Distributed multi-view image coding with learned dictionaries

Ivana Tošić
Ecole Polytechnique Fédérale de Lausanne
Signal Processing Laboratory (LTS4)
Lausanne, Switzerland
ivana.tosic@epfl.ch

Pascal Frossard
Ecole Polytechnique Fédérale de Lausanne
Signal Processing Laboratory (LTS4)
Lausanne, Switzerland
pascal.frossard@epfl.ch

ABSTRACT

This paper addresses the problem of distributed image coding in camera neworks. The correlation between multiple images of a scene captured from different viewpoints can be effiiciently modeled by local geometric transforms of prominent images features. Such features can be efficiently represented by sparse approximation algorithms using geometric dictionaries of various waveforms, called atoms. When the dictionaries are built on geometrical transformations of some generating functions, the features in different images can be paired with simple local geometrical transforms, such as scaling, rotation or translations. The construction of the dictionary however represents a trade-off between approximation performance that generally improves with the size of the dictionary, and cost for coding the atoms indexes. We propose a learning algorithm for the construction of dictionaries adapted to stereo omnidirectional images. The algorithm is based on a maximum likelihood solution that results in atoms adapted to both image approximation and stereo matching. We then use the learned dictionary in a Wyner-Ziv multi-view image coder built on a geometrical correlation model. The experimental results show that the learned dictionary improves the rate-distortion performance of the Wyner-Ziv coder at low bit rates compared to a baseline parametric dictionary.

Categories and Subject Descriptors

E.4 [**Data**]: Coding and Information Theory—*Data compaction and compression*

Keywords

Distributed source coding, sparse approximations, multi-view images

1. INTRODUCTION

Multi-view images are captured by a network of cameras distributed in a 3D scene. Compared to conventional 2D images, multi-view images offer a richer description of the captured scene because they convey both the texture and the 3D scene information. Camera networks have found usage in applications such as surveillance, 3D television and robotics.

One of the main challenges in multi-view imaging has been to come up with an efficient way to compress these images by exploiting the multi-view correlation, without communication between cameras. Distributed source coding (DSC) can offer here an elegant way to solve this problem. Namely, it is possible to exploit the correlation between sources without communication between encoders, as long as the decoding is performed jointly [1, 2]. Distributed multi-view coding relies on the knowledge of an appropriate multi-view correlation model, whose estimation represents a difficult and widely investigated problem. Simple block-translational models, as the ones used for video compression, are suboptimal in the multi-view case because images from different cameras are rather correlated by more diverse local transforms of objects in the scene, such as translation, scaling or rotation.

We have previously proposed a geometry-based correlation model for multi-view images that relates image features in different views by local transforms, such as translations, rotations or scaling [3]. We have proposed to capture these features by sparse image expansion with geometric atoms taken from a redundant dictionary. The correlation model is applied to the design of a DSC method with side information for multi-view omnidirectional images mapped to spherical images. The Wyner-Ziv coder is designed by partitioning the dictionary into cosets based on atom dissimilarity. The joint decoder uses the proposed correlation model to select the best candidate atom within the coset and to find corresponding features in two views. Since the disparity information can be estimated from atom transforms, the decoder transforms the reference image using the disparity in order to obtain an estimation of the Wyner-Ziv image and improve the final decoded image. The DSC scheme has been extended to handle a certain number of occlusions [4]. However, the choice of the dictionary in [3, 4] is empirical and not optimized for multi-view image representation.

This paper proposes to learn stereo dictionaries and to use them in distributed multi-view coding. Adapting the dictionary to a specific task or imposing a proper structure in the dictionary can yield significant performance improvement. We design stereo dictionaries that have the optimal properties for both image approximation and scene geometry estimation. We first propose a maximum likelihood (ML) method for learning stereo dictionaries, where the epipolar geometry is included in the probabilistic modeling. The

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7466
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7466

learned atoms present high anisotropy characteristics and substantially differ from atoms in single view learning. We then use the learned dictionaries in the Wyner-Ziv coding method proposed in [3]. The experimental results show that the learned dictionary offers better estimation of the Wyner-Ziv image. This leads to improved rate-distortion performance of the Wyner-Ziv coder at low bit rates.

We first overview the related work on multi-view DSC coding in Sec. 2. The Wyner-Ziv coder based on overcomplete geometric image representation [3] is briefly described in Sec. 3. Sec. 4 introduces the new dictionary learning method for stereo image representation. Experimental results for learning and distributed coding with the learned dictionary are presented in Sec. 5. Sec. 6 concludes the paper.

2. RELATED WORK

The application of DSC principles for multi-view coding is generally based on the disparity correlation between views, defined by the epipolar constraint. One part 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 [5]. The Wyner-Ziv images are decoded using the interpolated image obtained by disparity compensation from independent views. Wyner-Ziv coding of stereo images with unsupervised learning of the disparity between two stereo views has been proposed by Chen et al. [6]. This scheme requires a feedback channel from the decoder to the encoder to ensure enough information for disparity learning. Other solutions employ symmetric distributed source coding that equally balances the bit rate among different cameras. Gehrig and Dragotti have proposed to model the multi-view correlation by relating the locations of discontinuities in the polynomial representation of images [7]. Their scheme considers translations as correlation in multiple views, corresponding to the shifts of the discontinuities of the piecewise polynomials. A symmetric DSC scheme for coding of multi-view omnidirectional image proposed by Thirumalai et al. [8] is achieved using source partitioning. Images are transformed using a spherical Laplacian pyramid and progressively encoded with SPIHT. However, the Laplacian Pyramid is very redundant and cannot reach high compression gains.

Disparity-based solutions have been proposed also for distributed multi-view video compression. Several works build on the advantages of distributed video and multi-view coding for exploiting both temporal and inter-view correlation [9, 10, 11]. They take different approaches for modeling the correlation among views, like the disparity-based model [9], affine model [11], or homography-based model [10]. Another direction for the distributed multi-view video compression is based on classical motion compensated video encoding at each camera, while the inter-view correlation is exploited in a distributed manner [12, 13]. The study on the influence of multiple side information for distributed video coding has been presented by Maugey et al. [14]. They show that multiple side information can increase the overall performance, but at a price of higher decoding complexity.

The work presented in this paper differs from most of the previous work since it exploits a more diverse types of geometric correlation between multi-view images, such as translations, rotations and anisotropic scaling. Moreover, the disparity estimation of scene geometry can be performed using a single reference frame that has been highly compressed. Finally, a very important property of the proposed method is that it does not require a special camera arrangement in a camera network, since the geometric correlation model can cope with various local transforms. This makes the applicability of our method more generic than that of distributed coding methods designed for camera arrays.

3. GEOMETRIC WYNER-ZIV CODER

3.1 Correlation model by sparse approximations

The correlation model between multi-view images introduced in [3] relates image components that approximate the same 3D object in different views, by local transforms that include translation, rotation and anisotropic scaling. Given a redundant dictionary of atoms $\mathcal{D} = \{\phi_k\}, k = 1, ..., N$, in the Hilbert space H, we say that the image y has a *sparse* representation in $\mathcal{D}$ if it can be approximated by a linear combination of a small number of vectors from $\mathcal{D}$. Therefore, sparse approximations of two[1] multi-view images can be expressed as $y_L = \mathbf{\Phi_{I_L}} \mathbf{a} + \eta_L$ and $y_R = \mathbf{\Phi_{I_R}} \mathbf{b} + \eta_R$, where $I_{L,R}$ labels the set of atoms $\{\phi_k\}_{k \in I_{L,R}}$ participating in the sparse representation, $\mathbf{\Phi_{I_{L,R}}}$ is a matrix composed of atoms ϕ_k as columns, and $\eta_{L,R}$ represents the approximation error. Since y_L and y_R capture the same 3D scene, their sparse approximations over the sets of atoms I_L and I_R are also correlated. Our geometric correlation model makes two main assumptions in order to relate the atoms in I_L and I_R:

1. The most prominent (energetic) features in a 3D scene are present in the sparse approximations of both images, with high probability. The projections of these features in images y_L and y_R are represented as subsets of atoms indexed by $J_L \in I_L$ and $J_R \in I_R$ respectively.
2. These atoms are correlated by local geometric transforms. We denote by $F(\phi)$ the transform of an atom ϕ between two image decompositions that results from the change of viewpoint.

Under these assumptions the correlation between the images is modeled as a set of transforms F_i between corresponding atoms in sets indexed by J_L and J_R. The approximation of the image y_R can be rewritten as the sum of the contributions of transformed atoms, remaining atoms in I_R, and noise η_R:

$$y_R = \sum_{i \in J_L} b_i F_i(\phi_i) + \sum_{k \in I_R \setminus J_R} b_k \phi_k + \eta_R. \qquad (1)$$

The model from Eq. (1) is applied in [3] to atoms from the sparse decompositions of omnidirectional multi-view images mapped on the sphere. The approach is based on the use of a parametric redundant dictionary of atoms that are derived from a single waveform that undergoes rotation, translation and scaling. More formally, given a generating function g defined[2] in H, the dictionary $\mathcal{D} = \{\phi_k\} = \{g_\gamma\}_{\gamma \in \Gamma}$ is constructed by changing the atom index $\gamma \in \Gamma$ that defines rotation (ψ), translation (τ, ν) and scaling parameters (α, β) applied to the generating function g. This is equivalent to

[1]Two images are taken for the sake of clarity, but the correlation model can be generalized to any number of images.
[2]In the case of spherical images g is defined on the 2-sphere

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7466
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7466

applying a unitary operator $U(\gamma)$ to the generating function g, i.e., $g_\gamma = U(\gamma)g$. The main property of the parametric dictionary is that the transformation of an atom by a combination of translation, rotation and anisotropic scaling transforms results in another atom in the same dictionary. 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_L and y_R. When the parametric dictionary is used for both images, the transform of the atom g_{γ_i} in image y_L to the atom h_{γ_j} in image y_R reduces to a transform of its parameters, i.e., $h_{\gamma_j} = F(g_{\gamma_i}) = U(\gamma')g_{\gamma_i} = U(\gamma' \circ \gamma_i)g$. Due to the geometric constraints that exist in multi-view images, only a subset of all local transforms between $\{g_\gamma\}$ and $\{h_\gamma\}$ are feasible. This subset can be defined by identifying two constraints between corresponding atoms, namely *shape similarity* constraint and *epipolar* constraint.

First, we assume that the change of viewpoint on a 3D object results in a limited difference between shapes of corresponding atoms since they represent the same object in the scene. 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 atoms h_{γ_j} in y_R 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

$$\Gamma_i^\mu = \{\gamma_j | h_{\gamma_j} = U(\gamma')g_{\gamma_i}, \mu(i,j) > s\}. \quad (2)$$

Second, pairs of atoms that correspond to the same 3D points have to satisfy epipolar geometry constraints, which represent one of the fundamental relations in multi-view analysis. The decision on the epipolar matching of two corresponding atoms is taken when their epipolar distance $d_{EA}(g_{\gamma_i}, h_{\gamma_j})$ is smaller than a certain threshold κ (for more details on the epipolar atom distance we refer the reader to [3]). The set of possible candidate atoms in y_R, that respect epipolar constraints with the atom g_{γ_i} in y_L, called the *epipolar candidates set*, is then defined as the set of indexes $\Gamma_i^E \subset \Gamma$, with:

$$\Gamma_i^E = \{\gamma_j | h_{\gamma_j} = U(\gamma')g_{\gamma_i}, d_{EA}(g_{\gamma_i}, h_{\gamma_j}) < \kappa\}. \quad (3)$$

Finally, we combine the epipolar and shape similarity constraints to define the set of possible parameters of the transformed atom in y_R as $\Gamma_i = \Gamma_i^E \cap \Gamma_i^S$.

3.2 Wyner-Ziv coding

Based on the above geometric correlation model, a Wyner-Ziv coding scheme for multi-view omnidirectional images has been proposed in [3]. For the sake of completeness, we briefly overview here this scheme, shown in Fig. 1. The Wyner-Ziv coder is based on coding with side information, where image y_L is independently encoded, while the Wyner-Ziv image y_R is encoded by coset coding of atom indexes and quantization of their respective coefficients. The approach is based on the observation that when atom h_{γ_j} in the Wyner-Ziv image y_R has its corresponding atom g_{γ_i} in the reference image y_L, then γ_j belongs to the subset $\Gamma_i = \Gamma_i^E \cap \Gamma_i^S$. Since Γ_i is usually much smaller than Γ, the Wyner-Ziv encoder does not need to send the whole γ_j, but can transmit only the

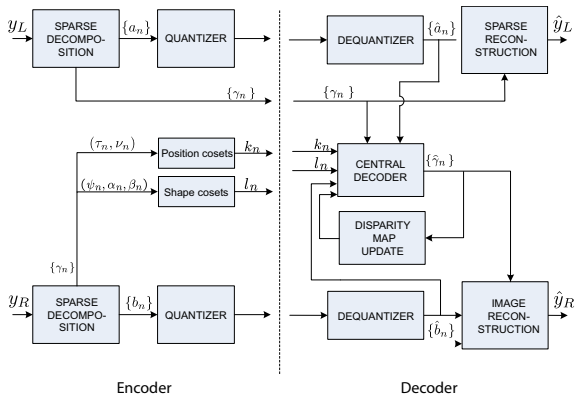


Figure 1: Occlusion-resilient Wyner-Ziv coder

information that is necessary to identify the correct atom in the transform candidate set given by Γ_i. This is achieved by coset coding based on partitioning of Γ into distinct cosets that contain dissimilar atoms with respect to their position (τ, ν) and shape (ψ, α, β). Two types of cosets were constructed: Shape cosets and Position cosets. The encoder eventually sends for each atom only the indexes of the corresponding cosets (i.e., k_n and l_n in Fig. 1). We design Shape cosets by distributing all atoms whose parameters belong to Γ_i^μ, for all i, into different cosets. The Position cosets are designed as VQ cosets [3], which are constructed by 2-dimensional interleaved uniform quantization of atom positions (τ, ν) on a rectangular lattice.

The decoder matches corresponding atoms in the reference image and atoms within the cosets of the Wyner-Ziv image decomposition using the correlation model described earlier. The atom pairing is facilitated by the use of quantized coefficients of atoms, which are sent directly. Each identified atom pair contains the information about the local transform between the reference and Wyner-Ziv image, which is exploited by the decoder to update the disparity map between them. The transformation of the reference image with respect to the disparity map provides an approximation of the Wyner-Ziv image that is used as a side information for decoding the atoms without a correspondence in the reference image. These atoms are decoded based on the minimal mean square error between the currently decoded image and the side information. Finally, the WZ image reconstruction $\hat{y}_R$ is obtained as a linear combination of the decoded image, reconstructed by decoded atoms from $\mathbf{\Phi_{I_R}}$, and the projection of the transformed reference image y_{tr} to the orthogonal complement of $\mathbf{\Phi_{I_R}}$ [3].

4. STEREO DICTIONARY LEARNING

The described Wyner-Ziv coding method has been implemented in [3] using the overcomplete parametric dictionary designed by uniform sampling of the transform parameters. Namely, the translations and rotations have been uniformly sampled on a linear scale, while the scaling parameters have been sampled uniformly on a logarithmic scale. Even if this choice has shown to result in dictionaries with good approximation properties for images [15], it is certainly not optimal especially for multi-view image representation with the pro-

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7466
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7466

posed geometric correlation model.

We therefore propose a stereo dictionary learning method, which is built upon the ML dictionary learning for monocular images introduced by Olshausen and Field [16]. The stereo image case is considered, where the stereo images follow the model in Eq. (1) and have the same number of sparse components related by local transforms. We consider a slightly more generic setting where the dictionaries $\mathbf{\Phi} = \{\phi_k\}, \mathbf{\Psi} = \{\psi_k\}$ used to represent stereo images y_L and y_R, respectively, are different. The maximum likelihood learning of overcomplete dictionaries $\mathbf{\Phi}, \mathbf{\Psi}$ maximizes the probability that stereo images captured by two cameras with a relative pose $(\mathbf{R}, \mathbf{T})$ are well represented by a set of atom pairs related by geometric transforms, under the sparsity prior. In other words, we want to simultaneously learn the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$ that well approximate the stereo images y_L and y_R, given the sparse stereo image model in Eq. (1). Moreover, we want to maximize the probability that the stereo images given by this model satisfy the disparity relation, i.e., that the epipolar constraint between all corresponding points on y_L and y_R is satisfied. Maximization of the disparity relations is crucial for learning dictionaries that have atoms with good epipolar matching properties, which is important in applications involving multi-view feature matching.

Formally, we want to solve the following optimization problem:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\max_{\mathbf{\Phi}, \mathbf{\Psi}} \langle \max_{\mathbf{a}, \mathbf{b}} \log P(y_L, y_R, D = 0 | \mathbf{\Phi}, \mathbf{\Psi}) \rangle, \quad (4)$$

where $D = 0$ denotes the event when the epipolar distance between all corresponding points on y_L and y_R is equal to zero (i.e., the epipolar constraint is satisfied). Marginalizing over $\mathbf{a}$ and $\mathbf{b}$ we have that:

$$P(y_L, y_R, D = 0 | \mathbf{\Phi}, \mathbf{\Psi}) = \\ = \int\int P(y_L, y_R, D = 0 | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}) \mathrm{d}\mathbf{a}\mathrm{d}\mathbf{b}. \quad (5)$$

We first need to define the joint distribution of coefficients $\mathbf{a}$ and $\mathbf{b}$, given dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, denoted as $P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi})$. Let us assume that pixels keep their intensity values under the local transforms induced by the viewpoint change. This assumption holds in multi-view images when the scene is assumed to be Lambertian, and when atom transforms correctly represent the local object transforms. Under this assumption, for a stereo atom pair ϕ_l, ψ_r linked with a transform F_{lr} the following equality holds (see Lemma 1 in [17]):

$$\langle y_R, \psi_r \rangle = \frac{1}{\sqrt{J_{lr}}} \langle y_L, \phi_l \rangle, \quad (6)$$

where J_{lr} is the Jacobian of the transform F_{lr}. Using the sparse image model and Eq. (6) we obtain the following probabilities:

$$\begin{aligned} P(b_r | a_l, \phi_l, \psi_r) &= P(a_l | b_r, \phi_l, \psi_r) \\ &= \frac{1}{z_b} \exp\left(-\frac{1}{2\sigma_b^2} (b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 \right), \end{aligned} \quad (7)$$

where z_b is the normalization factor and σ_b^2 is the variance of the zero-mean Gaussian noise that models the difference between b_r and $a_l / \sqrt{J_{lr}}$. We further assume that pairs of coefficients (a_l, b_r) are pair-wise independent, which is usually the case when image decompositions are sparse enough. Then, the distribution $P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi})$ is factorial, i.e.:

$$P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}) = \prod_{l=1}^{M} \prod_{r=1}^{M} P(a_l, b_r | \phi_l, \psi_r) = \\ P(\mathbf{a}) P(\mathbf{b}) \prod_{l=1}^{M} \prod_{r=1}^{M} \sqrt{P(b_r | a_l, \phi_l, \psi_r) P(a_l | b_r, \phi_l, \psi_r)}, \quad (8)$$

where we assume that priors on coefficients in each image $P(a_l)$ and $P(b_r)$ are independent of the atoms. Although in reality the distribution of the coefficients would depend on an arbitrarily chosen dictionary, imposing the independence of the coefficients with respect to the dictionary during learning would actually lead to inferring a dictionary that gives the same prior distribution of coefficients for all types of images.

For modeling the priors on coefficients, we assume that the coefficients a_l and b_r are i.i.d. and drawn from a Bernoulli distribution over the activity of coefficients, where a coefficient is different from zero with probability p and equal to zero with probability q. Thus, for $p \ll q$ the Bernoulli distribution can well model the prior on the sparse coefficients $\mathbf{a}$ and $\mathbf{b}$. If we take $p = 1/(1 + e^{1/\lambda})$, we have:

$$P(\mathbf{a}) = \frac{1}{z_\lambda} \exp\left(-\frac{\|\mathbf{a}\|_0}{\lambda} \right) \text{ and } P(\mathbf{b}) = \frac{1}{z_\lambda} \exp\left(-\frac{\|\mathbf{b}\|_0}{\lambda} \right),$$

where $\| \cdot \|_0$ denotes the l_0 norm and λ controls the level of "sparseness" of coefficients. For sparse vectors $\mathbf{a}$ and $\mathbf{b}$, the probabilities $P(\mathbf{a})$ and $P(\mathbf{b})$ are highly peaked at zero. Thus, we can approximate the probability $P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi})$ by its value at the maximum, since it is a product of a zero-mean Gaussian distribution and discrete distributions tightly peaked at zero. Eq. (5) then becomes:

$$P(y_L, y_R, D = 0 | \mathbf{\Phi}, \mathbf{\Psi}) \approx P(y_L, y_R | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) \cdot \\ \cdot P(D = 0 | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) P(\mathbf{a}, \mathbf{b} | \mathbf{\Phi}, \mathbf{\Psi}), \quad (9)$$

where we have used the fact that $D = 0$ does not bring more information to y_L, y_R than $\mathbf{\Phi}, \mathbf{\Psi}$. To evaluate our likelihood function, we next need to find the probability that the epipolar distance D is equal to zero given the stereo image model, i.e., we need to find $P(D = 0 | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$. The probability of epipolar matching for the stereo image pair can be modeled by the product of probabilities of epipolar matching for pairs of atoms that participate in sparse decompositions of the left and the right image, i.e., whose coefficients a_l and b_r are different from zero. If the epipolar estimation error is assumed to be Gaussian with zero mean and variance σ_D^2, we can model the probability $P(D = 0 | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$ as:

$$P(D = 0 | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) = \\ \frac{1}{z_D} \exp\left(-\frac{1}{2\sigma_D^2} \sum_{l=1}^{M} \sum_{r=1}^{M} \mathcal{I}(a_l) \mathcal{I}(b_r) D_E(\phi_l, \psi_r) \right). \quad (10)$$

where $\mathcal{I}$ denotes the indicator function, z_D is the normalization factor and $D_E(\phi_l, \psi_r)$ is the epipolar distance between stereo atoms. This distance can be easily evaluated by summing over the epipolar distances between points paired by the local transform between the corresponding atoms.

At this point, we have defined all components of the objective ML function in Eq. (9), except $P(y_L, y_R | \mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi})$. This probability can be modeled by a Gaussian white noise

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7466
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7466

of variance σ_I^2:

$$P(y_L, y_R|\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) = P(e_L + e_R)$$
$$= \frac{1}{z_I} \exp\left(-\frac{1}{2\sigma_I^2}(\|y_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|y_R - \mathbf{\Psi}\mathbf{b}\|_2^2)\right), \quad (11)$$

where we have used the fact that the sum of two zero-mean Gaussian random variables is also a zero-mean Gaussian random variable, and z_I is the normalization factor. We can now rewrite the ML learning problem in Eq. (4) as the following energy minimization problem:

$$(\mathbf{\Phi}, \mathbf{\Psi})^* = \arg\min_{\mathbf{\Phi}, \mathbf{\Psi}} \langle \min_{\mathbf{a}, \mathbf{b}} E(\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) \rangle, \quad (12)$$

where E denotes the energy function given as:

$$E(\mathbf{a}, \mathbf{b}, \mathbf{\Phi}, \mathbf{\Psi}) = \frac{1}{2\sigma_I^2}(\|y_L - \mathbf{\Phi}\mathbf{a}\|_2^2 + \|y_R - \mathbf{\Psi}\mathbf{b}\|_2^2) +$$
$$+\frac{1}{2\sigma_D^2} \sum_{l=1}^{M} \sum_{r=1}^{M} \mathcal{I}(a_l)\mathcal{I}(b_r) D_E(\phi_l, \psi_r)$$
$$+\frac{1}{2\sigma_b^2} \sum_{l=1}^{M} \sum_{r=1}^{M} (b_r - \frac{a_l}{\sqrt{J_{lr}}})^2 + \frac{1}{2\lambda}(\|\mathbf{a}\|_0 + \|\mathbf{b}\|_0). \quad (13)$$

The energy function thus consists of four main summation terms: 1) the approximation error term; 2) the epipolar constraint term; 3) the coefficient similarity term; and 4) the sparsity term. Unfortunately, the energy function is not convex, and we can find only the local minimum. We propose to use the Expectation-Maximization (EM) algorithm, which alternates between two steps:

1. **E step**, which minimizes the energy over the coefficients $\mathbf{a}$ and $\mathbf{b}$, while keeping the dictionaries fixed. Coefficients are found using a modified version of the Matching Pursuit (MP) algorithm. It selects the atoms that give the minimal value of the energy function, and then removes the contribution of those atoms from the stereo images. Thus, it selects m atoms for each of the stereo images.
2. **M step**, which minimizes the energy over the dictionaries $\mathbf{\Phi}$ and $\mathbf{\Psi}$, while keeping the coefficients fixed. Given the coefficients, the energy is a continuous function of $\mathbf{\Phi}$ and $\mathbf{\Psi}$, and can be minimized using the conjugate gradient method.

In the first iteration, the dictionaries are initialized randomly. The following iterations take the fixed values from the previous iteration. The E step and M steps are iteratively repeated until the convergence is achieved. The learning should be performed from a large set of different multi-view images with different camera poses.

5. EXPERIMENTAL RESULTS

This section first shows the results of stereo dictionary learning. We then use the learned dictionary in the Wyner-Ziv coding scheme described in Section 3.2 and compare its performance to the Wyner-Ziv coder with the uniformly sampled dictionary parameters.

5.1 Learning results

Since the scaling parameters are the most important for stereo matching, while translations and orientations are highly dependent on the position of the sensors, we choose here to focus on learning only the scaling parameters of the atoms. A parametric dictionary can then be constructed by applying to the generating function the learned scales and a discretized set of translations and orientations.

Table 1: Initial and learned scale parameters for the left and the right image, for different values of ρ.

Initial dictionary		learned dictionary			
		$\rho = 0$		$\rho = 1$	
$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$	$\alpha^{(L)}$	$\beta^{(L)}$
13.15	5.98	8.61	6.34	10.82	8.68
14.06	7.78	22.19	7.30	16.92	13.72
6.27	10.47	3.40	3.56	3.81	5.05
14.13	14.58	25.88	22.95	26.00	19.73
11.32	14.65	14.52	14.78	5.57	11.25
$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$	$\alpha^{(R)}$	$\beta^{(R)}$
6.58	6.42	2.94	2.69	3.58	4.73
14.71	9.22	12.18	5.04	11.72	8.43
14.57	14.16	25.93	20.30	25.57	18.94
9.85	12.92	6.60	6.80	5.70	10.56
13.00	14.59	15.87	16.05	15.08	14.52

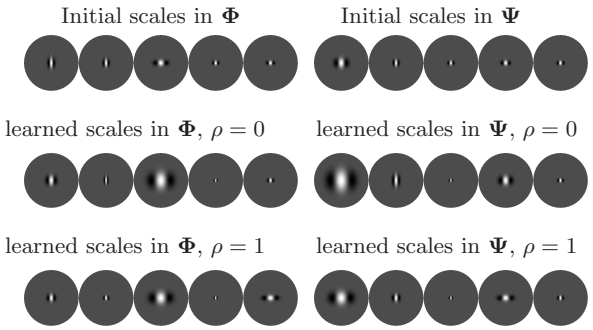


Figure 2: Subset of atoms in the initial and learned dictionaries for the left and right images.

We perform learning of stereo dictionaries for omnidirectional images mapped to spherical images. For representing spherical images, we use the formulation of a dictionary on the 2-D unit sphere [3]. The generating function is a Gaussian in one direction and its second derivative in the orthogonal direction. We have tested the proposed stereo dictionary learning algorithm on our "Mede" omnidirectional multi-view database, which consists of 54 omnidirectional images of an indoor environment. We have formed 216 pairs of images with different translation $\mathbf{T}$ between cameras, while the rotation $\mathbf{R}$ is identity.

We have learned five pairs of scaling parameters. The initial values of scales $\alpha^{(L)}$, $\beta^{(L)}$, $\alpha^{(R)}$ and $\beta^{(R)}$ have been chosen randomly, and they are given in the first two columns in Table 1. The atoms of the initial scales are shown in the first row in Fig. 2. The whole dictionary is built from these atoms by shifting them at all pixel locations and rotating in four orientations. To see the influence of the part of the objective function that relies on the multi-view constraint, we have introduced a factor ρ that multiplies the second and the third term in the energy function. When $\rho = 0$ the learning takes into account only the image approximation term,

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7466
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7466

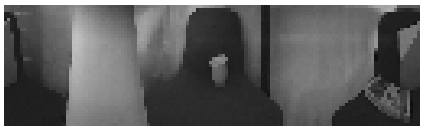
while increasing ρ puts more importance on the multi-view correlation term. From Fig. 2 we can see that for $\rho = 0$, the learned atoms are more elongated along the Gaussian direction, and more narrow in the direction of the second derivative of the Gaussian. These results are consistent with the previous work on dictionary learning for single-view image representation. However, for $\rho = 1$ we obtain different results for atoms scales (see Table 1). The atoms become more elongated along the direction of the Gaussian second derivative and narrower in the direction of the Gaussian. In addition, for $\rho = 1$ the learned scales generally tend to give smaller atoms than for $\rho = 0$. Finally, we observe that the dictionaries for both images are very similar since the learning strategy is symmetric.

5.2 DSC with the learned dictionary

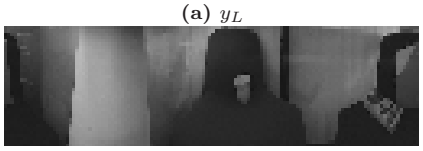
Since stereo learning results in two dictionaries that are very similar, we will use the parameters only of the dictionary $\mathbf{\Phi}$ for sparse approximation of two Lab images shown in Fig. 3. These are the same images used for evaluation in [3], for the camera distance of 10 cm. Lab images do not belong to the multi-view image database used for learning.

The Wyner-Ziv coder with the learned dictionary is essentially the same as the one in [3], but with a different dictionary used in the Matching Pursuit decomposition of both images. The learned dictionary is built on a generating function that is a Gaussian in one direction and the second derivative of a 2D Gaussian in the orthogonal direction (i.e., edge-like atoms). The position parameters τ and ν can take 128 different values, while the rotation parameter uses 16 orientations. We have used the learned scales of $\mathbf{\Phi}$, for $\rho = 1$. The image y_L is encoded independently at 0.21bpp with a PSNR of 30.61dB. The atom parameters for the expansion of image y_R are coded with the proposed scheme. The number of position cosets is the same as in [3], while we have used a smaller number of shape cosets (64 instead of 128) since we have a smaller total number of scales. The coefficients are obtained by projecting the image y_R on the atoms selected by MP, in order to improve the atom matching process, and they are quantized uniformly.

We have introduced one additional change with respect to the previous WZ coder. Namely, since the learned dictionary is built only on edge-like atoms, many atoms in the MP expansion align with the borders of the omnidirectional image when mapped to a spherical image, due to the mirror boundary. These atoms carry no geometry information, thus we need to encode them independently. Encoder distinguishes them from the other atoms when their translation parameter along the θ axis on the sphere corresponds to the mirror boundary. Therefore, the encoder sends an additional bit per atom to indicate which atom is encoded by coset coding and which one is independently encoded. Note that this was not required in [3] since the dictionary used there included also the 2D Gaussian atoms, which have been selected in the beginning of MP. Since they are low-frequency atoms, they do not align on the mirror boundary.



(a) y_L



(b) y_R

Figure 3: Original Lab images. The natural omnidirectional images partially cover the sphere due to the boundaries of the mirror in the omnidirectional camera. The images are cropped to focus on the captured part of the scene.

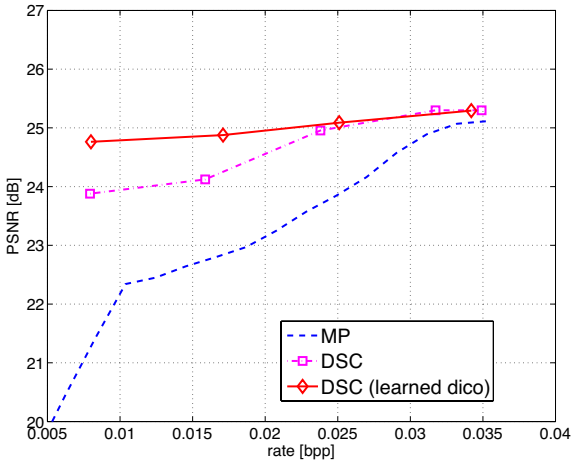
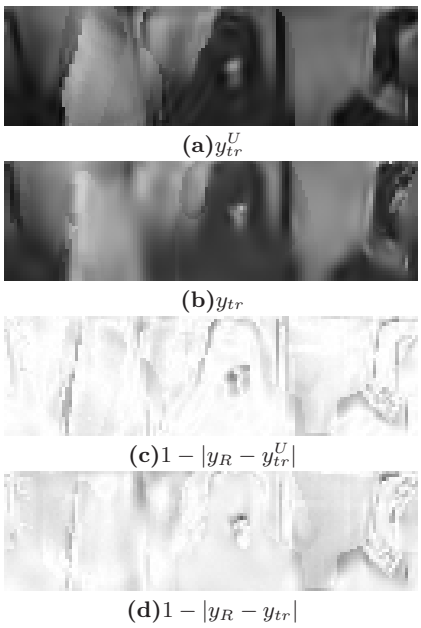


Figure 4: Rate-distortion performance of the Wyner-Ziv coding for image y_R.

Fig. 4 shows the rate-distortion (RD) curves for the Wyner-Ziv coding of image y_R, where the dash-dotted line corresponds to the uniformly sampled dictionary [3], and the solid line corresponds to the learned dictionary. The dashed line presents the RD performance of independent coding with MP. We can see that the learned dictionary improves the DSC performance only at low rates, while in the saturation zone of WZ coding, it performs the same as uniformly sampled dictionary. This suggests that the learned dictionary is mostly beneficial for improving the estimation of the WZ image obtained by disparity mapping from the reference image y_L, denoted as y_{tr}. However, learning the dictionary cannot correct the saturation effect of the proposed coder, and one has to employ the error resilient solution presented in [4].

The influence of the learned dictionary on the disparity compensated estimation of the Wyner-Ziv image (y_{tr}) is shown in Fig. 5. The y_{tr} image for the case of the uniformly sampled dictionary is shown in Fig. 5(a) (denoted as y_{tr}^U), while for the case of the learned dictionary is shown in Fig. 5(b) (denoted as y_{tr}). Both images correspond to the last point on the RD curves in Fig. 4. Figures 5(c) and (d) show respectively the differences between the images y_{tr}^U and y_{tr} and the original WZ image y_R, where white pixels correspond to zero. We can see that the learned dictionary results in a better estimation of the disparity map, which is especially visible in the center of the image. In this area, the

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7466
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7466



(a) y_{tr}^U

(b) y_{tr}

(c) $1 - |y_R - y_{tr}^U|$

(d) $1 - |y_R - y_{tr}|$

Figure 5: Comparison of the estimated WZ image y_{tr}, for the cases of uniform and learned dictionary. (a) y_{tr}^U for the case of uniform dictionary, (b) y_{tr} for the case of learned dictionary, (c) difference between y_{tr}^U and the original WZ image, (d) difference between y_{tr} and the original WZ image.

disparity estimation using the uniform dictionary still leaves small errors, while the atoms from the learned dictionary correctly recover the disparity. This shows the advantage of using the learned dictionaries for geometry based disparity matching between stereo images, since it results in improved RD performance of the proposed WZ coder.

6. CONCLUSIONS

This paper proposes the use of learned stereo dictionaries for the Wyner-Ziv multi-view coding. We have presented a maximum likelihood method for learning overcomplete dictionaries adapted to multi-view image representation. The learned dictionaries exhibit higher anisotropic atom properties and differ from the dictionaries used for single-view image representation. We have applied the learned dictionary to the Wyner-Ziv multi-view image coder previously proposed in [3]. The experimental results show that the new dictionary is mostly beneficial for improving the estimation of the Wyner-Ziv image obtained by disparity mapping from the reference image. Since this estimated image is used for the final Wyner-Ziv image reconstruction, the learned dictionary improves the RD performance of the Wyner-Ziv coder at low bit rates compared to a uniformly sampled dictionary.

7. ACKNOWLEDGMENTS

This work has been supported by the Swiss National Science Foundation under grant 200020-120063, and by the EU under the FP7 project APIDIS (ICT-216023). The authors would like to thank the members of the Redwood Center for Theoretical Neuroscience at UC Berkeley, for the fruitful discussions on the dictionary learning research.

8. REFERENCES

[1] Slepian D. and Wolf J. K. Noiseless coding of correlated information sources. *IEEE Trans. on Information Theory*, 19(4):471–480, 1973.

[2] Wyner A. D. and Ziv J. The rate-distortion function for source coding with side-information at the decoder. *IEEE Trans. on Information Theory*, 22(1):1–10, 1976.

[3] Tošić I. and Frossard P. Geometry-Based Distributed Scene Representation With Omnidirectional Vision Sensors. *IEEE Trans. on Image Processing*, 17(7):1033–1046, 2008.

[4] Tošić I. and Frossard P. Geometry-based distributed coding of multi-view omnidirectional images. In *Proc. of IEEE ICIP*, pages 2220–2223, 2008.

[5] Zhu X., Aaron A. and Girod B. Distributed compression for large camera arrays. In *Proc. of the IEEE PSSP*, 2003.

[6] Chen D., Varodayan D., Flierl M., and Girod B. Distributed stereo image coding with improved disparity and noise estimation. In *Proc. of IEEE ICASSP*, pages 1137–1140, 2008.

[7] Gehrig N. and Dragotti P. L. Geometry-Driven Distributed Compression of the Plenoptic Function: Performance Bounds and Constructive Algorithms. *IEEE Trans. on Image Processing*, 18(3):457– 470, 2009.

[8] Tošić I., Thirumalai V. and Frossard P. Symmetric distributed coding of stereo omnidirectional images. *Signal Processing: Image Communication*, 23(5):379 – 390, 2008. Special issue on distributed video coding.

[9] Flierl M. and Vandergheynst P. Distributed Coding of Highly Correlated Image Sequences with Motion-Compensated Temporal Wavelets. *EURASIP Journal on Applied Signal Processing*, Article ID 46747:10 pages, 2006.

[10] Ouaret M., Dufaux F. and Ebrahimi T. Fusion-based multiview distributed video coding. In *Proceedings of ACM International Workshop on Video Surveillance and Sensor Networks*, pages 139 – 144, 2006.

[11] Guo X., Lu Y., Wu F., Gao W. and Li S. Distributed multi-view video coding. In *Proceedings of the SPIE VCIP*, 2006.

[12] Song B., Tuncel E. and Roy-Chowdhury A. K. Towards A Multi-Terminal Video Compression Algorithm By Integrating Distributed Source Coding With Geometrical Constraints. *Journal Of Multimedia*, 2(3):9–16, 2007.

[13] Xiong Z., Yang Y., Stanković V. and Zhao W. Two-terminal video coding. *IEEE Trans. on Image Processing*, 18(3):534–551, 2009.

[14] Maugey T. and Pesquet-Popescu B. Side information estimation and new symmetric schemes for multi-view distributed video coding. *J. Vis. Comun. Image Represent.*, 19(8):589–599, 2008.

[15] Figueras i Ventura R. M., Vandergheynst P. and Frossard P. Low rate and flexible image coding with redundant representations. *IEEE Trans. on Image Processing*, 15(3):726–739, 2006.

[16] Olshausen B. and Field D. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–25, 1997.

[17] Tošić I. and Frossard P. Conditions for recovery of sparse signals correlated by local transforms. *Proc. of IEEE ISIT*, 2009.

Digital Object Identifier: 10.4108/ICST.MOBIMEDIA2009.7466
http://dx.doi.org/10.4108/ICST.MOBIMEDIA2009.7466