JOINT DECODING OF STEREO JPEG IMAGE PAIRS

Markus B. Schenkel[1,2], *Chong Luo*[1], *Pascal Frossard*[2] *and Feng Wu*[1]

[1] Microsoft Research Asia, Beijing, 100190, China
[2] Signal Processing Laboratory (LTS4), EPFL, Lausanne, 1015, Switzerland

ABSTRACT

This paper addresses the problem of joint decoding of stereo JPEG image pairs. Such images typically contain a high degree of redundancy. Predictive coding could efficiently capture this redundancy, but cameras would have to implement proprietary encoding solutions in this case as no such standard technology is available. We propose to rather use the popular JPEG compression tools in the cameras, and focus on the joint decoding problem for quality enhancement. We formulate this as a constrained optimization problem and show how regularization leads to more consistent results. It is similar to a distributed source coding framework, where the exploitation of the correlation at the decoder permits to save on the overall bandwidth. Experiments on natural stereo images show an improvement in both visual quality and PSNR when compared to separate decoding.

1. INTRODUCTION

Stereo perception and 3D rendering are usually achieved by simultaneously presenting two images of the same scene taken from two slightly displaced positions to both eyes. We will denote these two images with left (L) and right (R). Because they both represent the same reality they are highly redundant and accordingly we can expect significant compression ratios to be achievable. The depth information directly translates into the disparity that relates the two images and in the case of perfectly rectified views the possible transformations reduce to horizontal shifts only. Therefore a predictive coding scheme could be employed by intra coding the first image, followed by the disparity field for a predictor and the residue.

Such a traditional coding has previously been proposed by Perkins [1] for example. But the proprietary transmission format required by such an approach is unlikely to replace established general purpose image formats and it is worth investigating how much we could still improve without departing from JPEG encoding. Indeed, today's digital cameras are widely equipped with a JPEG encoder and stereoscopic image pairs are encoded as separate JPEG images by most applications, be it JPEG Stereo (JPS), Multi-Picture Object (MPO) [2] or others; thus they directly double the bandwidth requirements. Cameras also have a limited amount of processing power when compared to the corresponding decoders; this further motivates a distributed coding scheme.

A joint decoding strategy is proposed in this paper, in order to enhance the quality of the reconstructed image pairs that have been compressed independently with JPEG. Overall, the joint decoder allows for a lower overall bitrate for a given quality for both views. We cast the reconstruction problem as a regularized convex optimization problem that is constrained by consistent reconstruction conditions. We show that regularization permits to increase the accuracy of the disparity estimation, hence to obtain better reconstruction quality.

In particular, we consider an asymmetric coding scheme where one of the images is encoded at high quality, the other one at a reduced quality. Such an asymmetric scheme is of particular interest to applications where the second view is not always required, notably if no stereo display is available. In addition, controlled asymmetry does not really penalize 3D perception. Studies by Seuntiens et al. [3] and others indicate that the human brain can tolerate a fair amount of asymmetric image quality for stereo viewing such that the perceived quality lies between that of the two views.

This work is related to the distributed coding framework, where joint decoding is used to reconstruct correlated signals that have been independently encoded. However, we do not work here on the coding strategy, but rather rely on joint decoding approaches of signals that have been encoded with classical solutions. The joint reconstruction of compressed images has been considered also in the compressed sensing community, where different approaches have been proposed to represent images or parts of them as a sparse linear combination of other images assembled in a dictionary. If such a representation exists, it can under some conditions also be recovered in a stable way from linear projections onto a set of random vectors that reduces its dimensionality. This has led to applications in video coding where the dictionary is composed of blocks from a previous frame [4], face recognition where the candidate faces build the dictionary [5], or multiview representations [6]. In our solution a reconstructed block will be based on local dictionaries of candidate blocks. Compressed sensing in combination with total variation minimization has also been used for the compression of depth maps [7]. Alternatively, super-resolution reconstruction from image sequences has a similar objective of quality enhancement with multiple compressed images. It tries to either estimate a dense displacement field or to fuse different frames of a video together to enhance its quality. This is often formulated as an inverse optimization problem with a smoothness constraint on the displacement field (e.g. [8] and [9]), but unlike here more than two images are usually involved where reconstruction at higher resolution is the main target.

2. PROPOSED SCHEME

JPEG is a block-based still image compression scheme that compacts the image energy in a small number of coefficients and introduces losses mostly at high frequencies where they are visually more acceptable [10]. This is done by applying the two-dimensional discrete cosine transform (DCT) denoted by $\mathbf{D}$, followed by scalar quantization with up to ten times bigger step sizes at the highest frequencies than for the lower ones. The quantization step sizes are given by a table $\boldsymbol{q}$. If $\boldsymbol{b}$ is an image block and $\boldsymbol{y} = \mathbf{D}\boldsymbol{b}$ its representation in the transform domain, then quantization can be written as $\overline{y}_i = q_i\,[\,y_i/q_i\,]$ where $[\,\cdot\,]$ denotes rounding to the nearest integer. Because of the lossy nature of JPEG encoding we have a certain freedom to fill in the coarsely quantized coefficients from the other image at higher quality.

In the following we will always use the left image as the intra coded reference and the right one as the compressed image that we want to enhance. We limit our studies to luminance images only, but this extends to the other color components as well, because they are coded in the same fashion.

The compressed version of the right image defines a set of possible solutions for approximating the original image version, which are all consistent with respect to the compressed one, meaning that they would yield exactly the same JPEG bitstream after a recompression using the same quantization matrix. This is the range we will operate in to reconstruct the image. Although a midpoint or a centroid dequantization followed by an inverse DCT will likely minimize the reconstruction error if no further information is present, they are not the only choices within the aforementioned admissible region. Thus we will use the compressed image only to formulate a constraint on the output image. Our scheme operates on the blocks of 8×8 pixels defined by JPEG. We can formulate this elementwise constraint in the transform domain as

$$\left|\mathbf{D}(\hat{\boldsymbol{b}}^{(i)} - \overline{\boldsymbol{b}}^{(i)})\right| \preceq \frac{1}{2}\boldsymbol{q}\,, \tag{1}$$

where $\overline{\boldsymbol{b}}^{(i)}$ is the i^{th} block in the pixel domain after JPEG compression, $\hat{\boldsymbol{b}}^{(i)}$ is the estimate of that block after enhancement and the element-wise inequality $|x_j| \leq y_j \;\forall j$ is written as $|\boldsymbol{x}| \preceq \boldsymbol{y}$.

Now we build a dictionary $\boldsymbol{\Psi}^{(i)}$ composed of possible candidate blocks $\boldsymbol{\psi}_j^{(i)}$ from the reference image. Figure 1 illustrates the origin of the dictionary for block i. They are gathered from a range from 0 disparity up to the maximum disparity D_h in horizontal direction around the location in question. We further extend this range to $2D_v + 1$ shifts in the vertical direction to accommodate slight misalignments of the two cameras. Each of the $D = D_h(2D_v + 1)$ dictionary elements $\boldsymbol{\psi}_j^{(i)}$ now has an associated disparity vector $\boldsymbol{d}_j \in \{0, \dots, D_h\} \times \{-D_v, \dots, D_v\}$. They will be used later on to build the disparity fields $\boldsymbol{v}_x$ and $\boldsymbol{v}_y$ which associate to each block a horizontal and vertical shift respectively.

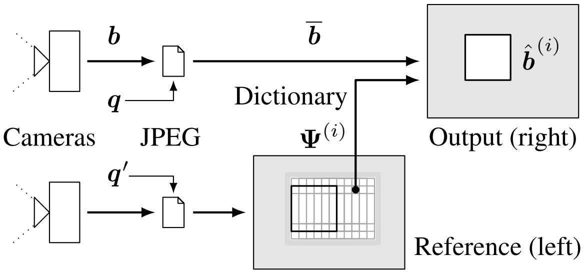


Fig. 1. Schematic view of the proposed scheme. Illustrating the separate encoding (with a fine $\boldsymbol{q}'$ and a coarse $\boldsymbol{q}$), the composition of the dictionary for a given block i from the reference view and the reconstruction of that block.

We now want to represent a block $\boldsymbol{b}^{(i)}$ as a linear combination of dictionary elements with the coefficient vector $\boldsymbol{s}^{(i)} \in \mathbb{R}^D$ as

$$\boldsymbol{b}^{(i)} \approx \sum_j \boldsymbol{\psi}_j^{(i)} s_j^{(i)} = \boldsymbol{\Psi}^{(i)} \boldsymbol{s}^{(i)}.$$

In general it is not possible to find such a decomposition that also satisfies (1). Thus we introduce the slack variables $\hat{\boldsymbol{b}}^{(i)}$ that are constrained as above and approximated by a linear combination over $\{\boldsymbol{\psi}_j^{(i)}\}$.

However, this inverse problem is still ill posed and we can regularize it in two ways. First, only a small number of the dictionary elements will contribute – ideally only a single one – and we can thus require $\boldsymbol{s}$ to be sparse. Although a high sparsity is best described with a low ℓ_0 pseudo-norm $\|\boldsymbol{s}\|_0 = |\{s_j | s_j \neq 0\}|$ this is not a convex function and would lead to a problem of combinatorial nature. Consequently we approximate it by the convex ℓ_1 norm $\|\boldsymbol{s}\|_1 = \sum_j |s_j|$.

Second, we can assume the disparity field $\boldsymbol{v}$ to be piecewise smooth because disparity discontinuities will occur only at object boundaries. Furthermore, not every block contributes an equal amount of depth information. Hence we can improve the reconstruction by enforcing a low total variation $\|\cdot\|_{\text{TV}} = \frac{1}{N}\sum_{j,k} |(\nabla \cdot)_{j,k}|$ of the disparity field $\boldsymbol{v}^{(i)} = [v_x^{(i)}\, v_y^{(i)}]^T$. We calculate $\boldsymbol{v}^{(i)}$ for each block as the weighted sum of the contributing disparities $\boldsymbol{v}^{(i)} = \sum_j \boldsymbol{d}_j\, s_j^{(i)}$. This step makes our problem a global one involving all blocks at once. In the following, all N blocks of an image are concatenated such that $\boldsymbol{b}^T = [\boldsymbol{b}^{(1)\,T} \cdots \boldsymbol{b}^{(N)\,T}]$ and the transform $\mathbf{D}$ becomes a block diagonal matrix.

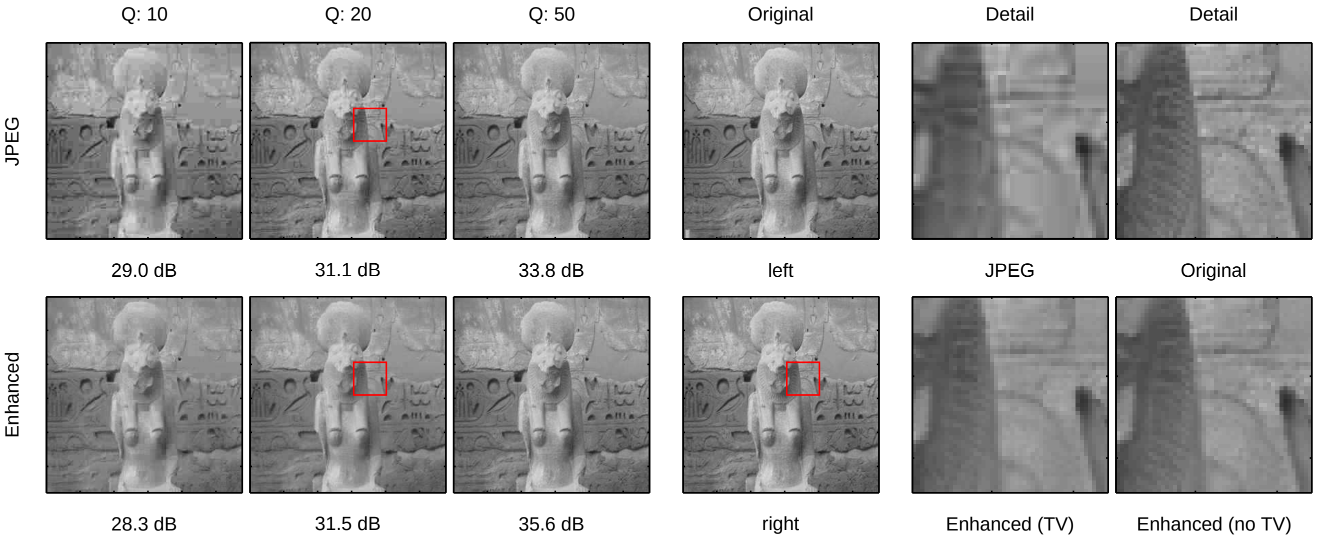


Fig. 2. This experiment on a 288×288 pixel image of *Sekhmet* taken with a commercial stereo camera illustrates our method. Results for three different JPEG quality settings with and without enhancement are followed by the original left and right view and enlarged details from the second image pair. We can improve the visual quality even if PSNR results drop and find less blocking artifacts if the total variation of the disparity field is minimized. Compare these results with Fig. 3(a).

Putting it all together we obtain the objective function

$$L(\hat{\boldsymbol{b}}, \boldsymbol{s}) = \left\| \hat{\boldsymbol{b}} - \boldsymbol{\Psi} \boldsymbol{s} \right\|_2^2 + \lambda_v \left(\|\boldsymbol{v}_x\|_{\mathrm{TV}} + \|\boldsymbol{v}_y\|_{\mathrm{TV}} \right) + \lambda_s \, \|\boldsymbol{s}\|_1$$

$$\hat{\boldsymbol{b}} = \arg\min_{\hat{\boldsymbol{b}}, \boldsymbol{s}} \; L(\hat{\boldsymbol{b}}, \boldsymbol{s}) \quad \text{s.t.} \quad \left| \mathbf{D} \left(\hat{\boldsymbol{b}} - \overline{\boldsymbol{b}} \right) \right| \preceq \frac{1}{2} \boldsymbol{q}. \qquad (2)$$

The parameter λ_v is the Lagrange multiplier weighting the importance of a smooth disparity field. The second parameter λ_s acts on the sparsity of the signal and trades off fidelity versus sparsity. All three terms in (2) scale with the number of blocks N, hence we can set the relative values of λ_s and λ_v independently of the image size.

The choice of λ_s is crucial, because on one hand a big value makes $\boldsymbol{s}$ approach $\boldsymbol{0}$, while on the other hand a small value can lead to a non-sparse $\boldsymbol{s}$ and a blurred result. We choose it such that the terms of the objective function are of similar magnitude and verify empirically that $\lambda_s = 5 \times 10^{-2}$ leads to good results for the tested images and a dictionary of size $D = 13 \times 9 = 117$. The choice of λ_v is less critical and we set it to $\lambda_v = 1 \times 10^{-2}$. All results presented in the following were obtained with this same set of parameters.

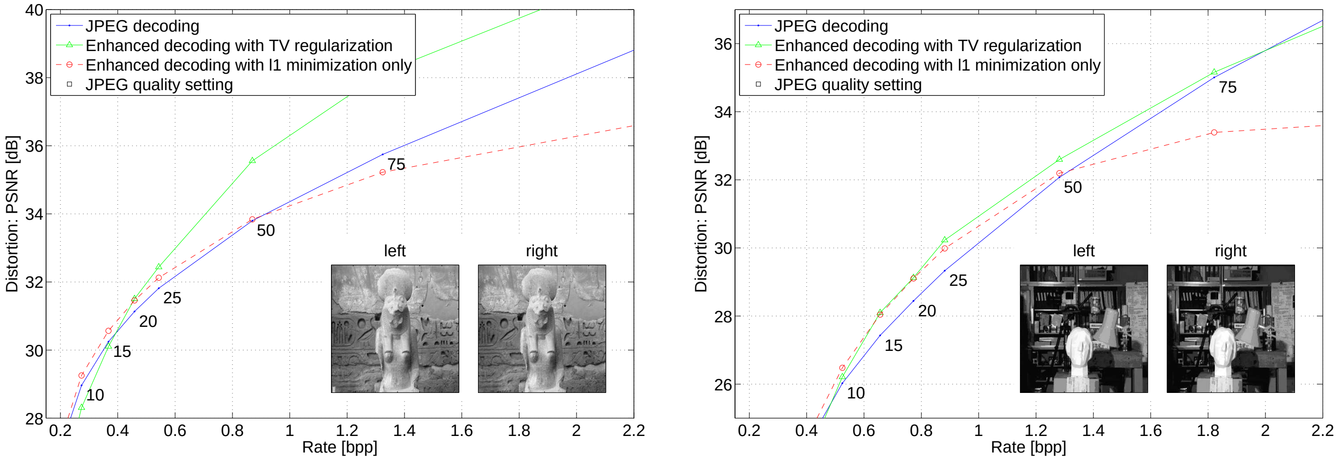
In order to solve (2) we use the `cvx` package for Matlab [11, 12] which applies an interior point method to it. Larger images are further partitioned into independently reconstructed regions of smaller size to keep the memory requirements manageable. Solving this problem is quite involved, but it scales linearly with the image size. An accurate knowledge of the maximum disparity reduces the number of variables and thus the runtime of the optimization.

3. EXPERIMENTAL RESULTS

We analyze in this section the performance of the joint decoding algorithm. As we can see from Fig. 2 we are able to enhance an image with correctly positioned details even if the right image is highly compressed. The ubiquitous blocking artifacts introduced by JPEG disappear and texture is added. Despite the small improvement in peak signal-to-noise ratio (PSNR) at low bitrates, the visual quality of the images improves clearly and in a consistent way. The enlarged areas show that even small details can be recovered that would simply be blurred out by JPEG. For medium bitrates (JPEG quality around 50) the PSNR improves by about 1 dB on average and more as the rate-distortion comparison in Fig. 3 highlights. In the optimal region we can accommodate a bitrate saving of 20% and above for the right view at a similar decoding quality.

Regions that are occluded in the reference view cannot possibly be reconstructed by this method. In the middle range (Quality 50 – 75) only few blocks have a PSNR decrease and they lie usually in such regions around disparity discontinuities as well as the very right border of the image. Nevertheless, ghosting artifacts appear seldom. Even though the results shown here were obtained with well aligned image pairs, additional experiments after multiple pixel shifts and a 2° rotation still exhibit good performance.

Furthermore, we study the influence of the individual parts of the optimization. First, we can set $\lambda_v = 0$ to remove the regularization of the disparity field with a total variation constraint. We find that at low bitrates the compressed image



(a) The *Sekhmet* image from Fig. 2 (288 × 288 pixels).

(b) The *Tsukuba* stereo test set (192 × 192 pixels).

Fig. 3. Rate-distortion comparison of JPEG (—•—) and joint decoding (—△—) for the right view of two image pairs. Also shown is the curve for unconstrained ℓ_1 minimization (-○-) given by Eq. (3).

might not contain enough details to reliably find a corresponding block in the reference view and it is in this region where the additional regularization of the disparity field leads to a further improvement. Although the increase in PSNR is only little, less artifacts are visible. The last part of Fig. 2 shows such a case. Second, we can also compare our method with the unconstrained ℓ_1 minimization

$$\hat{\boldsymbol{b}} = \boldsymbol{\Psi}\hat{\boldsymbol{s}}, \quad \hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} \left\|\overline{\boldsymbol{b}} - \boldsymbol{\Psi}\boldsymbol{s}\right\|_2^2 + \lambda_s \left\|\boldsymbol{s}\right\|_1 . \qquad (3)$$

This gives an improvement at low rates as seen from Fig. 3, but also tends to saturate at some PSNR level. We can expect this since the output is a sparse linear combination over the reference image which does not contain all the details. With the consistency constraint (1) we can overcome this gap.

Finally, we should note that in these experiments the reference image was always given at full quality. If the reference itself is compressed the improvements will naturally decrease; however, reference images at a quality setting of 80 and 90 could still be used successfully, as it has been confirmed by additional experiments.

4. CONCLUSIONS

We have presented a joint decoding solution for stereo image pairs. This method permits to reduce the bitrate of one view of a stereo image pair that is based on two separately coded, standard compliant JPEG images, but produces visually much better results than separate decoding. The only assumptions we make about the image pair is a relatively good alignment and a known maximum disparity. However both are only required to reduce the runtime of the algorithm.

We showed that it is possible to bring the quality of a second view closer to that of the reference image and simultaneously mitigate the effect of JPEG compression artifacts. The presented scheme provides good results for a viewing application because the two images will be of comparable quality. On the other hand it might not be a good preprocessing step for vision applications although the coarse depth maps obtained as a side product indicate that a fair amount of disparity estimation can still be done after compression.

5. REFERENCES

[1] M.G. Perkins, "Data Compression of Stereopairs," *IEEE Transactions on Communications*, vol. 40, no. 4, pp. 684–696, 1992.

[2] "Multi-Picture Format," *Camera & Imaging Products Association Standardization Committee*, 2009.

[3] P. Seuntiens, L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: effects of symmetric and asymmetric JPEG coding and camera separation," *ACM Transactions on Applied Perception*, vol. V, 2006.

[4] J. Prades-Nebot, Y. Ma, and T. Huang, "Distributed video coding using compressive sampling," *Proceedings of the Picture Coding Symposium*, 2009.

[5] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, and Y. Ma, "Robust face recognition via sparse representation.," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 2, pp. 210–27, 2009.

[6] V. Thirumalai and P. Frossard, "Motion estimation from compressed linear measurements," *Proceedings of ICASSP*, 2010.

[7] M. Sarkis and K. Diepold, "Depth map compression via compressed sensing," in *Proceedings ICIP*, November 2009.

[8] B.K.P. Horn and B.G. Schunck, "Determining Optical Flow," *Artificial Intelligence*, vol. 17, 1980.

[9] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers, "An improved algorithm for TV-L1 optical flow," *Statistical and Geometrical Approaches to Visual Motion Analysis: International Dagstuhl Seminar*, pp. 23–45, 2009.

[10] W. Pennebaker and J. Mitchell, *JPEG still image data compression standard*, New York, 1993.

[11] S. Boyd and M. Grant, "CVX: Matlab software for disciplined convex programming," *http://cvxr.com/cvx*, 2009.

[12] M. Grant and S. Boyd, *Graph implementations for nonsmooth convex programs, Recent Advances in Learning and Control (a tribute to M. Vidyasagar)*, pp. 95–110, Springer, 2008.