MOTION ESTIMATION FROM COMPRESSED LINEAR MEASUREMENTS

Vijayaraghavan Thirumalai and Pascal Frossard

Ecole Polytechnique Fédérale de Lausanne (EPFL)
Signal Processing Laboratory - LTS4, Lausanne, 1015 - Switzerland.
{vijayaraghavan.thirumalai, pascal.frossard}@epfl.ch

ABSTRACT

This paper presents a novel algorithm for computing the relative motion between images from compressed linear measurements. We propose a geometry based correlation model that describes the relative motion between images by translational motion of visual features. We focus on the problem of estimating the motion field from a reference image and a highly compressed image given by means of random projections, which are further quantized and entropy coded. We capture the most prominent visual features in the reference image using geometric basis functions. Then, we propose a regularized optimization problem for estimating the corresponding features in the compressed image, and eventually the dense motion field is generated from the local transform of the geometric features. Experimental results show that the proposed scheme defines an accurate motion field. In addition, when the motion field is used for image prediction, the resulting rate-distortion (RD) performance becomes better than the independent coding solution based on JPEG-2000, which demonstrates the potential of the proposed scheme for distributed coding algorithms.

1. INTRODUCTION

Distributed processing has recently found applications in vision sensor networks due to the low complexity encoding stage. One of the most important and challenging tasks in such a scenario is to estimate the correlation between the signals or images captured by different sensors, so that the information can be efficiently processed, coded or rendered. In this paper, we tackle the problem of estimating the correlation between a pair of images, where the common objects in different images are displaced due to the motion in the scene or positioning of the vision sensors. In particular, we are interested in computing this correlation when images are highly compressed and given under the form of few quantized linear measurements. This permits to have a low complexity acquisition that consists in computing inner products with a random projection matrix, instead of acquiring the entire image [1, 2].

We consider here the estimation of a motion field between the compressed image and the reference image. We model the motion between images as the geometric transformation of visual features. We first compute the most prominent visual features in the reference image and approximate them with geometric functions drawn from a parametric dictionary. We then formulate an optimization framework whose objective is to compute the corresponding features in the compressed image along with the relative geometric transformation. We add a regularization constraint in order to ensure that the estimated motion field is consistent and corresponds to the actual motion of visual objects. We show by experiments that the proposed algorithm accurately estimates the transformation between the pair of images. In particular, we show that dictionary based on geometric basis functions permits to capture the correlation more efficiently than the dictionary built on patches from the reference image [4]. In addition, we show that the motion field can be used to estimate the compressed image by motion compensation. Such reconstruction strategy outperforms independent coding scheme like JPEG 2000 in terms of RD performance, which outlines the potential of the proposed algorithm for distributed coding applications.

The concept of random projections in distributed scenarios has been previously studied in [3], where three joint sparsity models are designed and used in joint signal reconstruction algorithms. The problem of signal recovery based on random projections has recently been extended to distributed image or video coding, in an effort to reduce the complexity of the encoding stage [4, 5]. However, most works assume that the signal of interest is sparse in an orthonormal basis (e.g., DCT or wavelet) and fail to exploit the advantage of structured geometric dictionary in capturing the correlation between images. Few works have been reported in the literature in effort to build a correlation model for the images [7] or video [8] using a redundant structured dictionary. But these works developed the model using the approximated image but not from the linear measurements. However, we focus on estimating the motion from the random projections and the correlation model is built using the geometric transformation captured by the structured dictionary. Such a scheme is shown to be accurate in capturing the motion and provides an interesting alternative for distributed video or image coding with a simple encoding stage.

2. PROPOSED FRAMEWORK

We consider a framework where a pair of images I_1 and I_2 represents a scene at different time instants, or from different viewpoints, are correlated through the motion of visual objects. The images are transmitted to a joint decoder that estimates the relative motion between the received signals. The framework is illustrated in Fig. 1.

One of the images is encoded and decoded independently and serves as a reference image for motion estimation. While this image could be encoded with any coding algorithm, we choose here to represent the reference image I_1 by random linear measurements $y_1 = \psi\, I_1$ with a projection matrix ψ. The measurements are used by the decoder to reconstruct an approximation $\hat{I}_1$ using a convex optimization algorithm [9] under the assumption that I_1 is sparse in particular basis (e.g., a wavelet basis).

The second image I_2 is also projected on a random matrix ψ. The measurements $y_2 = \psi\, I_2$ are further quantized and are optionally entropy coded. The decoder performs the reverse operations (de-

This work has been partly supported by the Swiss National Science Foundation, under grant 200021-118230.

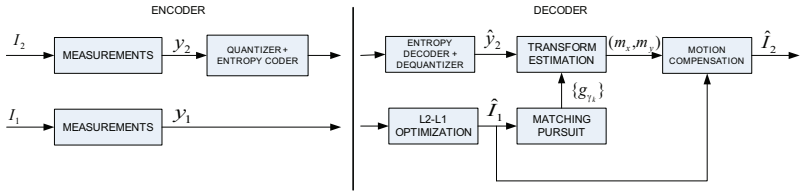


Fig. 1. Schematic representation of the proposed scheme.

quantization and entropy decoding) to form the measurement vector $\hat{y}_2$. This measurement vector is finally used by the joint decoder to estimate the relative transformation between the images I_1 and I_2.

We propose to model the correlation between the images by relative transformation between prominent visual features in both images. We assume that images I_1 and I_2 can be represented by sparse linear expansion of geometric function g_γ taken from a parametric and overcomplete dictionary $D = \{g_\gamma\}$. The geometric function g_γ in D is usually called as *atom*. The dictionary is constructed by applying set of geometric transformations to the generating function g. These geometric transformations can be represented by a family of unitary operator $U(\gamma)$, so that the dictionary spanning the input space takes the form $D = \{g_\gamma = U(\gamma)g, \gamma \in \Gamma\}$ for a given set of transformation indexes Γ. Typically this transformation set consists of scaling s_x, s_y, rotation θ, and translation t_x, t_y operators, defined as

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} 1/s_x & 0 \\ 0 & 1/s_y \end{bmatrix} \begin{bmatrix} cos\,\theta & sin\,\theta \\ -sin\,\theta & cos\,\theta \end{bmatrix} \begin{bmatrix} x - t_x \\ y - t_y \end{bmatrix}$$

where x, y defines the image coordinate. Thus, each of the transformation is indexed by five parameters.

We can then write the approximation of the first image with functions or atoms in D as

$$\hat{I}_1 \approx \sum_{k=1}^{N} c_k \, g_{\gamma_k}. \tag{1}$$

The approximation of $\hat{I}_1$ can be computed by sparse algorithms such as Matching Pursuit [6], which greedily pick up the N atoms $\{g_{\gamma_k}\}$ that best match the image $\hat{I}_1$. The second image I_2 can be described similar to equation 1. Under the assumption that the images I_1 and I_2 are correlated, the second image I_2 could be approximated with transformed version of the atoms used in the approximation of $\hat{I}_1$. We can thus write

$$I_2 \approx \sum_{k=1}^{N} c_k \, F^k(g_{\gamma_k}), \tag{2}$$

where $F^k(g_{\gamma_k})$ represents a local geometrical transformation of the atom g_{γ_k}. Due to the parametric form of the dictionary, the effect of F^k corresponds to a geometrical transformation of the atom g_{γ_k} that results in another atom in the same dictionary. Therefore, it is interesting to note that the transformation F^k on g_{γ_k}, boils down to a transformation of the atom parameters, i.e., $F^k(g_{\gamma_k}) = U(\delta\gamma)g_{\gamma_k} = g_{\gamma_k+\delta\gamma} = g_{\gamma'_k}$. Interestingly, local transformation of the atoms g_{γ_k} therefore lead to atoms that are part of a subdictionary gathering neighbours of g_{γ_k} and given as

$$D' = \{g_{\gamma'_k} \; : \; \gamma'_k \; \in \; [\gamma_k + \delta\gamma, \gamma_k - \delta\gamma]\} \tag{3}$$

Now the main challenge in the joint decoder is to estimate the local geometrical transformation F^k for each of the atoms g_{γ_k} from the linear measurements $\hat{y}_2$. In the next section, we formulate a regularized optimization problem in order to estimate F^k, or equivalently the relative motion between images I_1 and I_2.

3. MOTION ESTIMATION FROM COMPRESSED SIGNALS

Given the set of N atoms $\{g_{\gamma_k}\}$ that approximate the first image, the motion estimation problem consists in finding the corresponding visual patterns in the second image, while the latter is only given by compressed random measurements $\hat{y}_2$. This is equivalent to finding the correlation between the images, with the joint sparsity model described above.

We propose to estimate the transformation F^k iteratively, by deforming each of the N atom parameters γ_k by one increment in the parameter space. In particular, as we search for translational motion, we focus on the search space S that is given by perturbing each atom position by one unit, i.e., $t_x \pm 1$ and $t_y \pm 1$ for each atom γ_k. We initialize the algorithm with zero motion, i.e., the atoms $\{g_{\gamma_k}\}$ generated from $\hat{I}_1$ are used in the first iteration. Then at each iteration, we find the best N atoms $\{g_{\gamma'_k}\}$, or equivalently, the set of atom parameter Λ that minimizes the mean square error (MSE) w.r.t. the quantized measurements $\hat{y}_2$. More formally,

$$E_d = \min_{\Lambda \in S} \parallel \hat{y}_2 - \Psi_\Lambda \Psi_\Lambda^\dagger \hat{y}_2 \parallel^2 \tag{4}$$

where $\Psi = \psi[g_{\gamma_1}|g_{\gamma_2}|.....|g_{\gamma_N}]$. Then the next iteration is initialized based on the solution of the previous iteration, and the process is continued for T iterations or till convergence is reached.

The local motion of atoms are used to define a dense motion field. Given a pair of corresponding atoms g_{γ_k} and $g_{\gamma'_k}$ in images I_1 and I_2 respectively, we first calculate the mapping of each pixel $\mathbf{z} = (x_p, y_p)$ in g_{γ_k} to its corresponding pixel $\mathbf{z}' = (x_q, y_q)$ on $g_{\gamma'_k}$ using equation 1. This grid transformation $(x_p - x_q, y_p - y_q)$ corresponds to the amount of local motion captured by the pair of atoms g_{γ_k} and $g_{\gamma'_k}$. Using a similar process, the mapping is established for all the N atom pairs from the respective transform parameters γ_k and γ'_k. Then the grid transformation captured by all the N pairs of atoms are merged together to estimate the motion field. One possibility is to take the weighted average of grid transformation induced by all the N atoms. One can assign the relative weight based on the response of the atom at the pixel location $\mathbf{z} = (x_p, y_p)$. Mathematically, the horizontal component of the motion field at the location $\mathbf{z}$ is given as

$$m_{x_p} = \frac{\sum_{k=1}^{N} w_{\mathbf{z}}^{(k)} (x_p^{(k)} - x_q^{(k)})}{\sum_{k=1}^{N} w_{\mathbf{z}}^{(k)}}. \tag{5}$$

where $w_{\mathbf{z}}^{(k)}$ is the response or value of the k^{th} atom at the location $\mathbf{z}$ i.e., $w_{\mathbf{z}}^{(k)} = g_{\gamma_k}(\mathbf{z}) = g_{\gamma_k}(x_p, y_p)$. The vertical component of the motion field is defined very similar to equation 5.

During motion estimation, we further enforce the smoothness of the motion field. The goal of the smoothness term is to penalize the atom deformation in the neighborhood, so that it results in coherent motion for the neighborhood atoms. We compute the smoothness cost function using

$$E_s = \sum_{p,q \in \mathcal{N}} V_{p,q} \tag{6}$$

where $\mathcal{N}$ is the usual 4 pixel neighborhood. The term $V_{p,q}$ is computed using $min((m_{x_p} - m_{x_q})^2 + (m_{y_p} - m_{y_q})^2, K)$, where m_{x_p} and m_{y_p} are the x and y component of the motion field at the pixel location $\mathbf{z} = (x_p, y_p)$. We then merge both cost functions E_d and E_s as a single cost function. The estimation of the motion field is finally given by the solution of the optimization problem given as

$$E = \min_{\Lambda \in S}(E_d + \lambda E_s) \tag{7}$$

We obtained the motion field in an iterative way where each iteration consists in the perturbation of one of the atoms from the previous iteration. The joint decoding algorithm is summarized in Algorithm 1.

Algorithm 1 Joint Decoder

1: Input N, λ, K, T
2: Generate $\{g_{\gamma_k}\}$ from $\hat{I}_1$ s.t. $\hat{I}_1 \approx \sum_{k=1}^{N} c_k \, g_{\gamma_k}$
3: Initialize $(m_x, m_y) = (0, 0)$ i.e., $\{\gamma'_k\} = \{\gamma_k\}$
4: **for** 1:T **do**
5: Generate index search space S as $\{\gamma'_k\} \cup \{\eta'_k\}$, with $\gamma'_k = (t_x^k, t_y^k, \theta^k, s_x^k, s_y^k)$ and $\eta'_k = (t_x^k \pm 1, t_y^k \pm 1, \theta^k, s_x^k, s_y^k)$
6: Evaluate data cost E_d
7: Compute motion field (m_x, m_y) and then calculate smoothness cost E_s
8: Find N atom indexes $\{\gamma'_k\}$ in S using Eq. 7
9: **end for**

4. EXPERIMENTAL RESULTS

4.1. Setup

The experiments have been performed on one synthetic image set (given in Fig. 2) and one natural image set with resolution 128×128. The natural image set is built from frames 1 and 18 of the container video sequence. The dictionary D is constructed using two generating functions, as explained in [6]. The first one consists of 2D Gaussian function to capture low frequency component. The second function represents Gaussian in one direction, and the second derivative of 2D Gaussian in the orthogonal direction to capture the edges. The translation parameters t_x and t_y varies from 1 to 128, while 10 rotation parameters are used between 0 and π. The scaling parameters are uniformly distributed on a logarithmic scale from one up to a sixth of the size of the image, with a resolution of one fifth of octave. The random projections are computed using Hadamard matrix of block size 8 [9]. The measurements y_2 are quantized uniformly using a two bit quantizer. The reference image I_1 is encoded independently using 3600, and 10,000 measurements for synthetic and natural image sets respectively. In both cases, the quality of $\hat{I}_1$ w.r.t. I_1 is approximately 30 dB. Matching Pursuit is carried out on $\hat{I}_1$, and the image $\hat{I}_1$ is approximated using $N = 10$ atoms for synthetic image set and $N = 50$ atoms for natural image set. The search window size is $\delta\gamma = 4$ pixels for the translation components t_x and t_y, and no changes in scale or rotation is considered.

4.2. Performance analysis

The transformation F^k is estimated using the algorithm described in Algorithm 1. The resulting dense motion field is used to warp the reference image $\hat{I}_1$ and the image thus reconstructed is represented by $\hat{I}_2$ (see Fig. 1). Fig. 3 and Fig. 4 show the comparison of the reconstructed image $\hat{I}_2$ w.r.t. I_2 and I_1 for synthetic and natural image sets respectively. It is clear that the MSE is small for $\hat{I}_2 - I_2$ compared to $\hat{I}_2 - I_1$, indicates that the image $\hat{I}_2$ is closer to I_2 than I_1. In other words the proposed scheme captures the correlation between the images efficiently. We also compare the accuracy of motion w.r.t. the ground truth. For the synthetic scene the normalized L1 norm error of the generated motion field (from 60 quantized measurements) w.r.t. ground is 0.02.

Furthermore, in order to demonstrate the effect of regularization, we estimate the transformation F^k based only on the data cost E_d, i.e., $\lambda = 0$, and Fig. 5 shows the reconstructed image $\hat{I}_2$ with and without regularization. It is clear that the regularization helps a lot to improve the quality of the reconstructed image $\hat{I}_2$.

Finally, Fig. 6 shows the RD comparison of the reconstructed image $\hat{I}_2$ with JPEG 2000 based coding strategy. The bit rate is computed by encoding the quantized measurements using an Arithmetic coder. From the Fig. 6 it is clear that the proposed scheme outperforms JPEG 2000 by a margin of almost 3 dB, especially at lower rates. Similar observation is made for synthetic image set. For example, to attain the reconstruction quality of 27.6 dB, JPEG 2000 requires approx 1800 bits, while proposed scheme requires only approx 150 bits. It is worth mentioning that the gain over independent coding scheme is achieved mainly by compensating the motion between the images, and further gain could be achieved by improving the reconstruction stage.

4.3. Benefit of structured dictionary

In order to demonstrate the benefit of geometric dictionary, we compared the results to a scheme, which adaptively constructs the dictionary using blocks or patches in the reference image [4]. As demonstrated in [4], we divide the image I_2 into 8×8 blocks and the measurements are generated for each blocks, and further they are quantized, and entropy coded. The decoder selects the best block within the search window that minimizes the MSE. The displacement between the corresponding blocks represents the motion field. The generated motion field is used to warp the reference image $\hat{I}_1$, and the image $\hat{I}_2$ is thus reconstructed. Fig. 6 compares the quality of reconstructed image $\hat{I}_2$ with our scheme, and it is clear that our scheme outperforms block based scheme, mainly due to rich representation of the visual information provided by structured dictionary.

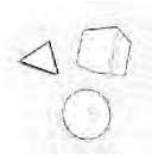
5. CONCLUSIONS

In this paper we have presented a method to estimate the dense motion field from the linear measurements. We have used structured dictionary to capture the prominent geometric features in the images. We relate the prominent features in both images using a geometry based correlation model. Then the motion field is computed using a regularized optimization under local transform constraints. Experimental results demonstrate that the proposed methodology is able to compute an accurate dense motion field, which opens an interesting perspective towards the design of distributed coding algorithms in vision sensor networks.

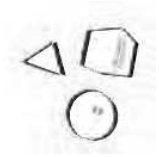
(a)

(b)

Fig. 2. Synthetic Image set (a) Image I_1 (b) Image I_2



$(a) MSE : 116$



$(b) MSE : 446$

Fig. 3. Synthetic Image set: Comparison of $\hat{I}_2$ with I_2 and I_1 (a) $1 - |\hat{I}_2 - I_2|$ (b) $1 - |\hat{I}_2 - I_1|$ (white pixels denotes no error). The image $\hat{I}_2$ is reconstructed using 60 quantized measurements.

$(a) MSE : 252$

$(b) MSE : 334$

Fig. 4. Natural Image set: Comparison of $\hat{I}_2$ with I_2 and I_1 (a) $1 - |\hat{I}_2 - I_2|$ (b) $1 - |\hat{I}_2 - I_1|$ (white pixels denotes no error). The image $\hat{I}_2$ is reconstructed using 1700 quantized measurements.

$(a) PSNR : 21.47\ dB$

$(b) PSNR : 27.63\ dB$

Fig. 5. Reconstructed image $\hat{I}_2$ using 60 quantized measurements. (a) without regularization ($\lambda = 0$) (b) with regularization

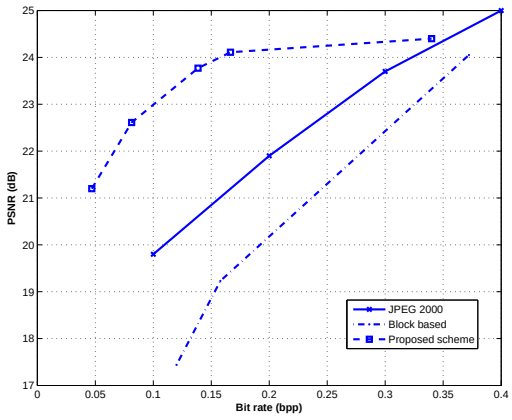


Fig. 6. Natural Image set: RD comparison of the proposed scheme with JPEG 2000 and the block based scheme [4].

6. REFERENCES

[1] D. Donoho, "Compressed Sensing," IEEE Trans. Infor. Theory, vol. 52(4), pp. 1289 - 1306, Apr. 2006.

[2] E. J. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," IEEE Trans. Infor. Theory, vol. 52, pp. 489 - 509, Feb. 2006.

[3] M. F. Duarte, S. Sarvotham, D. Baron, M. B. Wakin, and R.G. Baraniuk, "Distributed compressed sensing of jointly sparse signals," Proc. Asilomar Conf. on Sig. Sys. and Comp., Oct. 2005.

[4] J. P. Nebot, Y. Ma, and T. Huang, "Distributed video coding using compressive sampling," Proc. of Picture Coding Symp., May 2009.

[5] L. W. Kang and C. S. Lu , "Distributed Compressive Video Sensing," Proc. of Intl. Conf. on Acoustics, Speech, and Sig. Proc., Apr. 2009.

[6] R. M. Figueras, P. Vandergheynst, and P. Frossard, "Low-rate and flexible image coding with redundant representations," IEEE Trans. Imag. Proc., vol. 15, pp. 726 - 739, Mar. 2006.

[7] I. Tosic and P. Frossard, "Geometry based distributed scene representation with omnidirectional vision sensors," IEEE Trans. Imag. Proc., vol. 17, pp. 1033 - 1046, July 2008.

[8] O. D. Escoda, G. Monaci, R. M. Figueras, P. Vandergheynst and M. Bierlaire, "Geometric video approximation using weighted Matching pursuit," accepted to IEEE Trans. Imag. Proc..

[9] L. Gan, T. T. Do and T. D. Tran, "Fast compressive imaging using scrambled Hadamard ensemble," Proc. of European Sig. and Imag. Proc. Conf., Aug. 2008.