost of worse RD performance.

The most commonly used multi-view compression method is the multi-view video coding standard (MVC) [1]. This method finds correlated blocks in two views under geometric constraints. However, block-based matching limits the possible transforms between features in multiple views to simple translations. Although translations are usually sufficient to model temporal correlation, they are very restrictive in the multi-view case and limit the cameras to inline and parallel arrangements. Other coding solutions based on dense depth map or visual correspondences matching are often limited in capturing the true 3D geometry in arbitrary camera arrangements [2], or face the difficult problem of rate allocation between texture and disparity or depth information [3].

We present a new method to compute sparse approximations of pairs of stereo images, called Multi-view Matching Pursuit (MVMP). The proposed method exploits the multi-view correlation by finding a set of local transforms that relate pairs of geometric features selected from an overcomplete dictionary of edge-like atoms. To the best of our knowledge, MVMP is the first algorithm that approximates multi-view images by finding sparse, local transforms that satisfy the multi-view geometric constraints. The diversity of local transforms, such as translations, rotation and scaling, makes MVMP more flexible than block-based matching for multi-view coding. Another advantage of our algorithm over the existing multi-view coding schemes is that it gives a simple way to trade off approximation performance, geometric consistency, and coding rate in the image representation, and to adapt it to the target application. Experimental results show that geometric constraints in MVMP permit to increase the consistency of the geometry estimation and also reduce the total entropy of representation. On the other hand, rate constraints lead to compact representations (small entropy) but sacrifice the geometric consistency. Therefore, plenoptic geometry is certainly a very important constraint in building efficient multi-view coding schemes, where classical rate-distortion (RD) optimization fails to provide consistent descriptions of the 3D content.

2. MULTI-VIEW IMAGE MODEL

We first need to define the multi-view image model. For simplicity, we consider two images vectorized into column vectors: the left image $\mathbf{y}_L$ and the right image $\mathbf{y}_R$, but the model can be generalized to any number of images. The images $\mathbf{y}_L$ and $\mathbf{y}_R$ have m-sparse approximations in dictionary $\mathbf{\Phi}$ of size M, up to an approximation error $\mathbf{e}_L$, resp. $\mathbf{e}_R$. Hence, we have:

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

where the vectors $\mathbf{a}$ and $\mathbf{b}$ represent the coefficients for the left and right image, respectively. The index sets $\mathcal{L} = \{l_k\}, \mathcal{R} = \{r_k\}, k = 1, ..., m$ label the atoms that participate in the sparse decompositions of $\mathbf{y}_L$ and $\mathbf{y}_R$, respectively. In other words, $\{l_k\}, \{r_k\}, k = 1, ..., m$

This work has been supported by the Swiss National Science Foundation under grant PBELP2-127847.

denote the atoms with non-zero coefficients, i.e., $a_{l_k} \neq 0$ and $b_{r_k} \neq 0$. This model assumes that both stereo images are composed of m atoms, while the atoms in the left and the right image do not have to be identical ($\mathcal{L} \neq \mathcal{R}$). This is due to the fact that left and right images record the visual information from the same 3D environment and typically contain the image projections of the same number of 3D scene features recorded from different viewpoints. If the dictionary consists of localized and oriented atoms that represent well the edges and the objects geometry in general, stereo images are approximated with similar atoms, but locally transformed (shifted, rotated, etc.). Since sparse approximations in overcomplete geometric dictionaries result in atoms with a very small spatial overlap (different atoms approximate different scene features), it is reasonable to assume that all these atoms can generally undergo different local transforms[1]. Therefore, we further assume that signals $\mathbf{y}_L$ and $\mathbf{y}_R$ are correlated in the following way:

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

where $F_{l_k r_k}(\cdot)$ denotes a transform of an atom ϕ_{l_k} in $\mathbf{y}_L$ to an atom ϕ_{r_k} in $\mathbf{y}_R$. The transform differs for each $k = 1, ..., m$.

Since the parametric dictionary is built on rotation, translation and anisotropic scaling of the generating function, these are the types of atom transforms that we consider in this paper. However, these transforms are not arbitrary. As the corresponding atoms are images of the same feature in the 3D space, the atom transformations have to satisfy multi-view epipolar geometry constraints. Let two points $\mathbf{v}$ and $\mathbf{u}$ represent image projections of the same 3D point p on the left and right camera, resp. We denote the relative orientation between cameras as $\mathbf{R} \in SO(3)$ and the relative translation $\mathbf{T} \in \mathbb{R}^3$. Let further $\mathbf{v}$ lie on the atom ϕ_l and $\mathbf{u}$ lie on the atom $\phi_r = F_{lr}(\phi_l)$. In parametric dictionaries, transforming the atom ϕ_l with F_{lr} reduces to a linear transform of the coordinate system $Q_{lr}(\cdot)$, i.e., $\mathbf{u} = Q_{lr}(\mathbf{v})$. The epipolar geometry constraint is then $[Q_{lr}(\mathbf{v})]^{\mathsf{T}} \hat{\mathbf{T}} \mathbf{R} \mathbf{v} = 0$. The matrix $\hat{\mathbf{T}}$ is obtained by representing the cross product of $\mathbf{T}$ with $\mathbf{R}\mathbf{v}$ as a matrix multiplication.

The epipolar constraint is usually evaluated with a certain error in practice, so we have $d_l = [Q_{lr}(\mathbf{v})]^{\mathsf{T}} \hat{\mathbf{T}} \mathbf{R} \mathbf{v}$ where $d_l \neq 0$. Moreover, the epipolar constraint is not symmetric anymore, so $d_l \neq d_r = [Q_{lr}^{-1}(\mathbf{u})]^{\mathsf{T}} \hat{\mathbf{T}} \mathbf{R} \mathbf{v}$. The most likely transforms Q_{lr} (and F_{lr}) in pairs of stereo images are the transforms that give small epipolar errors on all points that lie on the spatial support of the atom (where atom has values $\neq 0$). We thus need to define the epipolar distance between two atoms as:

$$W_{lr} = \sum_{i=1}^{q} \left(w_l^{[i]} (d_l^{[i]})^2 + w_r^{[i]} (d_l^{[i]})^2 \right), \qquad (3)$$

where q is the number of points (pixels) for which we calculate the epipolar constraint. The role of the weights $w_l^{[i]}, w_r^{[i]}$ is to select the points that are on the spatial support of atoms. Moreover, they can be chosen such that they give more importance to the epipolar constraint for points where the epipolar estimation is more reliable.

[1]Note that the proposed model does not put an explicit assumption that neighboring atoms would undergo similar transforms, i.e., it does not perform any regularization. However, if such statistics exist in the space of transforms, we expect them to be captured by geometric consistency.

3. MVMP ALGORITHM

We describe here an algorithm for the joint sparse approximation of multi-view images, where geometric constraints can be enforced for improved consistency in image representations. Finding the best such approximation is a hard combinatorial problem, which requires testing of all possible combinations of atoms in the image decomposition. We propose to use a greedy approach based on a matching pursuit (MP) algorithm that iteratively selects atoms from the dictionary based on the constraint that the energy of the approximation error is minimized [4]. MP algorithms are suboptimal algorithms that are known to provide a good trade-off between approximation performance and complexity.

In [5] we have shown that one can arrive to a simple energy functional that trades off approximation performance $\mathcal{A}$ and geometry consistency $\mathcal{G}$, by maximizing the likelihood $P(\mathbf{y}_L, \mathbf{y}_R, D = 0 | \mathbf{\Phi})$, where $\mathbf{a}$ and $\mathbf{b}$ are hidden variables in the stereo image model in Eq. (1), and D is the total epipolar error for all points in the images. The general form of this energy can be written as $E = \mathcal{A} + \lambda \cdot \mathcal{G}$, where λ is a chosen trade-off parameter. We then define the multi-view (stereo) MP as an algorithm that selects at each iteration a pair of atoms (ϕ_l, ϕ_r) that give the minimal value of this energy. This leads to a greedy algorithm that jointly selects in different views the features that correspond to the same 3D features in space, because it incorporates the geometric constraint in the atom selection process.

3.1. Joint sparse approximation with consistent geometry

The proposed greedy algorithm chooses at each iteration k the pair of atoms ϕ_{l_k}, ϕ_{r_k} that gives the minimal value of the energy $E^{(k)}$ obtained from the ML objective in [5]. More formally, it chooses ϕ_{l_k}, ϕ_{r_k} such that:

$$\begin{aligned}(\phi_{l_k}, \phi_{r_k}) &= \arg\min_{\phi_l, \phi_r} E^{(k)} = \arg\min_{\phi_l, \phi_r} \\ &[(\|\mathbf{h}_l^{[k-1]} - \langle \mathbf{h}_l^{[k-1]}, \phi_l \rangle \phi_l\|_2^2 + \|\mathbf{h}_r^{[k-1]} - \langle \mathbf{h}_r^{[k-1]}, \phi_r \rangle \phi_r\|_2^2) \\ &+ \frac{\lambda}{z_D} W_{lr} + \frac{\lambda}{z_C} (\langle \mathbf{h}_r^{[k-1]}, \phi_r \rangle - \frac{\langle \mathbf{h}_l^{[k-1]}, \phi_l \rangle}{J_{lr}})^2], \end{aligned} \qquad (4)$$

where $\mathbf{h}_l^{[k-1]}$ and $\mathbf{h}_r^{[k-1]}$ are the residues of the left and right images respectively, after $k-1$ iterations, and J_{lr} is the Jacobian of the transform Q_{lr}. Initially, the residues are: $\mathbf{h}_L^{[0]} = \mathbf{y}_L$ and $\mathbf{h}_R^{[0]} = \mathbf{y}_R$, and they are updated at each step k as:

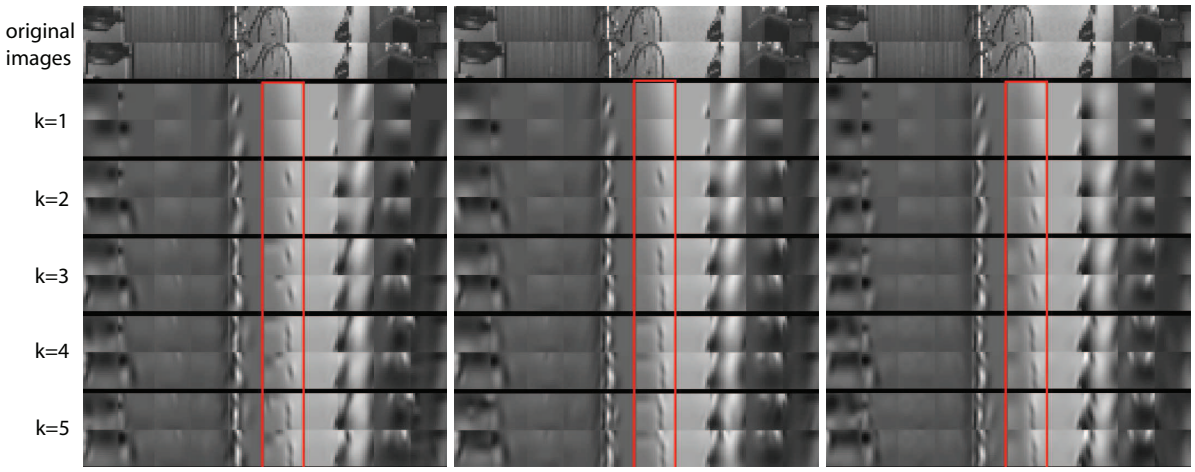
$$\begin{aligned} \mathbf{h}_L^{[k]} &= \mathbf{h}_L^{[k-1]} - \langle \mathbf{h}_L^{[k-1]}, \phi_{l_k} \rangle \phi_{l_k}, \\ \mathbf{h}_R^{[k]} &= \mathbf{h}_R^{[k-1]} - \langle \mathbf{h}_R^{[k-1]}, \phi_{r_k} \rangle \phi_{r_k}. \end{aligned} \qquad (5)$$

The coefficients a_{l_k} and b_{r_k} are simply evaluated as:

$$\begin{aligned} a_{l_k} &= \langle \mathbf{h}_L^{[k-1]}, \phi_{l_k} \rangle, \\ b_{r_k} &= \langle \mathbf{h}_R^{[k-1]}, \phi_{r_k} \rangle. \end{aligned} \qquad (6)$$

We refer to this algorithm as Multi-view Matching Pursuit (MVMP)[2]. We can immediately see that MVMP is a special case of the class of MP algorithms that we call the constrained MP. The convergence is guaranteed for this type of algorithm, even if the algorithm is prone to local minima. The convergence rate is typically smaller than the convergence rate of the MP algorithm, and the rate penalty depends on the weight λ for the additional constraint [5].

[2]Although we take here only two images, we can generalize the algorithm to more than two images by pairwise image correspondence.



(a) $\lambda = 0$**,** $\alpha = 0$ **(b)** $\lambda = 0$**,** $\alpha = 0.05$ **(c)** $\lambda = 1$**,** $\alpha = 0.05$

Fig. 1. Original and MVMP approximated images for different values of λ and α. First two rows in each set are original images. Each two successive rows represent MVMP approximations at iteration k.

3.2. Entropy-constrained approximation

The MVMP algorithm provides an approach for sparse approximation of multi-view images, but does not deal with the coding rate constraints. In coding applications, however, it is important to control the total rate of data, $\mathcal{R}$. The choice of atom pairs in MVMP certainly influences the total rate, since the parameters of atoms in a pair are correlated. The total rate to encode the parameters of both atoms would consist of the rate needed to encode parameters of the reference atom and the rate needed to encode the atom transform by differential coding. If we want the MVMP algorithm to select geometrically correlated atom pairs under a certain rate constraint, it should select at each iteration the atom pair that minimizes the energy $E = \mathcal{A} + \lambda \cdot \mathcal{G} + \alpha \cdot \mathcal{R}$. This amounts to adding a rate term $\alpha \cdot \mathcal{R}$ to the energy in Eq. (4). In this paper we do not fix an encoding scheme to find the total rate, and instead evaluate the performance of the MVMP algorithm by replacing the total rate by its entropy lower bound $H(\phi_l) + H(\phi_r|\phi_l)$ in the case of differential coding.

4. EXPERIMENTAL RESULTS

Parameters λ and α in the energy functional permit to choose different operating points of the MVMP algorithm, where each operating point gives different performance in terms of rate, distortion and geometry matching. This section provides experimental results that illustrate the performance of MVMP for different operating points. The experiments are designed to give insights into the trade-offs that occur for each point, but also to show how the geometric constraints influence the behavior of MVMP in terms of rate or entropy. In particular, we are interested to see if imposing the geometry constraint decreases the entropy of local atom transforms due to the consistency of geometric correlation.

We have used two omnidirectional images from our "Mede" database, and performed MVMP on a set of 16 blocks from each image. These blocks form two panoramas that are partially shown in the first two rows in Fig. 1 a)-c) (only 10 blocks from each panorama). Each block is of size 16×16 pixels, and thus the maximum disparity that can be captured is 16 pixels. If higher disparity is expected, one can choose a bigger block size, with the expense of increased complexity. The rotation between cameras is identity, and translation is $T = [0 \;\; 1 \;\; 0]^{\mathsf{T}}$. The dictionary is defined on the sphere, and it is built on a generating function that is a gaussian in one direction and its second derivative in the orthogonal direction [6]. We have used 16 translations in each direction, 4 orientations, and 5 pairs of anisotropic scales. For each different value of $\lambda = \{0, 1, 3, 5\}$ in MVMP, we have used a different set of anisotropic scales that are optimally learned for that particular value of λ. This way we make sure that the results are not biased by a fixed choice of the dictionary. The learning algorithm is explained in [5].

Let us first look at the approximated images with MVMP, for different values of λ and α. The pairs of rows Fig. (1) show the approximated left and right images at each iteration $k = 1, ..., 5$. Each consecutive pair of rows adds one atom per block. In Fig. (1) a) there is no geometry constraint and no rate constraint ($\lambda = \alpha = 0$), so this case reduces to classical MP. We can see that the atoms are not selected in pairs of features that correspond to the same 3D features. This is particularly evident on the bike wheel (outlined in red). Fig. (1) b) shows the approximated images when $\lambda = 0$ and $\alpha = 0.05$, i.e., there is no geometry constraint, but there is a constraint on the rate. This forces MVMP to select pairs of correlated atoms, but these pairs do not necessarily reflect the image geometry (see blocks 1 and 2 from the left). Finally, Fig. (1) c) shows the images when both geometry and rate constraints are included, $\lambda = 1$ and $\alpha = 0.05$. In this case MVMP selects correlated atoms that also reflect the image geometry.

In order to quantitatively evaluate the rate-geometry-distortion trade-offs, we have calculated the approximation error (distortion), the epipolar distance, and the total entropy that includes the entropy of atom parameters in the first (reference) image and the entropy of transforms. Transform parameters are simply calculated as a difference between atom parameters in $\mathbf{y}_L$ and $\mathbf{y}_R$. The entropy is calculated as the first order empirical entropy. We have evaluated

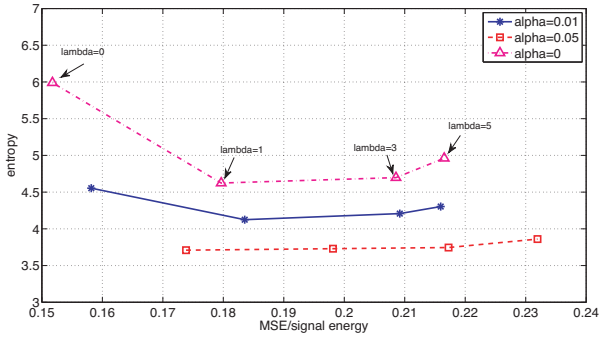


Fig. 2. MVMP performance: entropy vs distortion, $k = 8$.

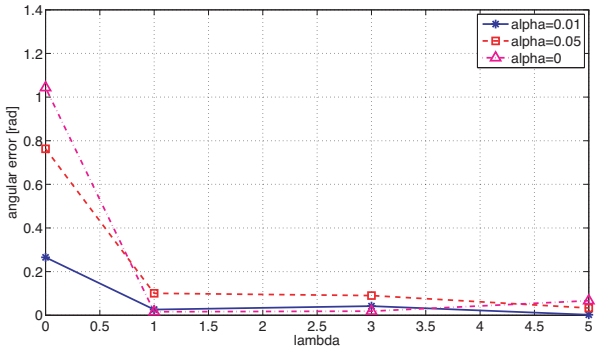


Fig. 4. Translation estimation error vs λ, $k = 8$.

these quantities for $\lambda = \{0, 1, 3, 5\}$ and $\alpha = \{0, 0.01, 0.05\}$. We first show the distortion-entropy curves in Fig. 2, where each curve corresponds to a different value of α. Each point on the curves corresponds to a different value of λ, i.e., from left to right λ changes from 0 to 5. Hence, the first points on each curve correspond to the case where no geometry constraint is included in MVMP. What we see from this graph is that increasing λ from 0 to 1 reduces the entropy even if we do not have any rate constraint ($\alpha = 0$). This confirms our intuition that the geometric constraint leads MVMP to find correlated image features. However, the distortion has to increase as a trade-off to geometry. When we increase α to impose rate constraints, the entropy for $\lambda = 0$ approaches the entropy of $\lambda = 1$ and becomes equal to it for $\alpha = 0.05$, while having a smaller distortion. In this case, having no geometry constraint is better in terms of RD performance.

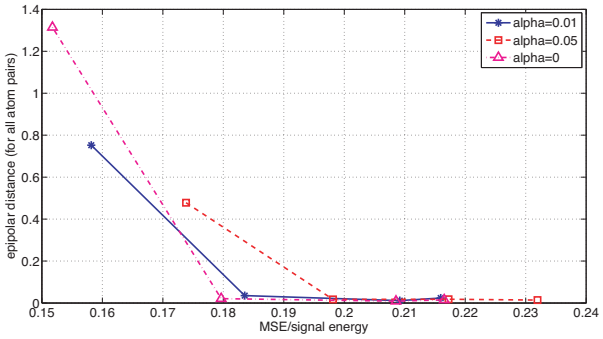


Fig. 3. MVMP performance: epipolar distance vs distortion, $k = 8$.

Independently of the value of α, cases where $\lambda = 0$ give worse epipolar matching, i.e., give higher epipolar distance than $\lambda > 0$. This is shown in Fig. 3, where we plot the distortion vs the epipolar distance between all pairs of atoms (sum of W_{lr} for all atom pairs). The curves are traced by changing the value of λ, similarly to Fig. 2. It is obvious that increasing λ from 0 to 1 decreases the epipolar distance drastically and thus improves the epipolar matching.

Finally, we want to show that increasing the weight on the geometrical part of the energy function actually contributes to a better geometric estimation from atom pairs. Therefore, we use the computed atom pairs to estimate the relative pose (rotation R and translation T) between two cameras from which the images were taken. The pose is estimated using the eight-point algorithm [7] from the coordinates of atom centers. Since the ground truth rotation between the cameras is identity, its estimation is usually good, so we do not report the errors on estimating R. The translation vector is harder to estimate, so we evaluate the estimation error in terms of the solid angle between the ground truth T and its estimate T_e. The angular error for different λ and α is shown in Fig. 4. We can clearly see that imposing geometry constrained (i.e., $\lambda > 0$) leads to multi-view image representations with meaningful geometrical information.

5. CONCLUSION

We have introduced a new multi-view image approximation algorithm MVMP, which exploits the geometric correlation of corresponding features in different views. MVMP can operate in different regimes, and the choice of a particular regime depends on the target application. If we are primarily concerned in the RD performance, without interest in geometry, then the geometric weight constraint should be dropped. On the other hand, if the target application includes any type of geometric estimation, we should impose a geometry constraint. MVMP is then able to simultaneously provide a sparse approximation of image pairs and a consistent description of the underlying geometry.

6. REFERENCES

[1] Chen Y., Wang Y., Ugur K., Hannuksela M., Lainema J. and Gabbouj M., "The Emerging MVC Standard for 3D Video Services," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–13, 2009.

[2] Yorozu N., Uematsu Y. and Saito H., "Multiple-view video coding using depth map in projective space," *Proc. of the 5th International Symposium on Advances in Visual Computing*, 2009.

[3] Magnor M., Ramanathan P. and Girod B., "Multi-view coding for image-based rendering using 3-d scene geometry," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 11, pp. 1092– 1106, 2003.

[4] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[5] Tošić I., *On unifying sparsity and geometry for image-based scene represenation*, PhD thesis, EPFL, 2009.

[6] Tošić I. and Frossard P., "Geometry-Based Distributed Scene Representation With Omnidirectional Vision Sensors," *IEEE Transactions on Image Processing*, vol. 17, no. 7, pp. 1033–1046, 2008.

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