Published in the Proc. of the Symposium on Application of subbands and wavelets , March 18, 1994

Multiresolution based hierarchical disparity estimation for stereo image pair compression

[**]Sriram Sethuraman, [*]Angel G.Jordan, [*]M.W.Siegel,
[**]Department of Electrical and Computer Engineering,
[*]The Robotics Institute / School of Computer Science,
Carnegie Mellon University, Pittsburgh, PA 15213

I Introduction

Stereo vision is the process of viewing two different perspective projections of the same real world scene and perceiving the depth that was present in the original scene. These projections offer a compact 2-dimensional means of representing a 3-dimensional scene, as seen by one observer. Different display schemes have been developed to ensure that each eye sees the image that is intended for it. Each image in the image pair is referred to as the left or right image depending on the eye it is intended for. The binocular cues contain unambiguous information in contrast to monocular cues like shading or coloring. Hence binocular stereo may be quite useful, for instance, in video based training of personnel. On the entertainment side, it can make mundane TV material lively. Though the concept has been around for more than half a century, only recently have technically effective ways of making stereoscopic displays and the usually required eyeware emerged. Despite this progress, stereo TV can be made a cost effective add-on option only if the increased bandwidth requirement is relaxed somehow. Since the two images are projections of the same scene from two nearby points of view, they are bound to have a lot of redundancy between them. By properly exploiting this redundancy, the two image streams might be compressed and transmitted through a single monocular channel's bandwidth.

The first step towards stereoscopic image sequence compression is 'still' stereo image pair compression that exploits the high correlation between the left and right images, in addition to exploiting the spatial correlation within each image. The temporal correlation between the frames can be taken advantage of, along the lines of the MPEG (Motion Picture Experts Group) standards, to achieve further compression. The final step would be to explore the correlation between left and right frames with a time offset between them.

In this paper a multiresolution based approach is proposed for compressing 'still' stereo image pairs. In Section II the task at hand is contrasted with the stereo disparity estimation problem in the machine vision community; a block based scheme on the lines of a motion estimation scheme is suggested as a possible approach. In Section III, the suitability of hierarchical techniques for disparity estimation is outlined. Section IV provides an overview of wavelet decomposition. Section V details the multiresolution approach taken. In section VI, the typical computational gains and compression ratios possible with this scheme are computed. Subjective and objective evaluations of several different compressed stereo image pairs highlight the efficacy of the proposed compression scheme. Possible extensions of this approach to stereo image sequence compression are discussed in the last section.

II Stereo disparity and estimation

The problem of finding two points in a stereo image pair that correspond to the same point in the 3D scene is called the *correspondence* problem. The distance between the two corresponding points when the two images are aligned one on top of the other is called the *disparity*[1]. The correspondence problem or the disparity estimation problem has been studied extensively by the computer vision community. The motivation there is that the disparity along with the knowledge of the camera geometry can be used to calculate the depth of each point; and depth

[1] The actual disparity depends on the camera geometry and on parameters such as interocular separation and convergence distance.

This research was supported by the Advanced Research Projects Agency under ARPA Grant No.MDA 972-92-J-1010.

information is essential, e.g., for collision avoidance in autonomous navigation. Hence the thrust in computer vision is to estimate *dense* depth maps which means estimating disparity for every pixel in the image. Computationally efficient and cost effective methods for doing this in real time are yet to come. Nor will, dense depth maps lead to compression. But fortunately the human eye is quite good in filling in missing information, so that *sparse* depth maps can effectively convey the desired stereo cues. Hence in applications where the stereo pairs are meant for human viewing rather than computer ranging, computing sparse depth maps is the key to compression

Assuming the disparity to be not varying much among adjacent pixels, a block-by-block disparity estimation can be carried out instead of a point-by-point estimation. This method is similar to the motion estimation scheme used in image sequence coding. For every block in an image, the best matching block within a neighborhood can be searched for in the other image based on some fidelity criterion like the maximum cross correlation (MCC), minimum mean squared error (MMSE) or the minimum absolute difference (MAD).

III Hierarchical estimation of disparity

Decomposition of an image at progressively lower resolutions is practically attractive because of the compatibility it provides across the range of envisioned transmission schemes, from Super HDTV to NTSC. With increasing resolution, the disparity (in pixels) between corresponding points increases. In a well composed stereo image, the vertical disparity can be made zero; nevertheless, even an exhaustive one-dimensional search in the maximum fusable horizontal disparity range would become a computationally formidable task at higher resolutions. Hence from a computational standpoint also, a hierarchical scheme that estimates on a coarse-to-fine scale is preferable.

The multiresolution theory of Mallat [5], unifies the concepts of subband coding and wavelet theory and also provides a psychophysical justification for operating at different scales. The wavelet decomposition of images provides an efficient way of representing information in an image, and the decomposed subimages can be quantized and coded very efficiently [1,2,5,11]. The context (significant objects in the scene) of the image is visible at all resolution levels; these progressively lower resolution images form a *pyramid* structure [3] that can be used for the hierarchical block disparity estimation. The quantized and reconstructed image retains most of the edge information crucial for stereo perception [6]. On the other hand, transform based compression methods that achieve compression by smoothing the image do not preserve the sharpness of the edges. Also, the subband coding method does not suffer from ‘tiling’ at low bit rates, which is a serious drawback in transform based methods that operate on blocks [9]. Keeping these factors in mind, a multiresolution based hierarchical disparity estimation scheme is considered here.

IV Wavelet decomposition

The wavelet decomposition of an image can be achieved using the equivalent filtering approach[5]. Figure.1.a shows a three level decomposition. The separable 2D filtering approach is used wherein the filtering is done along the rows and then along the columns. Daubechies’ compactly supported orthonormal filters [4] have been used to reduce the filtering complexity. It has been observed [12] that there is no appreciable improvement in using filter orders above six. Hence DAUB6 filters have been used for all the images in our experiments. The orthonormality of the filters permit downsampling by a factor of two, after filtering. Thus this decomposition preserves the total number of pixels after decomposition, unlike the Laplacian pyramid structure [3] which results in an one-third increase in the number of pixels. At each level there are four subimages that are one-fourth the size of the image at the level below.

The subimage obtained by lowpass filtering in both directions is called the *coarse* image; typically, it carries most of the energy present in the original image and has an intensity distribution similar to that of the original image. The other three subimages are called the *detail* images, because they contain the high frequency details in either the horizontal or vertical direction or both. *Only* the coarse image at each level is subjected to further decomposition to obtain the next level subimages.The detail images have an almost Laplacian intensity distribution ($Ae^{-\lambda|I|}$) [5], i.e, there are more pixels with intensities close to zero and only a few pixels with significantly high absolute intensities. Hence these detail images can be either scalar or vector quantized efficiently at very low bits per pixel (bpp) without any visually displeasing artifacts. The bit allocation for the detail image pixels increases at higher decomposition levels, due to the higher fraction of total energy usually contained in these levels. Compression ratios as high as 1:100 have been reported using entropy constrained lattice vector quantization of the detail images [2].

Figure.1.b shows the reconstruction process. The samples dropped during subsampling are filled with zeros and the synthesis filtering is carried out. The interpolated and filtered coarse and detail images at one level are summed to obtain the coarse image at the next lower level. In the absence of quantization, the reconstruction is perfect.

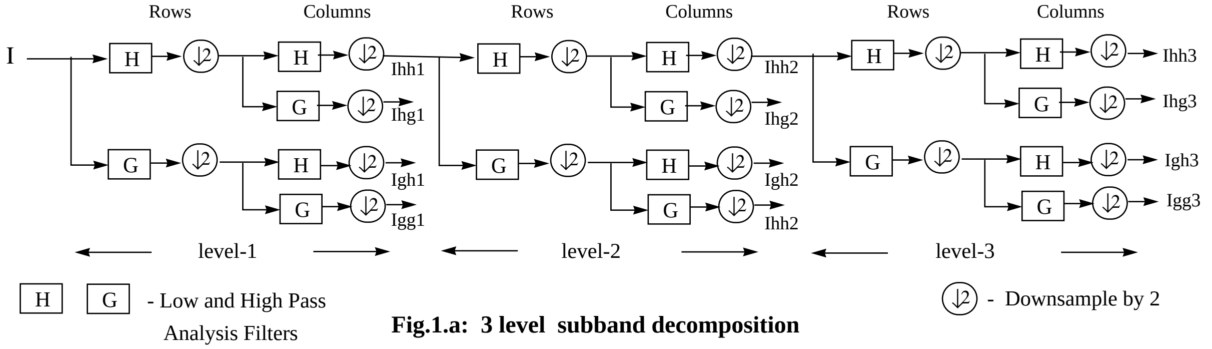


Fig.1.a: 3 level subband decomposition

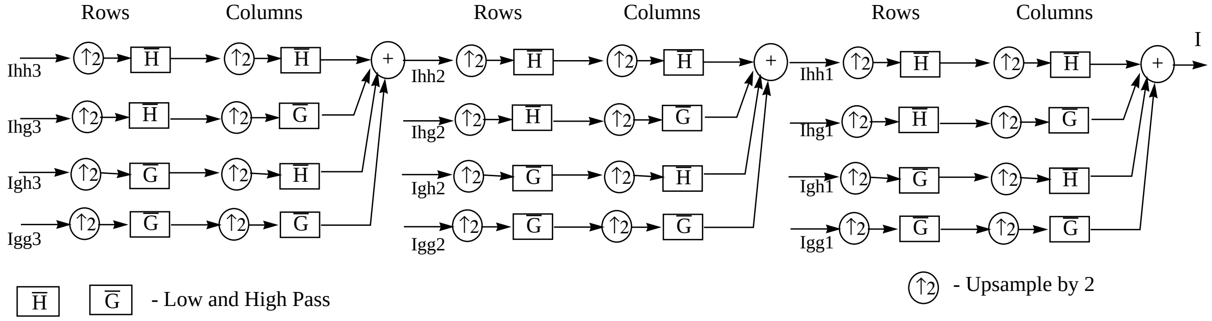


Fig.1.b: 3 level reconstruction

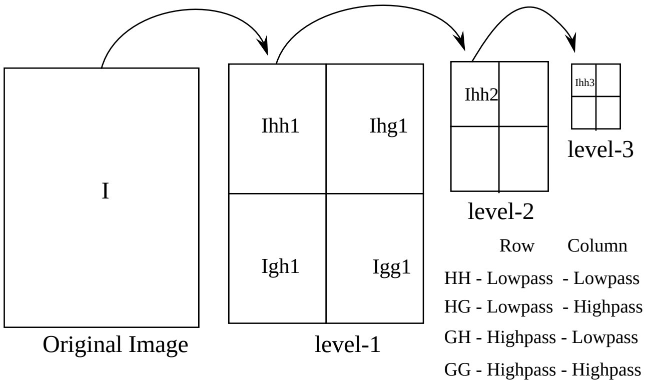


Fig.1.c: Multiresolution decomposition

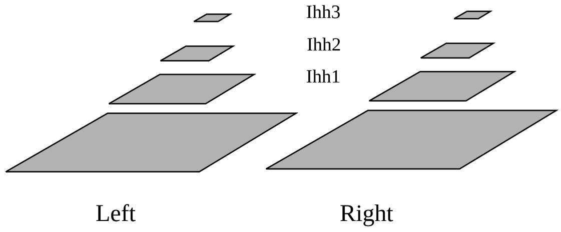
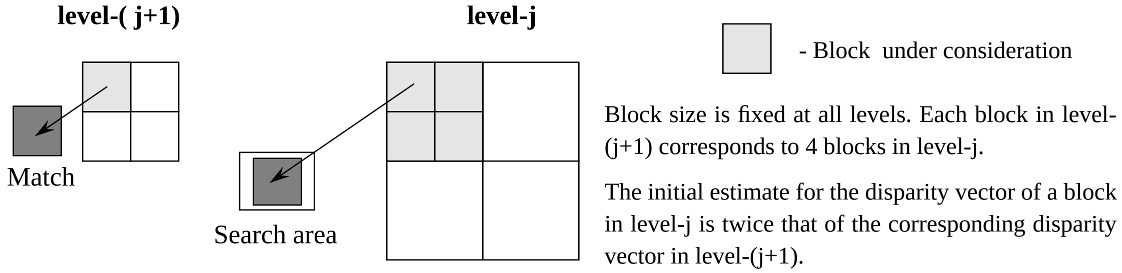


Fig.2: Pyramid for hierarchical block matching

V Hierarchical stereo matching

The left and right images are decomposed to level-3 and the two pyramids for hierarchical block matching are constructed using the coarse images at the different levels, as shown in Fig.2. The maximum horizontal disparity (HDMAX) is estimated (at present manually). The maximum vertical disparity (VDMAX) is within 3-4 pixels for reasonably composed image pairs. The search neighborhood in the left image[1] at level-n in one direction will be (maximum disparity in that direction)/2^n pixels (since disparity decreases with decreasing resolution). The minimum absolute difference (MAD) criterion is used to find the best matching block in the search neighborhood for every block in the right image.

Each block in level-3 corresponds to four blocks in level-2, 16 blocks in level-1 and 64 blocks in the original image, i.e., the block size (in pixels) remains a constant at all levels of resolution. The disparity vector for a block at level-3 is doubled and used as the initial estimate for the four corresponding blocks at level-2. The search neighborhood is formed around this initial estimate, and the block matching is carried out at level-2. Figure 3 illustrates this hierarchical block matching. This coarse-to-fine disparity estimation is carried out up to level-1. Instead of going down to level-0, the same disparities are applied to the corresponding blocks in all the four subimages at level-1. Given the disparities of the blocks at level-1 and the left image, the right image can be estimated by reconstructing from these four subimages. A schematic of this approach is shown in Fig.4. Since the four subimages at a level represent the same scene, applying the same disparity to all the four is valid. However, the maximum error in disparity is now ±2 pixels. The advantage of such estimation and its artifacts, when the left and estimated right images are viewed stereoscopically, are discussed in the next section.



Block size is fixed at all levels. Each block in level-(j+1) corresponds to 4 blocks in level-j.

The initial estimate for the disparity vector of a block in level-j is twice that of the corresponding disparity vector in level-(j+1).

Fig.3: Hierarchical disparity estimation

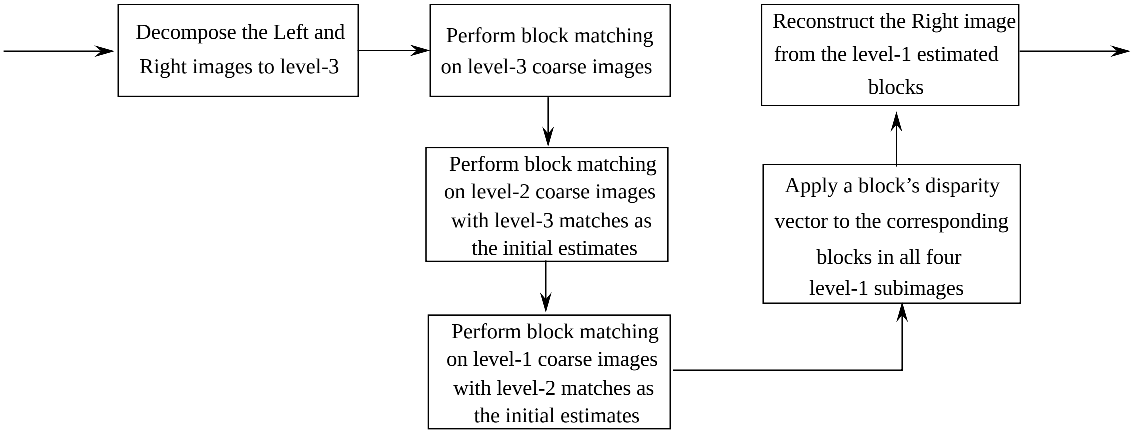

Decompose the Left and Right images to level-3
Perform block matching on level-3 coarse images
Perform block matching on level-2 coarse images with level-3 matches as the initial estimates
Perform block matching on level-1 coarse images with level-2 matches as the initial estimates
Apply a block's disparity vector to the corresponding blocks in all four level-1 subimages
Reconstruct the Right image from the level-1 estimated blocks

Fig.4: Schematic of the proposed algorithm
(for estimating the right image given the left image)

[1] One image in the image pair can be estimated given the other. Here it is assumed that the right is estimated given the left

VI Significant gains of the proposed approach

The following calculations elicit the computational simplicity of this method over an exhaustive search, and estimate the possible range of compression. Let the image size be (M x N) pixels at 8 bpp. A typical[1] block size of 6x4 is assumed; so the number of blocks (B) is (M/6)x(N/4). Assuming HDMAX to be 32 pixels and VDMAX to be three pixels,

Computational gain:

Compared to (32+1)x(3+1) = 132 searches for each of the 'B' blocks in the right image for an exhaustive search, following are the typical[2] number of searches using the proposed approach at the different levels.

level-3 - $(1+32/2^3)\text{x}1$ = 5 searches/block, B/64 blocks

level-2 - 4 (typical) x (1 top or bottom + 1) searches/block, B/16 blocks

level-1 - 6 (typical) x (1 top + 1 bottom + 1) searches/block, B/4 blocks

A total of (5/64 + 8/16 + 18/4) ≈ 5 searches for each of the B blocks, which amounts to a *factor of 26* decrease in the computational complexity.

Compression gain calculations:

Since HDMAX is 16 and VDMAX is 2 at level-1, the number of bits needed to represent the disparities (for the typical number of searches given above) will be, $(\log_2 16 + \log_2 2)B/4 = 1.25$ B bits. Thus to code the right image, in addition to the left image, only 1.25 B extra bits are needed. The compression factor (C) for the right image given the left image is (recalling that B = MN/24):

$$C = \frac{8MN}{(1.25MN)/24} \cong 153$$

If the left image at 8 bpp is compressed to 0.5 bpp (e.g., by wavelet coefficient quantization techniques [1]), then the excess bits required as a fraction of the coded left image will be, $B_e = (8/0.5)/153 \approx 10\%$. Since the disparity estimated from a higher level is used as the initial estimate for four blocks in the next lower level, the disparities at the different levels constitute a *reduced mean pyramid* [9,11]. This leads to an efficient coding of the disparity vectors.

The reconstruction of the right image from the block shifts applied to all the subimages at level-1 has a two-fold advantage. First, the number of blocks at level-1 is one-fourth the number at level-0. Second, the horizontal disparity range at level-1 is HDMAX/2. However, both these features have their demerits. The former might lead to visible blocking effects because the block size at level 0 is 12x8, instead of 6x4. For instance, anomalous large planar patches might appear when the reconstructed image pairs are viewed stereoscopically. Estimating the disparity only up to level-1 would result in larger errors in the disparity estimation.

It might be argued that the proposed approach is similar to choosing larger block sizes and performing a less precise block match at level-0. This is not true because the subband decomposition allows a controlled way of doing the match. Moreover, since most of the high frequency information is lost when the left image is compressed, the reconstructed left image is closer to the interpolated level-1 coarse image. Also, the interpolation smooths across the block boundaries, minimizing the patchiness due to block matching, which is not possible in level-0 matching. If large errors in disparity cannot be tolerated in a specific application, a subpixel matching can be done at level-1 using the samples of the level-1 coarse image before downsampling by two.

[1] The block size is chosen to strike a balance between the visible blocking effects due to increasing block sizes and the reduction in the compression factor with decreasing block sizes.

[2] The choice of the search neighborhoods at level-2 and -3 were made as a trade-off between the disparity estimation error and the computational complexity that a particular choice of neighborhood leads to.

An issue that has been ignored so far is that of *occlusions*: there exist some points in one image that have no correspondence in the other image. A prediction based scheme cannot detect such points. Under the intensity based block matching methods, gross errors due to occluded regions can be detected by thresholding the mismatch error of each block. If the mismatch error within the search neighborhood is above a threshold, then that block can be separately coded and transmitted.

VII Results

Seven image pairs with different types of scenes were chosen for evaluating the performance of the proposed algorithm. Table 1 shows the typical search neighborhoods and the block sizes used for disparity estimation for each of these images. The compression factor C that can be achieved for these neighborhoods is also listed in Table 1.

The left images were compressed using wavelet coefficient quantization techniques to approximately 0.5 bpp. The root mean square errors (RMSE) of the right image reconstructed from the uncompressed left image and the compressed left image are listed in Table 2. The uncompensated error term in the table is a measure of the disparity between the left and right images. It can be seen that the right image reconstructed from the compressed left image has an error close to that of the right image reconstructed from the original left image.

All these reconstructed image pairs were viewed stereoscopically to identify any visual artifacts that may be present. Subjectively, the compressed image pairs still have strong stereoscopic cues to stimulate the perception of depth. No patchiness was noticeable. However, spurious points appear at different depths due to reconstruction errors.

Figures 5 and 6 illustrate the results of estimation on two image pairs, named ‘Lake’ and ‘Juggler’.

VIII Conclusions and future work

The proposed stereo image pair compression algorithm is a first step towards stereo image sequence compression, which in turn could make stereoscopy available as a low cost add-on option on televisions. It has been shown that the multiresolution based hierarchical estimation of disparity vectors yields high compression factors while being computationally efficient as well. Subjective evaluations of stereo image pairs, compressed using the algorithm, indicate good performance.

Efforts are currently being made to achieve higher compressions than that are possible by MPEG coding of the two streams separately. The disparity estimation across the streams would augment the MPEG-type prediction and interpolation of frames to achieve this. The bidirectional interpolation would also help to reduce the occlusion overheads. Means for reducing other visual artifacts due to compression and reconstruction remain to be evaluated

References

[1] M.Antonini, et al, ‘Image coding using vector quantization in the wavelet transform domain', Proceedings of ICASSP, pp.2297-2300, 1990.

[2] M.Antonini, et al, ‘Image coding using lattice vector quantization of wavelet coefficients’, Proceedings of ICASSP, pp.2273-2276, 1991.

[3] P.J.Burt and E.H.Adelson, ‘The Laplacian Pyramid as a compact image code', IEEE Transactions on Communications, Vol.31, No.4, April 1983.

[4] Ingrid Daubechies, ‘Orthonormal bases of compactly supported wavelets', Communications of Pure and Applied Mathematics, vol.41, pp.909-996, 1988.

[5] Stephane G. Mallat, ‘A theory for multiresolution signal decomposition: The wavelet representation', IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.II, No.7, July 1989.

[6] David Marr, ‘Vision’, W.H.Freeman and Co., NewYork, 1982.

[7] M.G.Perkins, ‘Data compression of stereopairs’, IEEE Transactions on Communications, Vol.40, No.4, April 1992.

[8] A.Rosenfeld, (Ed.), ‘Multiresolution image processing and analysis', Springer Verlag, 1984.

[9] M.Rabbani and P.W.Jones, ‘Digital image compression techniques’, SPIE Optical Engineering Press, Bellingham, Washington, 1991.

[10] A.Tamtaoui and C.Labit, ‘Constrained disparity and motion estimators for 3DTV image sequence coding', Signal Processing: Image Communication, Vol.4, 1991.

[11] Q.Wang and R.J.Clarke, ‘Motion estimation and compensation for image sequence coding’, Signal Processing: Image Communication, Vol. 4, 1992.

[12] W.R.Zettler, et al, ‘Application of compactly supported wavelets to image compression’, SPIE Vol.1244, Image Processing Algorithms and Techniques, pp.150-160, 1990. .

Table 1: Search neighborhoods for different imagess

Image	Image Size MxN	HD MAX	VD MAX	Block Size	Search Neighborhood Level 3 T B L R	 Level 2 T B L R	 Level 1 T B L R	Compression factor for Right image
Crowd	420x280	18	2	8x4	0 0 3 0	0 0 4 3	1 1 3 4	208
Lake	549x329	7	2	8x4	0 0 1 0	0 0 2 1	1 1 1 2	256
Grand Canyon	545x380	25	3	7x7	0 0 3 0	0 0 4 3	1 1 3 4	320
Juggler	360x240	16	2	5x5	0 0 0 0	0 0 0 0	0 1 9 0	160
Clubs	382x287	14	4	4x8	0 0 0 0	0 0 0 0	0 2 9 0	195
Mt St Helens	512x512	10	1	6x6	0 0 0 0	0 1 3 0	1 1 4 3	256
Molecule	322x292	35	2	8x4	0 0 1 0	0 1 4 3	1 1 3 4	195

HD - Horizontal Disparity VD - Vertical Disparity T - Top B - Bottom L - Left R - Right

Table 2: Disparity Estimation Errors

Image	RMSE E_u	 E_e	 E_{rc}	 E_{lc}
Crowd	47.12	18.86	18.35	12.76
Lake	30.17	15.34	15.89	10.97
Grand Canyon	39.82	24.21	24.30	12.29
Juggler	47.71	20.84	21.25	11.00
Clubs	31.36	18.18	18.51	11.79
Mt St Helens	49.72	22.58	22.84	10.70
Molecule (synthetic)	50.44	7.54	7.70	3.68

$$\mathrm{RMSE} = \sqrt{\left(\sum_{i=1}^{M}\sum_{i=1}^{N}\left(I(i,j) - R(i,j)\right)^2\right) / (MN)}$$

I - Original image R - Reconstructed

E_u - Uncompensated error term (difference between left and right image)

E_e - Error after estimating right from left (difference between the estimated and original right images)

E_{rc} - Error after estimating the right from the compressed left image

E_{lc} - Error in compressing the left (difference between the compressed and original left image