BIT RATE ALLOCATION FOR DISPARITY ESTIMATION FROM COMPRESSED IMAGES

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 rate allocation scheme to compute the 3D structure of the scene from compressed stereo images, captured by the distributed vision sensor networks. The images captured at different view points are encoded independently with a balanced rate allocation. The central decoder jointly decodes the information from the encoders, and computes the 3D geometry of the scene in terms of depth map. We first consider the scenario of estimating the 3D geometry from the views, compressed using standard encoders, e.g., SPIHT. Unfortunately, we noticed that the depth value is not precisely reconstructed in the low contrast regions or region around weak edges. It is mainly due to the rate allocation scheme, that allocates the bits based on the variance of the coefficients. We therefore propose a rate allocation scheme, where each encoder first identifies the low contrast regions and then distributes the bits such that the visual information in the low contrast regions is preserved. At the same time, the approximation quality in the rest of the image should not be penalized significantly. We adapt the SPIHT coding scheme to implement the proposed rate allocation methodology. Experimental results show that for a given bit budget, the proposed encoding scheme reconstructs the 3D geometry with more accuracy comparing to SPIHT, JPEG 2000 and JPEG coding schemes.

1. INTRODUCTION

Vision sensor network usually consists of several cameras distributed in the 3D scene and are widely used in several applications that rely on the efficient 3D representation of the scene, e.g., 3DTV. These imaging systems pose several problems like multi-view coding, computing 3D structure or distribution of cameras in 3D space etc. In this paper, we consider the problem of reconstructing the 3D geometry of the scene in terms of depth map, from the multi-view images. The computed depth map of the scene could be further exploited for various applications like rendering, multi-view coding etc. In common practice these imaging systems are operated with limited bandwidth resources, and hence we restrict ourselves in encoding the visual information at low or medium rates. Also since the communication among the cameras consume power, we rely on the distributed processing, where each sensor encodes the information independently without the knowledge from the other sensor, and communicate only with the central point for joint decoding.

Fig. 1 shows a distributed vision network with two camera sensors C_1 and C_2, that encode the images I_1 and I_2 independently. In this setting, we are interested to compute the depth map at the joint decoder from the compressed stereo views, especially encoded at medium (or even low) bit rates. Several algorithms have been proposed in the literature to compute the depth map from the stereo images (see [3] for details), but these algorithms compute the depth map from the original images by assuming that the vision sensor encodes the information without any loss. In other words, the depth is computed at the joint decoder, by neglecting the distortion due to compression. In practice the images are often compressed to save bandwidth resources and the transmission costs. Unfortunately, standard encoding strategies (SPIHT, JPEG 2000 etc.) are not efficient in handling the images, when depth has to be computed from compressed images. In particular, when the images are approximated at low or medium bit rates, such schemes yield incorrect depth value, especially when the depth discontinuity occurs in the low contrast region (or region with comparable luminance value) of the image. We therefore propose a rate allocation scheme that gives importance to the reconstruction of the visual information in the low contrast regions, and at the same time does not penalize significantly reconstruction in the other regions (background). In particular, we adapt the SPIHT coding scheme to allocate enough bits in the low contrast regions, so that it preserves the weak edges in the approximated view, which facilitates the depth estimation algorithm to yield correct depth value in the low contrast regions. Such a rate allocation scheme is shown to lead to a more accurate depth map comparing the traditional coding schemes based on JPEG 2000, SPIHT and JPEG.

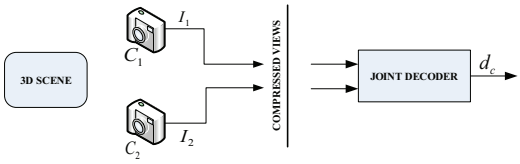

Fig. 1. Distributed vision sensor network with two cameras. The captured images I_1 and I_2 are encoded independently, and are transmitted to the joint decoder. The joint decoder computes the depth map d_c from the approximated views.

2. DEPTH ESTIMATION FROM COMPRESSED IMAGES

We consider the scenario shown in the Fig. 1, where the cameras C_1 and C_2 project the 3D visual information on the 2D plane I_1 and I_2 respectively, in different view points. The captured images I_1 and I_2 are encoded independently using b bits per image. Balanced rate allocation allows us to share the transmission and computational cost equally among the sensors and thus advantageously avoids the hierarchical relation among the sensors. At the joint decoder, the stereo image pair is first reconstructed independently, and is represented by $\hat{I}_1$, $\hat{I}_2$ respectively. Then the approximated images $\hat{I}_1$ and $\hat{I}_2$ are used to compute the dense depth map d_c, by assuming $\hat{I}_1$ as the reference.

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

(a) DE = 9.52 **(b) DE = 7.52**

Fig. 2. The images are encoded using SPIHT coding scheme and the depth map is computed at bit rates 0.4 bpp and 0.6 bpp per view. The corresponding disparity error is shown in (a) and (b) for the bit rates 0.4 and 0.6 bpp per view respectively. The white pixels correspond to disparity error > 1.

We first study the impact of the standard encoding procedures, on the performance of stereo matching algorithm. As an example, we select SPIHT coding scheme to encode the images, and Fig. 2 shows the disparity error w.r.t to the ground truth d_t on the Tsukuba test image [3] at medium bit rates 0.4 and 0.6 bpp per view. The disparity error (DE) is computed using

$$DE = \frac{1}{N_1 \times N_2} \sum_{x,y} (|d_c(x,y) - d_t(x,y)| > 1), \qquad (1)$$

where $N_1 \times N_2$ represents the dimension of the image [3]. For the sake of simplicity, compression is applied only on the luminance component of the image. As expected the overall quality in the depth map is better at higher bit rate (DE reduces from 9.4 to 7.52 respectively). However, it is quite important to observe that the quality of the depth map is improved only in the regions close to the strong edges (e.g., structure of the lamp, head), but the depth value remains unchanged in regions close to the weak edges or in low contrast regions (e.g., between the table leg and the background at the bottom most right corner). This behavior is observed because the SPIHT codec [4] minimizes the MSE in the reconstructed image, which obviously allocates more bits to the strong edges (high contrast regions) comparing weak edges (low contrast regions). Due to this rate allocation scheme, DE is significantly reduced in the high contrast regions, while the depth value remain unchanged in the low contrast regions. Similar observation is made with other standard encoding schemes like JPEG 2000, JPEG which also minimize the average MSE for a given target bit rate.

3. RATE ALLOCATION FOR IMPROVED DEPTH ESTIMATION

In the previous section, we discussed that the 3D structure of the scene is not reconstructed precisely near the weak edges or in the low contrast regions, when the stereo image pair is encoded using a traditional coding scheme. In order to improve the quality of the depth map at the joint decoder, we therefore propose a rate allocation scheme that preserves the weak edges (low contrast regions) in the reconstructed views $\hat{I}_1$ and $\hat{I}_2$, at the cost of a marginal penalization in the quality of the background. In the rest of the section, we discuss in detail about the proposed rate allocation scheme.

3.1. General Principle

Without loss of generality, we assume that the image $I \in \{I_1, I_2\}$ can be segmented into a low contrast region I^r and background I^b, such that $I = I^b \bigcup I^r$. Furthermore, the low contrast region I^r can be partitioned into k regions based on the strength of the edge, i.e., $I^r = \bigcup_{i=1:k} I^r_i$. Once the image I is partitioned into I^r and I^b, we distribute the bit budget b to the regions I^r and I^b based on principles of the ROI coding scheme [5], as described below.

1. Apply the desired transform on the image (e.g., Wavelets), and denote the transform coefficients by C.
2. Identify the set of transform coefficients $C^r_k \subset C$ from the region I^r_k.
3. Calculate the maximum value among the transform coefficients and then compute $M = \lceil log_2(max(C)) \rceil$. Also represent each transform coefficient using M bits with the bit value at position M as the most significant bit.
4. Compute $M_k = \lceil log_2(max(C^r_k)) \rceil < M$, and then calculate the scale $f_k = 2^{d_k}$ where $d_k = M - M_k$.
5. Multiply the coefficients C^r_k by f_k. As a result M_k is shifted to the bit plane M.

The steps explained above are illustrated in the Fig. 3 for $k = 2$. As shown in the Fig. 3(a), the transform coefficients C is partitioned into three regions, in which the left shaded region corresponds to the background coefficients C^b and the rest of them correspond to the low contrast regions C^r_1 and C^r_2. The bits in the top row represent the sign of the coefficients. The shift parameter is found to be $d_1 = 2$ and $d_2 = 4$ respectively, and the corresponding coefficients are scaled as shown in the Fig. 3(b). After scaling, the coefficients are progressively encoded starting from most significant bit plane M to the least significant bit plane along with the sign bit s. Also, the parameter d_k is transmitted to the decoder. While decoding, the coefficients are appropriately down scaled by f_k and the image is reconstructed by inverting the transform. The visual information in the low contrast region I^r is now preserved in the reconstructed image due to the scaling of corresponding coefficients C^r before encoding.

The shift parameter f_k proposed in our scheme for each region is certainly optimal. Suppose that $f_k = 2^{d_k+1}$, then after scaling the coefficients C^r_k are shifted to bit plane $M + 1$. In such cases the region I^r_k or in general I^r is decoded at higher quality, while penalizing the reconstruction in the background, since more bits are now spent in encoding the low contrast region than the background. This can be easily understood from the Fig. 3(b), by setting $d_1 = 3$ and $d_2 = 5$.

The steps outlined above illustrate the proposed rate allocation methodology. However, it is costly to implement such rate allocation methodology especially in sensor networks, due to the following reasons. Mask generation I^r involves a segmentation step that obviously increases the computational complexity and the power consumption at the sensor node. Also, the decoder needs a priori knowledge about the shape information I^r to start the decoding process, and thus requests the encoder to transmit the mask I^r which obviously increases the transmission cost. In order to alleviate these problems we propose to integrate the principles outlined above into a zero tree coder (e.g., SPIHT). It leads to a more practical solution that does not require any preprocessing nor segmentation step.

3.2. SPIHT - based rate allocation algorithm

In this section, we describe the proposed encoding scheme constructed using SPIHT coding principles. We denote C^l_i as the set of all coefficients in the i^{th} entry of the LIS (refer [4] for details). Due to the compact support of the Wavelet filter, the coefficients C^l_i represent a particular spatial region in the image. Also the magnitude of the coefficients C^l_i depend on the strength of the edge in

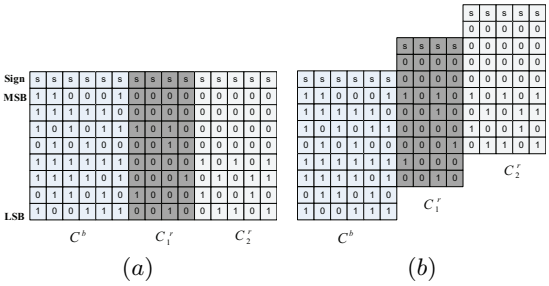

Fig. 3. Illustration of the proposed rate allocation scheme with $k = 2$. (a) Bit plane representation of original transform coefficients C with $M = 8$, and are partitioned into three regions. The bits in the top row represent the sign of the coefficients. (b) The coefficients C_1^r and C_2^r are shifted by $d_1 = 2$ and $d_2 = 4$ respectively.

the corresponding spatial region of the image, because the Wavelet coefficients are magnitude ordered according to the strength of the edge. i.e., for strong edges $M_i \approx M$, and for weak edges $M_i < M$, where $M_i = \lceil log_2(max(C_i^l)) \rceil$. Thus the spatial location of the low contrast regions can be identified directly from the magnitude of the Wavelet coefficients C_i^l. We then calculate the scale f_i for each i^{th} entry in LIS using $f_i = 2^{d_i}$ where $d_i = M - M_i$. It can be seen that the scaling parameter f_i is close to one for the entries in LIS which represent the strong edges, and it is close to 2^M for the weak edges. Thus the scaling parameter f_i is varied (between one to 2^M) inherently based on the strength of the edge in the image, more importantly without any preprocessing stage. Once the scale f_i is computed, the corresponding coefficients C_i^l are scaled, which obviously places M_i at the most significant bit plane M. As a result of this, the visual information in the low contrast regions are preserved in the reconstructed image, as explained in the section 3.1. Furthermore, as the number of coefficients in each entry of LIS is the same, the bit budget could be distributed equally among the entries of LIS.

It is well known that only the edges and the texture information are important to estimate the depth, and not the smooth regions. So the coefficients corresponding to the smooth regions are not encoded, and we propose to identify them in the transform domain based on thresholding. In more detail, we calculate the mean of the coefficients C_i^l in each i^{th} entry of the LIS and then compare it against a threshold T. If the mean value is smaller than T, then most probably the coefficients C_i^l are close to zero due to the smooth behavior, and it is not considered for rate allocation. So the bit rate is equally distributed among the remaining entries in LIS after thresholding. Also for correct decoding, the index i of the LIS that is not selected for rate allocation (whose mean value is less than the threshold) is signaled to the decoder, so that the decoder simply sets the coefficients C_i^l in the corresponding i^{th} LIS entry to zero.

Let us now discuss in more details about the rate allocation between the coefficients in the LL band and the descendant bands (entries in LIS). Let us denote b_p and b_d be the bits allocated to encode the coefficients in the LL band and its descendant bands respectively, such that the total bit rate $b = b_p + b_d$. As most of the signal energy is compacted in the highest level of the pyramid (LL band) enough bits must be spent to encode the coefficients in LL band, otherwise it degrades the image quality [1]. Based on experiments, we heuristically selected $b_p = 0.2b$ to encode the coefficients in the LL band. As the number of the coefficients in the LL band is 2^{2L} times (L denotes the wavelet decomposition levels) smaller than the size of the image, allocating $b_p = 0.2b$ bits may cause an overflow, especially at high rates. To avoid this we truncate b_p to $\frac{N_1}{2^L} \times \frac{N_2}{2^L} \times (M+1)$ at high rates, so that finally $b_p = min(0.2b, \frac{N_1}{2^L} \times \frac{N_2}{2^L} \times (M+1))$. The one extra bit plane $M+1$ is accounted to encode the sign of the coefficients.

[1] We will show that besides accurate depth map, it is important to attain good image quality in order to reduce the prediction error in view synthesis.

(a) **(b) DE = 5.58**

Fig. 4. Tsukuba Image set: (a) Computed depth map d_c at rate 0.6 bpp per view using the proposed encoding scheme. (b) Disparity error w.r.t to the ground truth. The disparity error for the SPIHT scheme for the same bit rate is shown in Fig. 2 (b).

Table 1. Venus Image set: Comparison of disparity error between the proposed scheme, SPIHT, JPEG 2000 and JPEG. Results are tabulated using five wavelet decomposition levels, and Threshold (T) = 3.54.

Bit rate per view	Disparity Error (DE)			
	Proposed	JPEG 2000	SPIHT	JPEG
0.4	6.56	8.92	11.99	9.61
0.5	5.67	7.54	10.91	9.24
0.6	4.89	7.64	9.06	8.37
0.7	4.14	7.22	8.1	6.91
0.8	3.91	6.93	7.7	5.91

4. EXPERIMENTAL RESULTS

4.1. Setup

We run experiments on two stereo image sets (Tsukuba and Venus [3]) and on the Flower garden sequence. A Wavelet transform is applied on the luminance component of the image using a Daub 9/7 filter. We compute the dense depth or disparity map (d_c) using α expansion algorithm in Graph cuts [2]. The pixel similarity between the two views is measured using Birchfield Tomasi cost function [1], which presents a great advantage in reducing the image sampling errors. We evaluate the penalty for assigning different labels between the neighbors in terms of intensity gradient (called the prior or the smoothness cost), and this choice is made mainly to improve the performance of the graph cut algorithm [3].

4.2. Performance analysis

Table 1 tabulates the disparity error for the Venus image set computed from the compressed images, at various target bit rates. In our proposed scheme, we heuristically selected the $T = 3.54$ (for Venus image set), and the effect of thresholding is not studied in detail due to limited space. Clearly, we could see that the proposed scheme performs better than the standard encoding schemes. Similar observation is made for the Flower garden sequence and for the Tsukuba image set. Fig. 4 (a) shows the disparity result for the Tsukuba image set computed at $b = 0.6$ bpp per view using the proposed coding scheme, and the corresponding disparity error is shown in Fig. 4

Table 2. Tsukuba Image I_1: Comparison of PSNR between the proposed and SPIHT coding schemes, using four wavelet decomposition levels.

Bit rate (bpp)	0.4	0.5	0.6	0.7	0.8
Proposed (dB)	28.14	29.02	29.7	30.34	30.9
SPIHT (dB)	30.31	31.16	31.95	32.8	33.25

Table 3. Comparison of the prediction error in the two novel views between the proposed and SPIHT based coding schemes. The view points are rendered by warping the approximated reference image $\hat{I}_1$ encoded at 0.5 bpp, and the quality of $\hat{I}_1$ is 28.4 dB (proposed) and 30.2 dB (SPIHT).

Novel View	Proposed		SPIHT	
	PSNR	% Holes	PSNR	% Holes
1	27.7	3.85	28.8	3.6
2	25.97	5.56	26.07	5.2

(b). Comparing the Figs. 2 (b) and 4 (b) it clear that the proposed encoding scheme computes more accurate depth map, comparing to the SPIHT scheme for the same target bit rate. Especially, the depth value is recovered with high fidelity in the regions close to the weak edges (see the bottom most right portion) without penalizing significantly the depth in the background region. Furthermore, we compare the quality of the independently reconstructed views $\hat{I}_1$ and $\hat{I}_2$ in terms of PSNR, between the proposed and the SPIHT based schemes. Table 2, compares the reconstruction quality of the image $\hat{I}_1$ in the Tsukuba image set between the proposed and SPIHT based schemes. We observe that our coding scheme penalizes the PSNR by 2 - 2.5 dB on the Tsukuba image (similar findings on other test images), mainly due to the encoding of weak edges in the image. However, the proposed scheme computes accurate depth map than SPIHT based scheme, which is significantly more important than the reference image quality in several applications, e.g., view synthesis (demonstrated in the next subsection). These results allow us to conclude that estimating the geometry of the 3D scene using high quality images (measured in terms of PSNR) does not guarantee to minimize the disparity error. A comprise has to be made between the image quality and depth accuracy.

4.3. View Synthesis

We synthesis novel views by warping (forward) the approximated reference view $\hat{I}_1$ using the associated depth map d_c. Table 3 and Fig. 5 shows the prediction error for the novel views measured in terms of PSNR for the Venus image set and the Flower garden sequence, respectively. Here, the position of the synthesized view point is numbered w.r.t to the distance from the reference image with the view point 1 is closer to the reference image. In computing the prediction error, we ignore the missing pixels due to occlusion, and we tabulate the percentage of the such missing pixels separately. We observe that in spite of accurate depth map in our scheme, the quality of the rendered view is degraded comparing to the SPIHT based scheme, for the view points close to the reference image. The degradation is mainly due to the poor reconstruction quality of the reference view $\hat{I}_1$ comparing to the SPIHT based scheme, as explained in the previous section. But, when the distance between reference and synthesized view increases, we could see that the gap decreases, and even that the proposed scheme outperforms the SPIHT based scheme (see Fig. 5) due to the accurate 3D geometry when views are far apart. Interestingly, we see that the prediction error in the synthesized view depends on the quality of 3D structure and the reference image. In particular, the prediction error for the views closer to the reference camera is significantly determined by the quality of the reference image used for warping, and on the 3D structure when the distance from the reference camera increases. Thus we conclude that the information in the 3D geometry and the reference view (used for warping) should be blended together with appropriate weights while predicting the appearance of the views at various distances from the reference camera, and such a methodology would bring further improvement in the quality of the synthesized views.

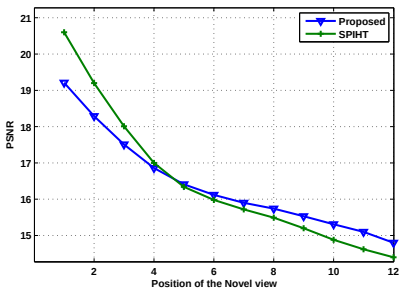

Fig. 5. Flower Garden sequence: Prediction error in the synthesized views for various novel view position. The novel view 1 is closer to the reference image and the view 12 is the farthest. The quality of the reference view $\hat{I}_1$ at 0.5 bpp is 20.6 dB and 23 dB for the proposed and SPIHT based schemes.

5. CONCLUSIONS

In this paper, we study the problem of 3D scene reconstruction in a distributed camera network, where the joint decoder computes the depth map from the approximated stereo images, encoded using medium (or low) bit rates. The proposed rate allocation scheme preserves low contrast regions in the image and is demonstrated with SPIHT coding principles. We show that comparing to the standard encoding schemes, the proposed scheme brings better 3D scene reconstruction at the cost of slightly penalizing the quality of the independently decoded image. However, we demonstrated that the quality of the reference view is less important than the 3D geometry of the scene while synthesizing novel views, in particular at a larger distance from the reference camera. Finally for optimal view synthesis, the information in the 3D geometry and the reference view should be blended with appropriate weights, and such a methodology would bring better quality in the synthesized views, that is of central importance in 3DTV and Distributed source coding applications.

6. REFERENCES

[1] S. Birchfield and C. Tomasi, "A pixel dissimilarity measure that is insensitive to image sampling," IEEE Trans. on Patt. Anal. and Mach. Intell., vol. 20(4), pp. 401-406, Apr. 1998.

[2] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," IEEE Trans. on Patt. Anal. and Mach. Intell., vol. 23(11), pp. 1222-1239, Nov. 2001.

[3] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense stereo," Intl. Journal of Comp. Vis., vol. 47(1), pp. 7-42, Apr. 2002.

[4] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec using set partitioning in hierarchical trees," IEEE Trans. on Cir. and Sys. for Video Tech., vol. 6(3), pp. 243-250, June 1996.

[5] E. Atsumi and N. Farvardin, "Lossy/lossless region-of-interest image coding based on set partitioning in hierarchical trees," Proc. IEEE Int. Conf. Image Proc., pp. 87-91, Oct. 1998.