OPTIMAL USAGE OF COLOR FOR DISPARITY ESTIMATION IN STEREO VISION

Eran D. Pinhasov, Nahum Shimkin and Yehoshua Y. Zeevi.

Department of Electrical Engineering, Technion – Israel Institute of Technology, Haifa, Israel
phone: +972-4-8545772, fax: +972-4-8551550, email: eran.pinhasov@zoran.com

ABSTRACT

Most of the stereo algorithms available today use only Luminance images, assuming that the chromatic channels are redundant. This is based on the assumption that the Luminance channel holds most of the information. We propose to improve the disparity estimation by working in a transformed color space. The optimal color space is found by minimizing the disparity estimation variance, which is calculated from the stereo-pair input images. The transformed images can be used as input for stereo vision algorithms. We examine both local and global versions of the proposed process and a local adaptive approach as well, which selects the number of channels to be used during the correspondence process. The improved performance of these methods is demonstrated using synthetic and real stereo pairs.

1. INTRODUCTION

Most of the stereo vision algorithms use intensity images only. Nevertheless some researchers suggested that the chromatic channels could improve disparity estimation [2,3,1]. The most common method to incorporate the chromatic channels is by using a straightforward extension of the intensity dissimilarity or difference metric [1,3,4], using a weighted-sum of the channels dissimilarity metrics.

A very common dissimilarity function for a single channel is the sum of square difference (SSD), calculated over a region around the point of interest for robustness. We note that the SSD metric is considered optimal when using Bayesian estimation methods [4,2], assuming no correlation between the chromatic channels noise.

Okutomi *et. al* [3] adopted an analytic approach for evaluating the benefits of using color. They formulated an analytic expression that relates the input observations to the disparity estimation variance and showed that the disparity estimation variance obtained by taking all three channels is lower than or equal to the one using only a single channel, or even an intensity channel. Both Jordan *et. al* [1], Okutomi *et. al* [3] and others assumed that the noise component in the color image is not correlated between the chromatic channels, i.e. the noise covariance matrix is diagonal.

Magarey [2] in his work on motion estimation reported that the noise in color images is correlated; hence, the usage of a straightforward SSD for all color channels is not optimal. Magarey formulated a color space transformation matrix, based on the inter-frame noise covariance matrix, transforming the RGB color space to a new "optimal" color space in which there is no correlation between the channels noise. Magarey claimed that only by using the "optimal" color space, the SSD metric could serve as an optimal metric.

The main disadvantage of full color methods is their complexity, without necessarily outperforming simple monochrome algorithms. On the other hand, common intensity images, like Y (out of the YCbCr color space), hold only part of the observations, which may reduce features and contrast leading to poor results compared to full color algorithms. Based on these insights we consider an adaptive approach that incorporates the color information into the correspondence process.

Our first goal is to find a linear combination of the color channels, serving as an optimal intensity image in the sense of minimum disparity estimation variance. A local solution for this problem is described in section 2. Section 3 extends the local linear combination to a local color space which simultaneously de-correlates both noise and relevant signal. We present an adaptive algorithm for locally selecting the number of channels to use in the new color space, in order to reduce complexity while maintaining good disparity estimations. Next, we discuss several options for finding a global solution, to the optimal channel linear combination. In section 5 we show experimental results for both synthetic and natural inputs demonstrating the superiority of the proposed methods.

2. THE LOCAL BEST COLOR VECTOR

We start with the idea presented by Okutomi *et. al* [3], assuming a pair of color stereo images with additive noise using an SSD error metric for finding correspondence. Other assumptions include: no occlusions, no color or image distortions by the cameras, static scene and additive Gaussian noise with zero mean which may exhibit correlation between the chromatic channels. The left and right images may have different noise statistics but they are assumed to be uncorrelated and the displacement between the left and right images is constant for small regions. We use 1-D images for simplifying the calculations, but it is the same for 2-D images (based on the epipolarity constraint).

Let us start with the original noise free left and right color images (i.e. before the digital acquisition process), $\mathbf{o}_L$ and $\mathbf{o}_R$ respectively. Every image location is represented by a 3×1 vector, holding the red, green and blue color components. Assuming disparity between the images we write

$$\mathbf{o}_R(x) = \mathbf{o}_L(x - d(x)), \tag{1}$$

where x is the current location and $d(x)$ is the real unknown disparity for location x. The final – digitally acquired pair is noisy i.e.,

$$\mathbf{f}_L(x) = \mathbf{o}_L(x) + \mathbf{n}_L(x), \quad \mathbf{f}_R(x) = \mathbf{o}_R(x) + \mathbf{n}_R(x), \tag{2}$$

where $\mathbf{n}_L(x)$ and $\mathbf{n}_R(x)$ are Gaussian random vectors (3x1), with a 3×3 covariance matrix each, $\mathbf{R}_{NL}$ and $\mathbf{R}_{NR}$ for the left and right images respectively, assumed to be constant for a small region W around location x. Assuming constant disparity for region W, we replace $d(x)$ with a regional disparity $d_{[W]}$ and create a pair of intensity images

$$\begin{aligned} I_L(x) &= \mathbf{c}^T \mathbf{f}_L(x) = \mathbf{c}^T (\mathbf{o}_L(x) + \mathbf{n}_L(x)), \\ I_R(x) &= \mathbf{c}^T \mathbf{f}_R(x) = \mathbf{c}^T (\mathbf{o}_L(x - d_{[W]}) + \mathbf{n}_R(x)), \end{aligned} \tag{3}$$

where $\mathbf{c}$ is a 3x1 local (for region W) Color Vector (CV) holding the RGB weights $\mathbf{c} = [C_R \;\; C_G \;\; C_B]^T$. Using the SSD as an error metric we obtain

$$e_I(x,d) = \sum_{j \in W} [I_L(x+j) - I_R(x+j+d)]^2. \tag{4}$$

The value of d with the lowest error is the estimate of the local disparity $d_{[w]}$. Substituting the intensity images of equation (3) into equation (4), we obtain

$$e_I(x,d)=\sum_{j\in W}\left\{\mathbf{c}^T\left[\mathbf{o}_L(x+j)-\mathbf{o}_L\left(x+j+d-d_{[W]}\right)+\mathbf{n}(x+j)\right]\right\}^2, \quad (5)$$

where $\mathbf{n}(x) = \mathbf{n}_L(x) - \mathbf{n}_R(x+d)$ is the difference noise, which is also an additive Gaussian noise whose covariance matrix is

$$\mathbf{R}_N = \mathbf{R}_{NL} + \mathbf{R}_{NR}. \quad (6)$$

Note that $\mathbf{R}_N$ is also constant for region W. By taking the Taylor expansion of $\mathbf{o}_L(x+j+d-d_{[W]})$ in (5) around $(x+j)$, assuming $(d-d_{[W]})$ to be small we obtain a quadratic form of $(d-d_{[W]})$

$$e_I(x,d)\approx a(x)\left(d-d_{[W]}\right)^2-2b(x)\left(d-d_{[W]}\right)+g(x), \quad (7)$$

where

$$a(x)=\sum_{j\in W}\left(\mathbf{c}^T\mathbf{o}'_L(x+j)\right)^2,\ b(x)=\sum_{j\in W}\left[\mathbf{c}^T\mathbf{o}'_L(x+j)\right]\left[\mathbf{c}^T\mathbf{n}(x+j)\right] \quad (8)$$

and g(x) is some function of the noise component. The value of d that minimizes equation (7) is the estimated local disparity for the color vector $\mathbf{c}$

$$\hat{d}_{[W]}=d_{[W]}+\frac{b(x)}{a(x)}. \quad (9)$$

Further assuming that $a(x)$ and $d_{[W]}$ are deterministic we get,

$$Var\left(\hat{d}_{[W]}\right)=\frac{1}{a^2(x)}Var(b(x)), \quad (10)$$

We thus obtain the following expression:

$$Var\left(\hat{d}_{[W]}\right)=\frac{Var\left(\mathbf{c}^T\mathbf{n}\right)}{\sum_{j\in W}\left[\mathbf{c}^T\mathbf{o}'_L(x+j)\right]^2}. \quad (11)$$

Both the numerator and denominator in (11) can be written as matrix quadratic forms as follows:

$$v(\mathbf{c})=Var\left(\hat{d}_{[W]}\right)=\frac{\mathbf{c}^T\mathbf{R}_N\mathbf{c}}{\mathbf{c}^T\mathbf{R}_D\mathbf{c}}, \quad (12)$$

where $v(\mathbf{c})$ is the calculated local disparity estimation variance, as function of the color vector $\mathbf{c}$, $\mathbf{R}_N$ is the local difference noise covariance matrix (calculated in (6)) and $\mathbf{R}_D$ is the local (for region W) noise-free left image Horizontal Derivatives Cross-correlation Matrix (HDCM) which is obtained as follows:

$$\mathbf{R}_D=\sum_{j\in W}\mathbf{o}'_L(x+j)\left[\mathbf{o}'_L(x+j)\right]^T. \quad (13)$$

Since we do not have the original, noise-free left image, an estimation of $\mathbf{R}_D$ should be calculated based on the final (noisy) left image. One possible way for calculating $\mathbf{R}_D$ is by noise suppression prior to the use of equation (13) (see [4] for details). The value of $\mathbf{c}$ that minimizes (12) is a local CV that should yield the lowest disparity estimation variance; referred to as the Local Best Color Vector (LBCV) given by

$$\mathbf{c}_{LBCV}=\arg\min_{\mathbf{c}}\left(\frac{\mathbf{c}^T\mathbf{R}_N\mathbf{c}}{\mathbf{c}^T\mathbf{R}_D\mathbf{c}}\right). \quad (14)$$

One possible solution of (14) is obtained as a solution of the following eigenvalue problem, assuming that $\mathbf{R}_D$ is invertible

$$\mathbf{R}_D^{-1}\mathbf{R}_N\mathbf{x}=\lambda\mathbf{x}, \quad (15)$$

where the eigenvector with the lowest eigenvalue serves as $\mathbf{c}_{LBCV}$. Let $\mathbf{V}_{LBS}$ hold the eigenvectors of $\mathbf{R}_D^{-1}\mathbf{R}_N$ in its columns in ascending order by their eigenvalues, from left to right. $\mathbf{V}_{LBS}$ represents a special color space transformation matrix which de-correlates both $\mathbf{R}_N$ and $\mathbf{R}_D$, as follows

$$\mathbf{V}_{LBS}^T\mathbf{R}_D^{-1}\mathbf{R}_N\mathbf{V}_{LBS}=\mathbf{D}_{LBS} \quad (16)$$

$$\mathbf{V}_{LBS}^T\mathbf{R}_N\mathbf{V}_{LBS}=\mathbf{D}_{Nlbs} \quad \text{and} \quad \mathbf{V}_{LBS}^T\mathbf{R}_D\mathbf{V}_{LBS}=\mathbf{I},$$

where both $\mathbf{D}_{LBS}$, $\mathbf{D}_{Nlbs}$ are diagonal matrices (see [4] for proof). Equation (12) is a theoretical calculation of the local disparity estimation variance for region W, using $\mathbf{c}$ as channel weights for creating the intensity image. Intuitively, equation (12) is like a Noise-to-Signal ratio. We summarize with the following algorithm for full image stereo correspondence:

Algorithm 1: *Full image correspondence using LBCV*

1. Divide the left image to small regions, small enough for having the same disparity. Repeat steps 2-4 for all regions.
2. Find $\mathbf{c}_{LBCV}$ for the current region, using (14).
3. Create an intensity left region, and a search area from the right image, using $\mathbf{c}_{LBCV}$ as the color vector in equation (3).
4. Use a gray-level stereo algorithm for finding the regional disparity, using region I_L and the I_R search area.

3. THE LOCAL BEST SPACE

We go back to the eigenvector matrix $\mathbf{V}_{LBS}$ from (16), this matrix is a color space transform matrix. The new color space, referred to as the Local Best Space (LBS), is obtained as follows

$$\begin{bmatrix}c_1\\c_2\\c_3\end{bmatrix}=\mathbf{V}_{LBS}^T\begin{bmatrix}r\\g\\b\end{bmatrix}, \quad (17)$$

where r,g and b are the input RGB pixel components, c_1, c_2 and c_3 are the transformed components. $\mathbf{D}_{LBS}$ from (16) is a diagonal matrix whose diagonal components are the calculated disparity estimation variance for the three channels in the LBS color space. For instance, the top-left eigenvalue in $\mathbf{D}_{LBS}$ represents the theoretical disparity estimation variance when using $\mathbf{c}_{\text{LBCV}}$. The other two diagonal components have higher values representing channels that give higher estimation variance. Because the diagonal components of $\mathbf{D}_{LBS}$ are ordered from low to high, the transformed channels are ordered from the one that is the best channel for disparity estimation (lowest variance), to the one which is the worst, with the highest variance.

Recalling the "optimal" color space suggested by Magarey [2], we observe that it de-correlates the frame difference noise, making $\mathbf{R}_N$ diagonal, without taking into account the signal itself. The new proposed space (LBS) takes into account both signal and noise during the de-correlating process what makes it suitable for adaptive usage of channels during the correspondence process. It should be noted that because the noise is not correlated in the LBS, the use of the SSD metric on all channels is considered optimal from a Bayesian point of view.

Obviously the "mono" algorithms are less expensive and faster than full color algorithms. Additionally, in many cases full color does not necessarily improve the estimation significantly compared to the increase in complexity. For this reason we consider a local adaptive approach that selects the number of channels to use during correspondence. We use the theoretical disparity estimation variance from (12) as input to the following algorithm

Algorithm 2: *Selecting the number of channels - locally*

if $(v(\mathbf{c})<2\text{ch\_thld})$ { nch = 1 } // Use C_1 only.
else if $(v(\mathbf{c})<3\text{ch\_thld})$ { nch = 2 } // Use C_1 and C_2
else { nch = 3 } // Use C_1, C_2 and C_3

The following error metric is used for correspondence:

$$e(x,d)=\sum_{i\in W}\sum_{q=1}^{\text{nch}}\frac{1}{\lambda_q}\left[\text{Cq}_L(x+i)-\text{Cq}_R(x+i+d)\right]^2,$$

where λ_1, λ_2 and λ_3, are the diagonal components of $\mathbf{D}_{LBS}$ from (16). For the tests we use 2ch_thld=0.135 and 3ch_thld=0.26 (see [4] for details). Later, in section 5 we refer to this method as Adaptive Local Best Space (ALBS).

4. A GLOBAL OPTIMAL COLOR VECTOR

A major strength of the LBCV is its locality, although this is also what complicates the algorithm. In this section we extend the LBCV to a global solution, which is inferior to the LBCV algorithm, with the benefit of simplicity, less storage and that it can be used as a separate pre-process stage to an existing stereo algorithm.

In [4], we presented several options for finding the global CV. One option minimizes the disparity estimation variance of all regions. Other options include a common CV, which is the most popular CV among all regions or a global CV based on a constant $\mathbf{R}_D$ matrix for the whole image. We concentrate on one of the options, which we find preferable. For solving the global problem, we minimize the sum of all regional variances (12), as follows

$$v_G(\mathbf{c}) = \sum_{W \in B} \frac{\mathbf{c}^T \mathbf{R}_N^{[W]} \mathbf{c}}{\mathbf{c}^T \mathbf{R}_D^{[W]} \mathbf{c}}, \tag{18}$$

where B is the set of all image regions. For simplicity we assume $\|\mathbf{c}\| = 1$, since every local quotient in (18) is invariant to the scale of $\mathbf{c}$. The CV that minimizes (18) is referred to as the Global Best Color Vector (GBCV) given by

$$\mathbf{c}_{GBCV} = \min_{\mathbf{c}} \arg \left\{ \sum_{W \in B} \frac{\mathbf{c}^T \mathbf{R}_N^{[W]} \mathbf{c}}{\mathbf{c}^T \mathbf{R}_D^{[W]} \mathbf{c}} \right\}. \tag{19}$$

The minimization problem in (19) need not be convex like (12), hence, an iterative optimization process is needed. Based on our experiments [4], some regions should be excluded from the sum. Regions with a very small $\mathbf{R}_D$ matrix (usually homogeneous regions) should be excluded since they dominants the sum with a very high disparity estimation variance even when using their LBCV. On the other hand there are regions with strong features (high $\mathbf{R}_D$), giving good estimations even for their worst CV, these also should be excluded. Based on experiments we performed, two thresholds are used for excluding the non-relevant regions.

$$G = \left\{ W \in B : \frac{\mathbf{x}^T \mathbf{R}_N^{[W]} \mathbf{x}}{\mathbf{x}^T \mathbf{R}_D^{[W]} \mathbf{x}} > \text{WorstThld} \;\; , \;\; \frac{\mathbf{y}^T \mathbf{R}_N^{[W]} \mathbf{y}}{\mathbf{y}^T \mathbf{R}_D^{[W]} \mathbf{y}} < \text{BestThld} \right\},$$

where $\mathbf{x}$ is the local "worst" CV (the right-most column of $\mathbf{V}_{LBS}$), $\mathbf{y}$ is the LBCV and G is the set of valid regions, replacing B in equation (19) above. We use WorstThld = 0.2 and BestThld = 0.135 based on our experiments in [4]. The solution to equation (19) is found using an iterative optimization algorithm; we use the Nelder-Mead Simplex algorithm with several start locations.

5. EXPERIMENTAL RESULTS

In this section we test the performance of our proposed methods, compared to other intensity and full color methods. We use both synthetic and real color stereo image pairs, added with synthetic noise. We used stereo images from the Middlebury stereo vision page - www.middlebury.edu/stereo, see [5] for details. In order to compare and evaluate the performance of the proposed methods we used the RMSE (root-mean-square-error) quality metric, measured in disparity units (pixels), as suggested in [5],

$$R = \sqrt{\frac{1}{|B|} \sum_{W \in B} \left(d_{[W]} - \hat{d}_{[W]} \right)^2}, \tag{20}$$

where $d_{[W]}$ and $\hat{d}_{[W]}$ are the real and estimated disparity for region W, B is the set of all regions and $|B|$ is the total number of regions. For finding correspondence between regions we used a simple, full search algorithm with a gradient based fine-tune stage, using a ±10 pel horizontal search size for the synthetic tests, and ±55 pels for the real stereo tests. For the detailed algorithm, see [4].

5.1 Tests with synthetic stereo pairs

We used a single input image and created a synthetic pair, mainly for having a controlled and known disparity for the input images.

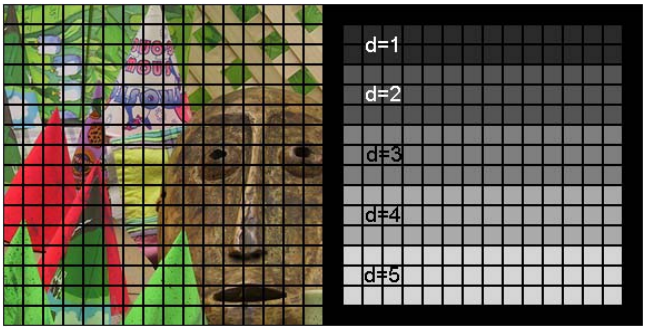

Fig. 1. Region partition (left) and synthetic disparity map (right) – for the "Cones" image.

The input image is divided to square regions, 16×16 pel each, giving every region a synthetic disparity value. Figure 1 show the original image divided to regions, and the used disparity map. The right image was created synthetically from the left image, using the disparity map in Figure 1. For checking the performance under noisy conditions, we add correlated noise to the left and right images, using different covariance matrices. Forty pairs of left and right noise covariance matrices ($\mathbf{R}_{NL}$ and $\mathbf{R}_{NR}$) are randomly selected referred to as varying-noise. For each pair we average the RMSE over 50 runs using the same $\mathbf{R}_{NL}$ and $\mathbf{R}_{NR}$ matrices. Figure 2 shows the sorted (by ALBS) RMSE results for varying-noise. The maximum channel noise variance is 0.008, for images in the range of [1,0]. We compare between: Y, GBCV, LBCV, ALBS, full RGB and Magarey's "optimal" space. Table 1 summarizes the RMSE improvement over the Y channel based the results in Figure 2.

Not only that the ALBS is superior to all other methods, it also reduces complexity during the stereo correspondence process. In the varying-noise experiment above ALBS need only ~43% of the calculations made by Magarey's "optimal" approach, on average.

	GBCV	LBCV	ALBS	RGB	Optimal
RMSE improve %	41.3	56.9	63.7	32.2	60.2

Table 1. Average RMSE improvement over the Y channel, in percent [%], for the "Cones" synthetic pair.

5.2 Tests with real stereo pairs

We use a single stereo pair called "Teddy" (from the Middlebury stereo vision page). "Teddy" includes two 450x375 color images and a ¼ pel resolution disparity map. The SNR of the pair is very high, for this reason we add synthetic noise in order to check the proposed methods under noisy observations. The disparity is calculated for a quarter of the left image pels, using a 15×15 region around every location for the search process, with a horizontal search size of ±55 pels. For having a close to real case, the regional noise covariance matrices and HDCM matrix are calculated from the noised images, (see [4] for detail). Figure 3 show the results for varying-noise for Y, LBCV, ALBS and Magarey's "optimal". The test is carried out using maximum channel variance of 0.004 (~24dB PSNR). Table 2 summarizes the RMSE improvement over the Y channel based on the results in figure 3. The results from the real scene "Teddy" are less distinctive than the synthetic tests. Based on additional tests reported in [4], we concluded that it is caused by in-accurate estimations for $\mathbf{R}_N$ and $\mathbf{R}_D$. This calls for better and robust methods for their calculation.

	LBCV	ALBS	Optimal
RMSE improve %	9.3	14.7	15.3

Table 2. RMSE improvement over the Y channel, for the "Teddy" stereo pair.

5.3 Real stereo pair visual demonstration

In order to visually demonstrate the performance of the LBCV, we use the same disparity estimation method as in section 5.2, this time on the "Cones" stereo pair. The "Cones" pair has very high SNR too; hence we add noise in order to check the LBCV performance under noisy conditions. We use the following covariance matrices for noising the left and right images (scaled by 1000):

$$\mathbf{R}_{NL} = s\begin{bmatrix} 5 & -1.63 & -1.21 \\ -1.63 & 4.04 & -0.29 \\ -1.21 & -0.29 & 0.99 \end{bmatrix}, \mathbf{R}_{NR} = s\begin{bmatrix} 4.16 & -1.49 & -0.69 \\ -1.49 & 5 & -1.7 \\ -0.69 & -1.7 & 4.11 \end{bmatrix}$$

were s is a scaling factor (s=1/1000). Figure 4 shows the estimated disparity maps using Y and LBCV. In figure 4, bright locations indicate scene locations which are closer, with a big disparity value. In this example, the LBCV improvement over the Y channel is clearly seen from the figures. A bad estimation is when disparity_error > 1pel.

6. SUMMARY AND CONCLUSIONS

In this paper, we have presented new pre-process methods for improving disparity estimation by using the color channels, for SSD based algorithms. Our main result is the Local Best Color Vector, which holds the local RGB weights, for creating an optimal image in the sense of low disparity estimation variance. We presented two extensions to the LBCV, one is an adaptive method for locally selecting the number of channels and the other finds a global optimal color vector. The performance of the proposed methods was compared to other known methods, using both synthetic and real color stereo images. The results indicate a significant improvement for average and high levels of noise.

We conclude that the ALBS algorithm should be considered instead of full color algorithms based on its ability to yield better estimations with reduced complexity. The LBCV and GBCV are less complex, suitable for existing single channel intensity algorithms. All of the proposed methods outperform the Y channel, hence, their usage should be considered based on the available computation power.

More work is needed for noise and derivative estimations from noisy observations and extensions of the LBS to other matching metrics like SAD, phase and 2D motion estimation.

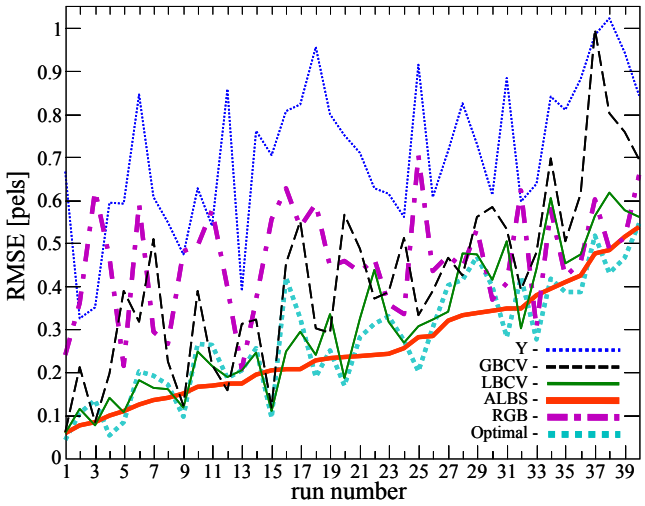

Fig. 2. Sorted averaged RMSE for varying-noise – "Cones" synthetic pair. Sorting by ALBS.

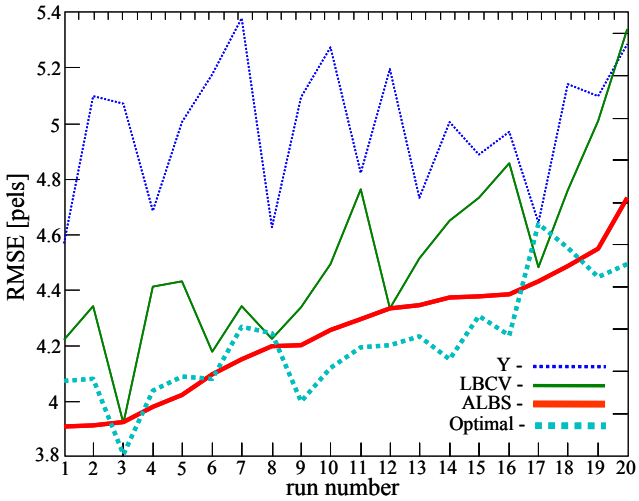

Fig. 3. Sorted RMSE for varying-noise – "Teddy" real noised pair. Sorting by ALBS.

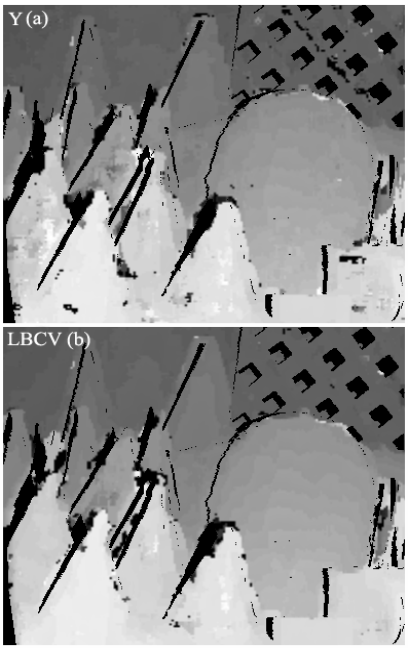

Fig. 4. (a) Y channel and (b) LBCV disparity maps. Y_RMS = 5.22 and LBCV_RMS = 4.89. Y has 23% of bad estimations, LBCV has 18% of bad estimations.

REFERENCES

[1] J. R. Jordan III and A. C. Bovik. "Using Chromatic information in dense stereo correspondence", *Pattern Recognition*, vol. 25, no. 4, pp. 367-383, 1992.

[2] J. Magarey. "Motion Estimation using Complex Wavelets" Ph.D. Thesis, Signal Proc. and communications Lab, University of Cambridge, 1997.

[3] M. Okutomi, O. Yoshizaki and G. Tomita. "Color Stereo Matching and its application to 3-D measurement of optic nerve head". *Proc. 11th Int. conf. On Pattern Recog.*, pp.509-513, 1992-I.

[4] E. D. Pinhasov. "Optimal Usage of Color for Stereo Vision", M.Sc. Thesis, Electrical Engineering faculty, IIT - Israel Institute of Technology, Jan. 2005

[5] D. Scharstein and R. Szeliski. "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms", *Int. Journal of Comp. Vis.*, vol. 47, No. 1/2/3, pp. 7-42, 2002.