RATE-DISTORTION OPTIMAL DEPTH MAPS IN THE WAVELET DOMAIN FOR FREE-VIEWPOINT RENDERING

Matthieu Maitre, Yoshihisa Shinagawa and Minh N. Do

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
{maitre,sinagawa,minhdo}@uiuc.edu

ABSTRACT

We consider the problem of estimating and encoding depth maps from multiple views in the context of 3D-TV with free-viewpoint rendering. We propose a novel codec based on the Rate-Distortion (RD) optimization of the Depth-Image-Based Representation (DIBR) in the wavelet domain. The rate constraint enforces the piecewise smoothness of the depth map, which improves the reliability of its estimation. We propose an efficient optimal solution for the joint estimation and coding of the depth map using dynamic programming along the tree of wavelet coefficients. It also provides an automatic bitrate allocation between images and depth maps. Experiments on real data show that the wavelet approach can improve RD performances over a state-of-the-art technique that uses quadtrees.

Index Terms— Free-viewpoint rendering, Depth-Image-Based Representation (DIBR), disparity-map estimation

1. INTRODUCTION

The success of three-dimensional (3D) video games and online 3D virtual worlds emphasizes the demand for viewing experiences beyond passively watching two-dimensional televisions. In particular, viewers look for increased interactivity inside four-dimensional (3D+t) environments. This demand is likely to be increased by the advent of autostereoscopic displays [1], which let user perceive the third dimension without wearing special glasses.

However, recording and broadcasting such 3D+t environments is still an open problem. In spite of an intense research effort in the domain of free-viewpoint 3D-TV [1], virtual worlds are still mostly synthetic, created using computer graphics tools. Indeed, allowing users to freely choose their viewpoints requires the encoding of all the possible views, i.e. the entire plenoptic function [2]. Such a massive amount of data is not compatible with the current broadcasting systems.

This issue is avoided in synthetic environments by encoding the underlying 3D geometry of the world, along with its photometric properties, and rendering views on demand at the decoder. This approach greatly reduces the amount of data to broadcast. However, it requires the knowledge of the 3D geometry, whose estimation from real data is still a difficult issue.

The Depth-Image-Based Representation (DIBR) tries to mitigate these issues by providing a trade-off between no 3D geometry and exact 3D geometry. In this representation, the plenoptic function is approximated locally by pairs of images and depth maps, arbitrary views being created on demand at the decoder using Image-Based Rendering (IBR) [2]. Since depth errors tend to be more conspicuous when the virtual viewpoint is far from the actual one, this approach can cope with approximate depth maps. At the same time, since the required number of recorded views is much reduced, the DIBR offers a compact data representation.

In this paper, we propose to study the encoding of the DIBR and the estimation of its depth maps from multiple views using a novel Rate-Distortion (RD) optimization. For simplicity, we limit the study to static DIBR with a unique pair of grayscale image and depth map. The proposed RD framework considers both the image and the depth map jointly to obtain an automatic allocation of the available bitrate between the two. The choice of the S transform, an integer version of the Haar transform [3], to represent the depth map allows us to use the rate constraint to obtain piecewise smooth depth maps and as a consequence to reduce spurious depth errors. Moreover, the S transform introduces a tree of dependencies between the wavelet coefficients, which allows an efficient solution of the RD optimization using dynamic programming [4]. Experimental results on real data confirm the efficiency of the wavelet-based smoothness and show that the RD performances of the proposed codec can outperform those obtained using quadtrees [5, 6].

Note that the problem of depth-map estimation is related to the one of motion-field estimation in video coding, where techniques based on RD-optimized quadtrees have also been proposed [7].

The remainder of the article is organized as follows. Section 2 gives an overview of the RD framework, while Section 3 presents an efficient solution and Section 4 gives an account of our experimental results.

2. RATE-DISTORTION FRAMEWORK

The encoder has access to a set of views. These views are represented by the column vectors I_s, $s \in [0, ..., N_v - 1]$, obtained by stacking all the pixels together. The view I_0, called the reference view, is the image in the DIBR. The views are assumed to be from coplanar viewpoints and to have been rectified [8], so that the motion vectors due to the motion parallax between the reference view and any other view are parallel to the baseline between these two views.

The decoder renders arbitrary views inside the plane by motion compensating the reference view using motion vectors obtained from the depth map. Since this is a forward motion compensation, it leaves holes in the rendered view which are filled using interpolation. The norm of the motion vectors is inversely proportional to the depth of the scene. It is therefore more practical to define the RD framework in terms of inverse depths, that is, disparities.

The goal of the encoder is then to find a DIBR such that the rendered views $\hat{\mathsf{I}}_s$ are as close as possible to the actual views I_s in the mean-square sense, under the constraint of the available bitrate. In order to reduce their entropy, both the reference view I_0 and the disparity map δ are encoded in the wavelet domain. We consider two different wavelet synthesis operators to obtain an efficient solution to the RD problem: a linear operator with matrix $\mathbf{T}$ for the image and an non-linear integer operator $\mathcal{T}$ for the disparity map. Let c and d be the vectors of their wavelet coefficients. We can then write

$$\hat{\mathsf{I}}_0 \triangleq \mathbf{T}\mathsf{c} \quad \text{and} \quad \delta \triangleq \mathcal{T}(\mathsf{d}). \tag{1}$$

Introducing the Lagrange multiplier λ [9], also known as RD slope, the RD optimization is given by

$$\min_{\mathsf{c},\mathsf{d}} \frac{1}{N_v} \sum_{s=0}^{N_v-1} \left\| \mathsf{I}_s - \mathcal{M}_s^f(\mathbf{T}\mathsf{c}; \mathcal{T}(\mathsf{d})) \right\|_2^2 + \lambda \left(R(\mathsf{c}) + R(\mathsf{d}) \right), \tag{2}$$

where N_v denotes the number of views, $\|.\|_2^2$ the norm 2, $R(.)$ the bitrate and $\mathcal{M}_s^f(\hat{\mathsf{I}}_0; \delta)$ the forward motion compensation that transforms the encoded reference view $\hat{\mathsf{I}}_0$ into the rendered view $\hat{\mathsf{I}}_s$ using the disparity map δ.

Ignoring the issues of occlusions and resampling, the Mean-Square Error (MSE) term can be defined either in terms of backward motion compensation or forward motion compensation. The latter is more practical since it decouples the encoded reference view from the motion compensation. Using this approximation, the optimization becomes

$$\min_{\mathsf{c},\mathsf{d}} \frac{1}{N_v} \sum_{s=0}^{N_v-1} \left\| \mathcal{M}_s^b(\mathsf{I}_s; \mathcal{T}(\mathsf{d})) - \mathbf{T}\mathsf{c} \right\|_2^2 + \lambda \left(R(\mathsf{c}) + R(\mathsf{d}) \right), \tag{3}$$

where $\mathcal{M}_s^b(\mathsf{I}_s; \delta)$ denotes the backward motion compensation that transforms the rendered view $\hat{\mathsf{I}}_s$ into the encoded reference view $\hat{\mathsf{I}}_0$ using the disparity map δ.

3. EFFICIENT OPTIMIZATION

3.1. Overview

The MSE term of (3) depends on the wavelet vectors c and d in very different ways: it is quadratic in c but non-linear in d. Therefore, the problem is solved using successive optimizations, first optimizing c and then d. This way, we can design optimization techniques adapted to each case. The optimization is initialized at high bitrate where the MSE is virtually null, that is, $\mathbf{T}\mathsf{c} \approx \mathsf{I}_0$ and $\mathcal{M}_s^b(\mathsf{I}_s; \mathcal{T}(\mathsf{d})) \approx \mathsf{I}_0$.

3.2. Reference view

Assuming that the wavelet transform $\mathbf{T}$ is nearly orthonormal, like the 9/7 wavelet used in JPEG 2000 [3] for instance, the optimization (3) with regard to the image wavelet coefficients c can be rewritten as

$$\min_{\mathsf{c}} \left\| \mathbf{T}^{-1}\mathsf{I}_0 - \mathsf{c} \right\|_2^2 + \lambda R(\mathsf{c}), \tag{4}$$

for which JPEG2000 provides a near-optimal solution [3].

3.3. Disparity map

For simplicity, we present the case of one-dimensional disparity maps. The procedure can be generalized to two-dimensional maps by applying it alternatively along the horizontal and vertical axes.

Since motion compensation is a non-linear function of the disparity map, we cannot rely on the wavelet transform being nearly orthonormal to simplify the problem. Instead, we take advantage of the fact that quantized disparity maps take only a finite number of disparity values. We then choose an integer wavelet transform and a rate model whose graph of dependencies is a tree, as shown in Figure 1.

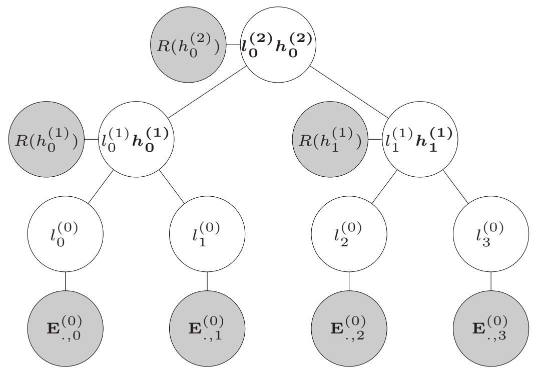


Fig. 1. Dependency graph of a two-level S transform. The coefficients in bold are those included in the wavelet vector d. Gray nodes represent the MSE and rate terms of the RD optimization.

Disparity maps tend to be piecewise constant, so the Haar wavelet transform is suitable to provide a compact representation. Since quantized disparities are discrete, we choose an integer version of the Haar transform, that is, the S transform [3]. It is defined as follows. Let $l_n^{(j)}$ and $h_n^{(j)}$ be two integer coefficients of the S transform at level j, respectively low-pass and high-pass. Let $l_{2n}^{(j-1)}$ and $l_{2n+1}^{(j-1)}$ be two low-pass integer coefficients at the next finer level $j-1$. The analysis operator of the S transform relates these quantities by

$$\begin{cases} l_n^{(j)} = \left\lfloor \dfrac{l_{2n}^{(j-1)} + l_{2n+1}^{(j-1)}}{2} \right\rfloor \\ h_n^{(j)} = l_{2n}^{(j-1)} - l_{2n+1}^{(j-1)} \end{cases} \tag{5}$$

where $\lfloor x \rfloor$ denotes the largest integer less or equal to x. At the finest level, the high-pass coefficients are not defined and the low-pass coefficients are equal to the disparity map, that is, $l^{(0)} = \delta$. The wavelet vector d is made of the low-pass coefficients at the coarsest level and all the high-pass coefficients.

For the rate model, we assume that all the wavelet coefficients are independent of one another, the low-pass coefficients l at the coarsest level following a uniform distribution and the high-pass coefficients h following a discrete truncated Laplace distribution with zero mean and scale parameter b. Therefore, the rate (in bits) is approximated by

$$R(\mathsf{d}) = \frac{1}{b \log 2} \sum_{j=1}^{L} \sum_{n=0}^{N_n(j)-1} |h_n^{(j)}| + cst \tag{6}$$

where cst is a term independent of d, L is the index of the coarsest level and $N_n(j)$ the number of high-pass coefficients at level j.

Let the quantized disparity map take integer values in the range $[0, N_\delta - 1]$. The backward motion compensation of view I_s corresponding to the pixel n of $\hat{\mathsf{I}}_0$ when its disparity is m takes the simple form

$$\mathcal{M}_{s,n}^b(\mathsf{I}_s; m) = \mathsf{I}_{s,n+\alpha_s m+\beta_s}, \tag{7}$$

where α_s and β_s depend on the camera parameters of the views.

For a fixed c, and thus a fixed $\hat{\mathsf{I}}_0 \triangleq \mathbf{T}\mathsf{c}$, Equation (3) can be written as

$$\min_{\mathsf{d}} \frac{1}{N_v} \sum_{s=0}^{N_v-1} \sum_{n=0}^{N_n-1} \left(\mathcal{M}_{s,n}^b(\mathsf{I}_s; \delta_n) - \hat{\mathsf{I}}_{0,n} \right)^2 + \lambda R(\mathsf{d}), \tag{8}$$

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix}$$

(a) Rate matrix $\mathbf{R}$

$$\begin{array}{cc} & \begin{pmatrix} 10 & 15 & 8 & 0 & 14 \end{pmatrix} \\ \begin{pmatrix} 5 \\ 8 \\ 0 \\ 12 \\ 20 \end{pmatrix} & \begin{pmatrix} 15 & 21 & 15 & 8 & 23 \\ 19 & 23 & 17 & 10 & 25 \\ 12 & 16 & 8 & 1 & 16 \\ 25 & 29 & 21 & 12 & 27 \\ 34 & 38 & 30 & 21 & 34 \end{pmatrix} \end{array}$$

(b) Error matrix $\mathbf{J}$ and error vectors $\mathbf{E}^{(0)}_{\cdot,0}$ and $\mathbf{E}^{(0)t}_{\cdot,1}$

Fig. 2. Example of node-wise minimization for $l_0^{(1)} = 2$ with $N_\delta = 5$ and $\mu = 1$. The error matrix $\mathbf{J}$ is the sum of the error vectors and the rate matrix. The minimum is searched for among the values in the gray boxes and is found in the dark-gray box.

where N_n is the number of pixels.

The MSE term can be factorized into a product of two matrices: an error matrix $\mathbf{E}^{(0)}$, which shall serve as initialization of the dynamic programming presented in the following section, and a selection matrix $\mathbf{S}(\delta)$. The entry $\mathbf{E}^{(0)}_{m,n}$ of the error matrix gives the square error that the pixel n of $\hat{\mathsf{I}}_0$ would be associated with if it had disparity m. That is,

$$\mathbf{E}^{(0)}_{m,n} \triangleq \frac{1}{N_v} \sum_{s=0}^{N_v-1} \left(\mathcal{M}^b_{s,n}(\mathsf{I}_s; m) - \hat{\mathsf{I}}_{0,n} \right)^2. \tag{9}$$

This error matrix is also called "disparity space image" [8] and is independent of the disparity map δ. The selection matrix $\mathbf{S}(\delta)$ is made only of zeros and ones with exactly one one along each row. The locations of the ones are given by the values of the disparity map δ. The optimization (3) with regard to the disparity wavelet coefficients d can then be written as

$$\min_{\mathsf{d}} \operatorname{tr}(\mathbf{S}(\mathcal{T}(\mathsf{d}))\mathbf{E}^{(0)}) + \mu \sum_{j=1}^{L} \sum_{n=0}^{N_n(j)-1} |h_n^{(j)}(\mathsf{d})| \tag{10}$$

where tr denotes the trace operator and $\mu \triangleq \lambda/(b \log 2)$.

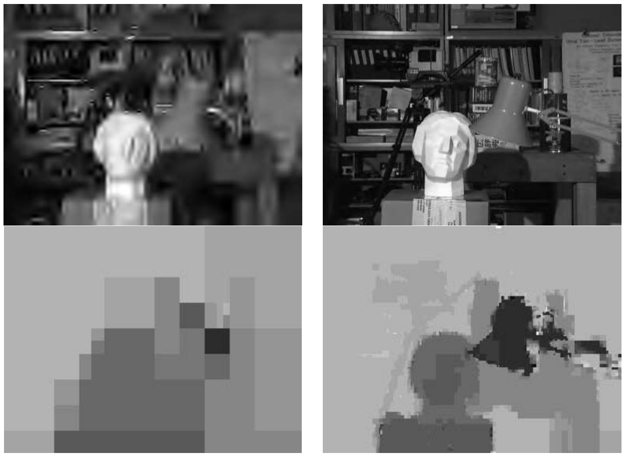
3.4. Dynamic programming

Since the graph does not contain loops, the optimal solution can be efficiently computed by grouping the terms of the summation in (10) such that the large minimization becomes a recursion of small minimizations. This dynamic-programming approach consists of two passes: one bottom-up which performs minimizations at each node of the graph, followed by one top-down which backtracks through the node-wise minimizations to find the globally optimal solution.

Let us define a rate matrix $\mathbf{R}_{n,m} \triangleq \mu|n-m|$, as shown in Figure 2(a). At level j, the error matrix $\mathbf{E}^{(j)}$ is known and the error matrix $\mathbf{E}^{(j+1)}$ at the coarser level is calculated. A matrix $\mathbf{H}^{(j+1)}$ of high-pass coefficients is also computed to prepare the backtracking.

At the node connecting the low-pass coefficients $l_{2n}^{(j)}$ and $l_{2n+1}^{(j)}$, the algorithm creates a node-wise error-matrix $\mathbf{J}$ such that

$$\mathbf{J}_{l_{2n}^{(j)}, l_{2n+1}^{(j)}} \triangleq \mathbf{E}^{(j)}_{l_{2n}^{(j)}, 2n} + \mathbf{E}^{(j)}_{l_{2n+1}^{(j)}, 2n+1} + \mathbf{R}_{l_{2n}^{(j)}, l_{2n+1}^{(j)}}. \tag{11}$$



(a) $\lambda = 5.9 \times 10^{-2}$

(b) $\lambda = 6.5 \times 10^{-4}$

Fig. 3. DIBR of the Tsukuba set estimated at two RD slopes. The bitrate constraint limits spurious noise during the disparity estimation. The quality of both the image and the disparity map varies according to the available bitrate.

For each value m of the low-pass coefficient $l_n^{(j+1)}$ at the coarser level, it performs the minimization

$$\min_{l_{2n}^{(j)}, l_{2n+1}^{(j)}} \mathbf{J}_{l_{2n}^{(j)}, l_{2n+1}^{(j)}} \text{ such that } \left\lfloor \frac{l_{2n}^{(j)} + l_{2n+1}^{(j)}}{2} \right\rfloor = m, \tag{12}$$

as shown in Figure 2(b). The value of the minimum is stored in the entry $\mathbf{E}^{(j+1)}_{m,n}$ of the error matrix at the coarser level. The coefficients $l_{2n}^{(j)*}$ and $l_{2n+1}^{(j)*}$ achieving this minimum give the optimal value $\mathbf{H}^{(j+1)}_{m,n}$ that $h_n^{(j+1)}$ would take if the optimal value of $l_n^{(j+1)}$ was m.

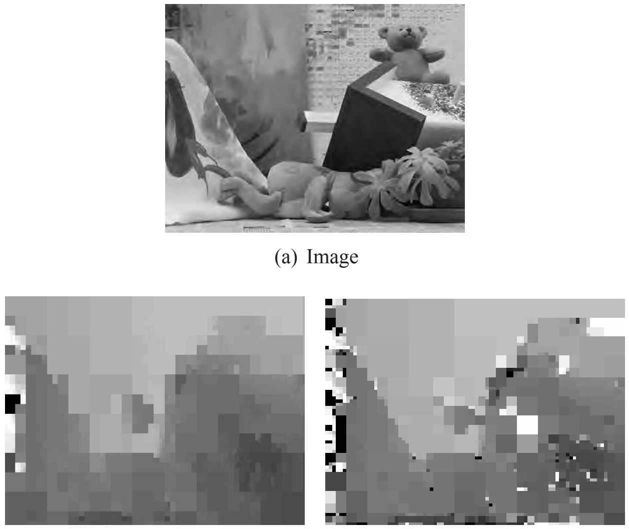
This process is repeated until the coarsest level L is reached. At this point, the optimal low-pass coefficients $l_n^{(L)*}$ are obtained by selecting the column-wise minimima of the error matrix $\mathbf{E}^{(L)}$. This starts the backtracking of the top-down pass. At level j, the optimal high-pass coefficients are given by $h_n^{(j)*} = \mathbf{H}^{(j)}_{l_n^{(j)*}, n}$. The synthesis operator of the S transform then gives the optimal low-pass coefficients $l_{2n}^{(j-1)*}$ and $l_{2n+1}^{(j-1)*}$ at the finer level. This procedure is repeated until the finest level is reached.

It remains to estimate the optimal scale factor b. It is obtained using a procedure akin to dichotomy. The procedure starts with a given range for μ and an initial value of μ. It then iteratively finds the optimal vector d, then the optimal scale factor b in the Kullback-Leibler sense and finally the actual Lagrange multiplier λ. Each time, the smoothness μ is adapted so that eventually the optimizations of both the image and the disparity map are performed using the same Lagrange multiplier λ.

4. EXPERIMENTAL RESULTS

We present experimental results on two image sets, Tsukuba and Teddy [8], displayed in Figures 3 and 4. Experiments were run in the grayscale domain with intensity values in the range $[0, 1]$. Nine views were used from the Tsukuba set and two from the Teddy set.

The proposed wavelet codec is compared to a codec based on quadtrees [5–7]. Both codecs rely on the Kakadu JPEG2000 codec [3]



(a) Image

(b) Disparity map – Wavelets

(c) Disparity map – Quadtree

Fig. 4. DIBR of the Teddy sequence at $\lambda = 4.5 \times 10^{-3}$: image (a) and disparity maps estimated using wavelets (b) or quadtrees (c). Unlike quadtrees, wavelets enforce inter-block smoothness which limits spurious noise.

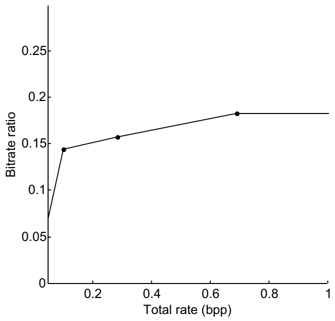


Fig. 5. Bitrate allocation on the Teddy set. The disparity map consistently represents around 15% of the total bitrate.

to code the reference view and use therefore the same error matrix **E**. They differ by their disparity model. Both favor variable-size blocks with constant disparities but wavelets also favor inter-block smoothness, while quadtrees do not. This explains why the disparity map in Figure 4 obtained using quadtrees contains much more spurious noise. Reduced noise allows the rendering of novel views further away from the reference view, therefore reducing the number of pairs of reference views and depth maps needed to represent the plenoptic function. Since neither quadtrees nor wavelets model occlusions, both have issues in regions where they happen, like on the left of the image or around the chimney for instance.

The improvement in disparity-map estimation brought by wavelets translates in improved RD performances with improvements of up-to 1.3dB, as shown in Figure 6. Finally, Figure 3 shows the automatic bitrate allocation: as the bitrate is reduced, the quality of both the reference view and the depth map is reduced. The allocation is actually fairly stable across the range of bitrates with around 15% dedicated to the disparity map, as shown in Figure 5.

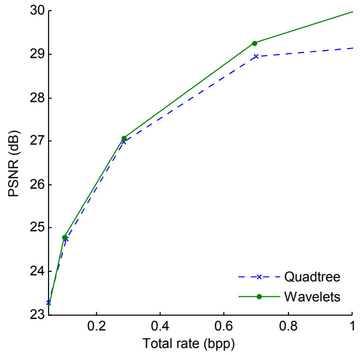


Fig. 6. RD performances of the DIBR codecs using either wavelets or quadtrees on the Teddy set. The wavelet-based codec offers improvements of up to 1.3dB.

5. CONCLUSION

We have proposed a novel codec of the depth-image-based representation based on rate-distortion optimization. Using wavelets lets the encoder enforce the piecewise smoothness of the disparity map, which reduces spurious noise. The optimization is efficiently solved using dynamic programming and provides an automatic bitrate allocation. The experiments on real data show a PSNR gain of up-to 1.3dB and a bitrate allocation stable across the range of bitrates with around 15% dedicated to the depth map.

6. REFERENCES

[1] C. Fehn, R. Barre, and R. S. Pastoor, “Interactive 3-D TV – concepts and key technologies,” *Proc. of the IEEE*, vol. 94, no. 3, pp. 524–538, 2006.

[2] H.-Y. Shum, S.-C. Chan, and S.B. Kang, *Image-Based Rendering*, Springer-Verlag, 2007.

[3] D. Taubman and M. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*, Springer-Verlag, 2001.

[4] D.P. Bertsekas, *Dynamic Programming and Optimal Control*, Athena scientific, 2005.

[5] J.N. Ellinas and M.S. Sangriotis, “Stereo video coding based on quad-tree decomposition of B-P frames by motion and disparity interpolation,” *IEE Proc.-Vis. Im. Sig. Proc.*, vol. 152, no. 5, pp. 639–647, 2005.

[6] J. D. Oh and R.-H. Park, “Reconstruction of intermediate views from stereoscopic images using disparity vectors estimated by the geometrical constraint,” *IEEE Trans. on Circuits and Systems for Video Tech.*, vol. 16, pp. 638–641, 2006.

[7] G. J. Sullivan and R. L. Baker, “Efficient quadtree coding of images and video,” *IEEE Trans. on Image Proc.*, vol. 3, pp. 327–331, 1994.

[8] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *Int. J. of Comp. Vis*, vol. 47, no. 1–3, pp. 7–42, 2002.

[9] G.J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression,” *IEEE Signal Proc. Mag.*, pp. 74–90, 1998.