EURASIP Journal on Image and Video Processing
a SpringerOpen Journal

RESEARCH Open Access

Efficient estimation of disparity statistics and their use as a predictor for perceived 3D video scene quality

Ömer C Gürol[1*], Bülent Sankur[1], Burak Acar[1] and Mehmet Güney[2]

Abstract

Excessive depth perception in 3D video is one of the major factors that causes discomfort to the viewer and that can decrease the viewer's quality perception of 3D video. With the idea of real-time quality control of 3D videos, we proposed an edge-based sparse disparity estimation algorithm with a novel similarity metric. The comparative assessment with other four state-of-the-art similarity metrics, implemented within the proposed edge-based disparity estimator, showed higher performance for the novel metric. User tests are conducted to assess the relation between certain disparity statistics and user perception of 3D scene quality that is a retrospective subjective experience of quality. Subjective tests indicate that the viewer discomfort can be predicted best by using maximum and slew rate of 95 percentile scene disparities together.

Keywords: Visual discomfort; 3D quality; Maximum disparity; Stereo video; Sparse disparity map

1 Introduction

The consumer market is moving rapidly toward 3D motion image delivery, and content providers, distributors, and equipment manufacturers see this as an opportunity. Consequently, there is an intense development effort in the field of 3D technologies, which range from compressive coding to 3D displays. In parallel, the 3D content production is growing rapidly in the form of 3D cinema and television programs. It is expected that 3D video will attain a large usage both at home theaters and mobile platforms in the coming years . The penetration of 3D video into our lives brings in concomitantly the question of multimedia user experience. In particular, the comfort level of 3D viewing will be of paramount importance in contrast to 2D video; in fact, negative aspects such as visual strain and viewer fatigue will curb the wide adoption of 3D. In this paper, we address two issues: (1) realizable and efficient estimation of large disparities in 3D video and (2) the impact of excessive disparities on 3D viewing comfort.

In 3D broadcasting besides having visually good images for each stereo channel, it is also important that the channels match with each other. Research in stereo quality control has identified a number of factors affecting viewing experience, such as parallax (disparity) irregularities, focus mismatch, color mismatch, geometry mismatch, vertical parallax, object edge tearing, cardboard effect, pincushion distortion, etc. Such distortion factors affecting 3D video quality are well documented [1].

The mechanism of depth perception in the human visual system is fairly well understood. It is known that depth perception uses both psychological and physiological cues. On the psychological aspect, the human visual system (HVS) uses cues related to perspective, such as overlap, shadow, apparent size, and texture; on the physiological aspect, the main cues include binocular parallax, motion parallax, accommodation, and convergence [1]. Among them, the binocular parallax, while being the most common method for 3D stereo rendering, is also one of the most dominant factors affecting the viewing experience. Binocular parallax is the relative spatial distance between similar points, which share the same physical origin, in the left and right stereo image pairs. It is governed by the binocular disparity, that is, the horizontal separation between the retinal images of the two eyes, when

*Correspondence: omer.gurol@boun.edu.tr
[1]Electrical and Electronics Engineering Department, Boğaziçi University, Bebek, Istanbul 34342, Turkey
Full list of author information is available at the end of the article

convergence to a specific distance is achieved. Hence, it can be conjectured that the quality of 3D video from the viewers' point of view is largely determined by the binocular disparity. This calls for an automated 3D video quality assessment tool based on the estimated disparities between stereo frame pairs as a function of time. Such a tool would not only provide an overall quality figure for a given video but could also be able to provide information regarding the frequency of parallax errors with respect to scenes as a guidance for video post-processing.

It is known that the brain uses binocular disp arity to infer depth information from the 2D retinal images resulting in 3D perception, that is, stereopsis. The creation of the sense of depth via binocular disparity in stereoptical screens is influenced by the size of the 3D display and the viewing distance, given the same relative parallax. Thus, the disparity requirements vary proportionally for cinema viewing (typically 20 m), home TV viewing (typically 1.5 m), and mobile device viewing (typically 0.2 m). In practice, smaller screens require a larger stereo baseline to provide more disparity as a fraction of the image width to retain a good impression of depth. It has been recommended that perceived depth range be upper bounded at a visual angle of 60 arcmin to ensure visual comfort for the majority of the viewers [2-5].

In order to create a satisfying sense of depth, the disparities between the image pairs should be made compatible with HVS 3D perception. Excessive disparity values correspond to an exaggerated depth range; they may strain the binocular fusion faculty of the subject and may cause the scene depth to be perceived inaccurately. Such defective image pairs can cause a weakened depth sense in the observer, and it can even result in headache or nausea when exposed for a long time. The effects of such defective image pairs on human visual system [3,6-8] and some quality assessment methods based on disparities [9] have been investigated in recent years.

The viewer discomfort is known to be affected by multiple factors and is observed in multiple ways. It is generally accepted that the vergence-accommodation conflict is a dominant factor in viewer discomfort and eye strain. Two major consequences of excessive disparities are vergence-accommodation conflict and double vision. Under fixed viewing conditions, the vergence-accommodation conflict can be related to the maximum disparity. The double vision is, however, affected by not only the stereo properties of 3D video, such as disparity, but also by its content and viewing environment. Lambooij et al. have noted that even with plausible disparity range, there are video characteristics such as fast motion or spatial and temporal inconsistencies that may contribute to the visual discomfort [8]. In this work, we focus solely on the effect of disparity, that can be measured efficiently with the proposed novel method, on discomfort.

In accordance with our goal of viewing comfort prediction using estimated excessive disparities, the main contributions of this paper are the following:

- We introduce a new edge-based efficient sparse disparity estimation approach.
- We introduce a novel similarity metric (correlation of gradient orientations (CGO)) for disparity estimation and carry out a comparative performance assessment with respect to the state-of-the-art metrics.
- We report the results of a pilot study exploring the correlation between the perceived 3D video quality and a number of statistical measures extracted from the sparse disparity estimates.

The disparity is estimated only on detected edge pixels; hence, the resulting disparity field is sparse in comparison to conventional dense disparity estimators in [10-13]. The sparse approach is chosen since the intention is not to reconstruct the entire disparity field but to find large range disparities in frames; as a byproduct, edge sparseness enables more rapid and efficient estimation of disparity statistics. We have introduced a new block similarity measure called CGO. This method is found to be a more efficient and reliable disparity estimator compared to several other block search methods. We also investigate the relationship between certain disparity statistics and user viewing comfort in order to develop a predictor of subjective 3D video quality. Such a quality indicator would help screen content provided by third parties prior to purchase decision; it would also be instrumental in quality-based scene selection in 3D video during post-processing and prior to broadcast.

The rest of the paper is organized as follows. In Section 2, we describe the proposed edge-based sparse disparity estimation algorithm and the similarity metrics used including the novel CGO. In Section 3, we explain the test material used together with the details of the similarity metrics we defined. The performance results on both reference databases, with ground-truth information, and on actual video streams from TV industry are presented and discussed in Section 4. The concluding remarks are given in Section 5.

2 Disparity estimation methods

Disparity estimation has been the subject of much interest in the last two decades, and a plethora of algorithms have been developed. These algorithms and their relative performances are well documented in the literature [14,15].

Several advanced algorithms have been proposed to reconstruct dense disparity fields through global optimization methods [12,13]. Such approaches, primarily due to their computational load, are not suitable for our

major goal of providing a 3D video quality metric fast enough to meet the needs of broadcast companies in selecting and/or post-processing the 3D video prior to purchase/broadcast. Furthermore, dense disparity fields are not required for our purpose. In order to differentiate our method from the dense disparity map methods in the literature, we will call ours the point disparity estimator and the outcome as sparse disparity map.

Our sparse disparity map algorithm, as illustrated in Figure 1, consists of image pre-processing, disparity search guided by the image edge field, and post-processing for error correction. These stages will be explained in detail in the following.

2.1 Image pre-processing for disparity estimation

We have confined our sparse disparity map estimation on image edges because (1) the block matching-based disparity estimation, irrespective of the similarity metric used, can best be done on non-flat image patches and (2) HVS is most sensitive to edges; hence, a computationally low-cost approach should be prioritized on edges. The well-known Canny edge detector is used for this purpose. Assuming that the cameras are rectified, we limit our search for stereo correspondences only in the horizontal direction. Furthermore, we eliminate all horizontal edges as the horizontal search along them yields ambiguous results. In short, we limit the search to edges with orientations within the 60° to 120° cone. The edge field is extracted in only one of the images in the stereo pair, and it is used to guide the placement of the disparity search window in the other image. The selected edges are dilated horizontally through morphological operations to make them s pixels wide (see Appendix). Illustration of these edge processing stages is given for the left image of a stereo pair in Figure 2.

Both left and right images are preprocessed to mitigate the illumination artifacts and to enhance the edge structures [15]. The rank filter is applied prior to disparity search whenever a pixel intensity difference-based similarity metric is used [16]. The rank filtering is omitted for other metrics as it decreases the dynamic range as explained in Section 2.2. The rank filtering simply considers a $W \times W$ window around each pixel, rank orders these pixel values, and assigns the rank of each pixel as its new pixel value $\in [1, W^2]$. This filtering is applied to both of the stereo images and the choice of $W = 15$ was found to be adequate, so that original gray values are mapped to the range $[1, 225]$.

2.2 Disparity search

The two images composing the stereo image pair are termed as the reference image (f_R) and the target image (f_T). Either left image or right image can be labeled as reference or target. The disparity search is done by means of sliding a search block, taken from f_R, over the search range defined on f_T. A search block, of size $N = h \times b$ pixels, is defined to be the $h \times b$ image patch centered at a predetected edge point (x, y) as $f_R(x - h/2 : x + h/2, y - b/2 : y + b/2)$. The matching block is searched for in f_T by sliding the aforementioned search block, with reference to its center point, over a range of horizontally shifted positions in f_T, namely over $(x, y - k : y + k)$. For each shifted position, a similarity cost is computed, i.e.:

$$\begin{aligned} C(d) = \psi(&f_R(x - h/2 : x + h/2, y - b/2 : y + b/2), \\ &f_T(x - h/2 : x + h/2, y - b/2 + d : y + b/2 + d)); \\ &d \in [-k, k] \end{aligned} \tag{1}$$

ψ is the similarity metric (similarity cost function), $C(d)$ is called the *cost profile*, and the estimated disparity is $d^E = \underset{d}{\operatorname{argmin}}(C(d))$. The parameters h and b are chosen empirically as the smallest block sizes providing robust estimation of similarity costs used in the study. The choice of k upper bounds the disparity estimates. Too small k values would result in underestimation of disparities, while too large k values would cause the algorithm to match unrelated image patches based on some structure or intensity similarity. We have $k = 80$, based on the properties of HVS and the datasets used, as detailed in Section 3.

We considered five state-of-the-art similarity cost functions including the proposed CGO. The definitions of the five similarity cost functions and the details of the cost aggregations are given in the following.

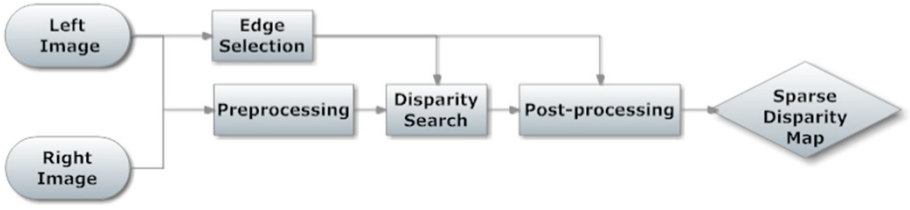

Figure 1 General flowchart of the algorithm. Both of the image pairs are subject to pre-processing. Edge selection is only applied to reference image (left image). The resulting edge map guides the disparity search and post-processing steps.

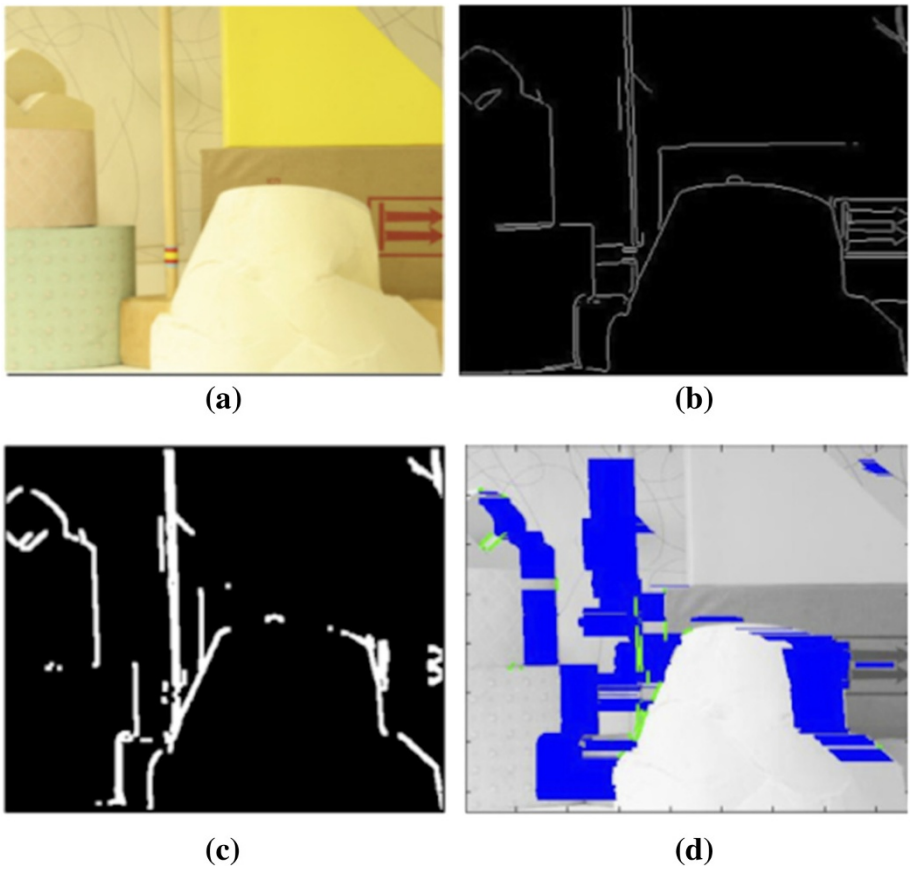

Figure 2 An example of pre-processing and estimated sparse disparity map. Left image of a stereo pair is given in **(a)** and its edge map in **(b)**. After removing the horizontally aligned edges and dilating the remaining ones the edge map becomes as in **(c)**. Estimated sparse disparity map is shown in **(d)** with blue lines. Green dots in **(d)** represent the *unreliable* (non-validated) disparities.

2.2.1 Sum of absolute differences

Sum of absolute differences (SAD) [10] is a well-studied matching cost calculation method for stereo matching and motion estimation tasks. SAD is defined as:

$$\mathrm{SAD}(d) = \frac{1}{N}\sum_{n=-h/2}^{h/2}\sum_{m=-b/2}^{b/2} |f_R(x+n, y+m) - f_H(x+n, y+m+d)|\,; \qquad -k \le d \le k \tag{2}$$

where $N = hxb$. The SAD value at candidate disparity d is obtained as the sum of the absolute difference of the two blocks from f_R and f_T at d units horizontal shift from each other.

2.2.2 Hermann Weyl's discrepancy measure

Hermann Weyl's discrepancy measure (HWDM) [17] is a similarity measure, which recently gained popularity for its usage in texture analysis tasks. HWDM uses the integral image concept. The pixel differences (not the absolute difference) of the two blocks are considered; these differences are integrated along four directions, namely, left to right and top to bottom (LRTB), left to right and bottom to top (LRBT), right to left and top to bottom (RLTB), and right to left and bottom to top (RLBT). For example, in the LRTB direction case, the integral image is obtained by summing the pixels from left to right and then from top to bottom; in other words, the image is first integrated horizontally and then vertically. Once the four integral images are obtained (each of size $h \times b$), the difference between maximum and minimum values in each integral image is calculated as $\max_{x,y}(\mathrm{I}_q) - \min_{x,y}(\mathrm{I}_q)$, and finally, the maximum among these four difference values from integral images is taken as the cost value. Accordingly, HWDM can be defined as:

$$\mathrm{HWDM}(d) = \max_{q}\left(\max_{x,y}(\mathrm{I}_q) - \min_{x,y}(\mathrm{I}_q)\right)\,; \qquad q \in \{\mathrm{LRTB}, \mathrm{LRBT}, \mathrm{RLTB}, \mathrm{RLBT}\} \tag{3}$$

Here, I_q represents the integral image obtained along the direction q. The minimum values of the integral images

are subtracted from the maximum ones in order to constrain the final costs to be positive. The coordinate location that yields the minimum HWDM is taken as the disparity estimate.

2.2.3 Adaptive support windows

In this method the search block is divided into five overlapping sub-blocks as shown in Figure 3. The central smaller sub-block is one third in size of the larger corner blocks. Accordingly, the final cost value at shift d, adaptive support windows ($ASW(d)$) [18], is calculated by adding the two smallest SAD costs of the four corner sub-blocks to the cost value of the center sub-block:

$$\begin{aligned}\mathrm{ASW}(d) = \mathrm{SAD}_5(d) + \min_{i=1}^{4} SAD_i(d) \\ + \text{second}\min_{i=1}^{4} \mathrm{SAD}_i(d)\end{aligned} \tag{4}$$

where $\mathrm{SAD}_i(d)$ represents the SAD cost of the ith sub-block at shift d.

2.2.4 Sum of absolute differences of scale invariant feature transform vectors

The images f_T and f_R are processed to extract their scale invariant feature transform (SIFT) fields [19]. The SIFT vectors are obtained by dividing the 16×16 neighborhood of each pixel into 4×4 cells and then quantizing the orientation in each cell into 8 bins [19]. Thus, each pixel in both of the images is replaced with a SIFT vector of size 128, and the corresponding SIFT vector images are obtained. The disparity search consists of matching blocks between the reference and target SIFT image blocks using simply the SAD criterion. Note that this algorithm does not need the pre-processing step of rank filtering (Section 2.1), since SIFT already uses the image contrast in the neighborhood. Again, the matching is performed only on edges of the reference image.

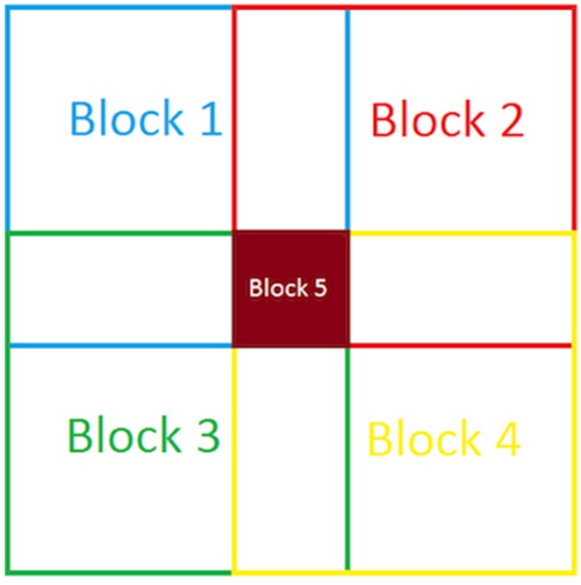

Figure 3 Overlapping sub-blocks in ASW. Blocks 1, 2, 3, and 4 are the support windows of block 5. Final cost value is obtained by adding the sum of the two minimum costs from the support windows to the cost of block 5.

2.2.5 Correlation of gradient orientations

In this method, first the complex gradient fields of the reference and target blocks are calculated [20]. The weakest pixels, that is, pixels with gradient magnitude below a given threshold are eliminated. The gradient orientation fields $\{O_R, O_T\}$ at the reference and target are calculated on the 'stronger' pixels. Finally, the SAD score of the two orientation fields is computed at shift d, that is:

$$\begin{aligned}\mathrm{CGO}(d) = \frac{1}{N} \sum_{n=-h/2}^{h/2} \sum_{m=-b/2}^{b/2} |O_R(x+n, y+m) \\ - O_T(x+n, y+m+d)|\,; \qquad k \le d \le k\end{aligned} \tag{5}$$

CGO algorithm can be summarized as:

1. Complex gradients of both of the images ($f_i(x,y)$ where $i \in \{R, T\}$) are calculated:
$$G_i(x,y) = \nabla_x f_i(x,y) + j\nabla_y f_i(x,y)$$
2. Pixels with sum of absolute gradient values less than some threshold Z are eliminated ($G_i(x,y) \rightarrow G_i'(x,y)$):
$$|\nabla_x f_i(x,y)| + |\nabla_y f_i(x,y)| \le Z \rightarrow 0$$
3. Gradient orientation maps ($O_i(x,y)$) are obtained:
$$O_i(x,y) = \frac{G_i'(x,y)}{\|G_i'(x,y)\|}$$
4. The correlation between gradient orientations at shift d, ($CGO(d)$) is calculated by taking SAD between O_R and O_T (Equation 4).

Notice that this algorithm also bypasses the rank filtering step due its use of gradients.

We determined the block sizes empirically. For fairness in the performance comparison of disparity estimation methods, we take the block size that yields the best results for each method. Thus, we used the following $h \times b$ figures: the size of the search block is 25×25 in SAD and ASW; in the latter, the four overlapping sub-blocks are of size 15×15 and the middle sub-block is of size 5×5. For HWDM, the block size is 11×11, for SADSIFT, 5×5, and for CGO, it is taken as 15×15.

As a post-processing step, each disparity estimate is validated by *cross-checking*. If a block at some position (x, y) in f_R finds its match in f_T at $(x, y + d^E)$, where d^E is the estimated disparity, then one searches for matching block

in f_R this time starting from the reference point in f_T. If the two estimations: $f_R \rightarrow f_T$ and $f_T \rightarrow f_R$ are consistent with each other, that is, they are within 2 pixels distance from each other, then the estimate is considered as valid. Otherwise, the pixel under investigation is labeled as *unreliable* in the disparity map. Such cross-checking is useful in determining the occluded regions and hence helps to reduce wrong disparity estimates.

3 Experimental setup

We describe here briefly the stereo image database, video test material, and the metrics used in experiments.

3.1 Stereo image database

We used 35 stereo image pairs, with known dense ground-truth disparity maps, in the Middlebury stereo image database for quantitative performance assessments [21,22]. These images are known to be rectified and have unidirectional disparity. The maximum disparity values occurring in Middlebury database are as follows: for 6 of the images, their absolute maximum disparities are less than 20 pixels ($\approx$ 5% of the scene), and in the remaining 27 images, the absolute maximum disparities vary between 38 to 71 ($\approx$ 10% to 16% of the scene). Accordingly, in all of our experiments with the Middlebury dataset, we set $k = 80$ ($\approx$ 18% of the scene), such that the disparity search ranges between $(x, y - 80)$ and $(x, y + 80)$ for any pixel at location (x, y). Thus, the search limit k is chosen larger than the maximum true disparity value of the image set, and hence, it allows for some overestimation. Any larger setting of the range k would have the potential to induce erroneous and somewhat unrealistic disparity estimates.

3.2 Video test material

In order to assess the potential of using the edge-based maximum disparity estimation for subjective 3D video quality prediction, we used a custom test video set consisting of 12 different stereo scenes from the footages provided by a commercial digital broadcasting company (Digitürk A.Ş.). Eight of these footages were taken in a soccer stadium by an expert 3D broadcasting crew; one of them is a computer animation, and three of them were taken in public locations around the city. The stereo shots have frame rate of 25 frames per second, and the scenes have durations that range from about 17 to 60 s. There is a 1-s length black screen between the scenes. The total length of the test video is 9,963 frames ($\approx$ 6.5 min). The scenes do not contain any subtitles.

The original resolution of the videos were $1,080 \times 1,920$ pixels, but we down-sampled them to 270×480 for practical purposes. The videos, shot by the professional crew, were not rectified; however, their vertical disparity was negligible. The stereo shots were taken with a slight angular shift between the cameras. Therefore, they have bi-directional disparities, such that when the left image is taken as the reference, then the resulting disparities for background objects can be expected to be *positive* and for foreground objects as *negative*. Some of these video shots contain large disparities as the offset between the cameras was intentionally and randomly modulated during the shootings.

Although these video scenes do not have ground-truth information according to which k could be set (as in Section 3.1), we have set $k = 80$ ($\approx$ 17% of the scene) manually, based on the observed maximum absolute disparity between the stereo pairs throughout the video. A disparity of 17% of the scene is equivalent to 64 arcmin angular disparity in our subjective test setup. This search range slightly exceeds the 60 arcmin of visual angle, which is the threshold for HVS to be able to fuse stereo images for 3D perception [2-5]. Therefore, $k = 80$ would be a suitable choice in our setup for the prediction of excessive disparity-related visual discomforts within the limits that still enable fusion.

3.3 Performance metric

Since the goal of our algorithm is to detect the largest disparities in the scene, we have developed performance measures to this effect. Our experience has shown that the absolute disparity estimation error is proportional to the size of the actual disparity. Furthermore, it can be conjectured that the largest disparities per frame and per scene affect the viewer comfort level the most. We therefore compute the mean of the largest 5% of the actual disparities, the true 95 percentile mean, $\mu 95$, for each image and use it as an indicator of estimation error. In fact, we rank the Middlebury images according to their ground-truth $\mu 95$ values in ascending order and reported the disparity estimation performance as a function of ascending $\mu 95$. The metrics we used are as follows:

Criterion 1: percentage of erroneous disparities (Erroneous%)

In this error criterion, we consider a disparity estimate as erroneous if it exceeds 10% of its true value. If d^G is the ground-truth and d^E is the estimated disparity, then the test $T(d^G, d^E)$ is expressed as:

$$T(d^G, d^E) = \begin{cases} 1 & if |d^G - d^E| > \lfloor \frac{d^G}{10} \rfloor \\ 0 & if |d^G - d^E| \leq \lfloor \frac{d^G}{10} \rfloor \end{cases} \quad (6)$$

The mapping between the disparity value and tolerable absolute disparity errors is given in Figure 4. Notice that the staircase behavior due to round-down is effectively a quantization.

This criterion tolerates errors in proportion to the actual disparity size; for example, the tolerance for disparity values $d^G < 15$ is 1, for $15 < d^G < 25$ is 2, while larger disparities allow for larger errors. A disparity estimation

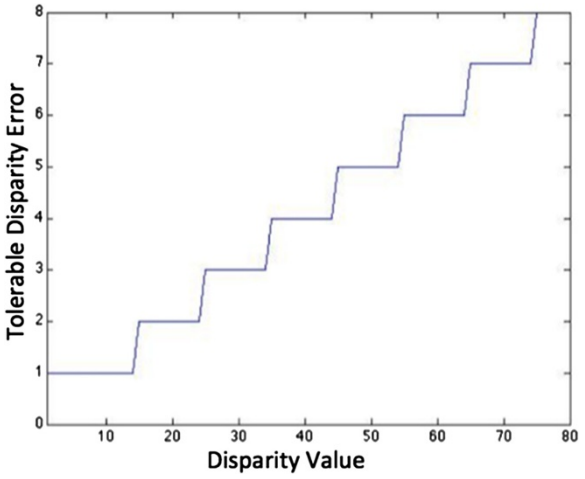

Figure 4 Threshold curve for calculating an erroneous disparity (Equation 7). Absolute disparity versus tolerable disparity estimation error. Disparity errors above the staircase function are penalized.

figure of merit can be obtained for the whole image in terms of the percentage of erroneous disparity estimates vis-*à*-vis the total number of pixels, C, at which disparity is estimated.

$$\text{Erroneous\%} = \frac{\sum_{i=1}^{C} T(d_i^G, d_i^E)}{C} \times 100 \tag{7}$$

This metric gives scores in the $[0, 100]$ range, with 0 corresponding to perfect estimation

Criterion 2: 95 percentile absolute error (Diff95%)
To put the disparity performance at large values into better evidence, for each image, we calculate the 95 percentile of disparities. In other words, we rank the disparities in ascending order and take the pixels corresponding to the highest 5% of disparities and then calculate the disparity errors at this particular cut point. More specifically, consider the rank ordered ground-truth disparities: $\{d_1^G, d_2^G, \ldots, d_C^G\}$, and consider the threshold value ν for the largest 95% disparities: $\nu = \lfloor 0.95 \times C \rfloor$. Then one has the following:

$$\text{Diff95\%} = |d_\nu^G - d_\nu^E| \tag{8}$$

where the notations d_ν^G and d_ν^E signify the νth rank-ordered true and estimated disparities (95 percentile disparities). Criterion 2 yields the absolute discrepancy between the 95 percentile values of the ground-truth and estimated disparities. Obviously, the range of this metric is between 0 for perfect estimation and k, the largest attainable error.

Criterion 3: 95 percentile ratios (Ratio5%)
In this measure, the disparities are again sorted as in 95% absolute error, and we consider the ground-truth and estimated disparities in the last 5 percentile (between 95 and 100 percentiles) sets, respectively, are calculated. We then consider the ratio of the means of the disparities d^E and d^G. If this ratio is close to 1, then there is a good agreement; ratio scores above 1 means that the algorithm overestimates the disparities, and the ones below underestimates the disparities. This criterion can be expressed as

$$\text{Ratio5\%} = \frac{\frac{1}{C-\nu}\sum_{i=\nu}^{C} |d_i^E|}{\frac{1}{C-\nu}\sum_{i=\nu}^{C} |d_i^G|} \tag{9}$$

4 Results and discussion

4.1 Performance results on stereo images

To assess the performance of the proposed similarity metric quantitatively and in comparison with state-of-the-art metrics, in this case of large disparities, the Middlebury dataset is used. The quantitative disparity estimation performance results are given in Figure 5 for three performance indicators and five similarity metrics (cost functions).

Figure 5a shows the scatter plot of the disparity estimation errors, quantified as the percentage of erroneous disparity estimates (Erroneous%), as a function of $\mu 95$. It is observed that the disparity estimation error is correlated with the amount of true disparity, as represented by $\mu 95$. Furthermore, this plot shows that CGO, on the average, performs better than the other metrics. Table 1 shows the rank sum performances of the five disparity estimation methods considered. For any one method, rank 1 corresponds to the number of images where it has performed best and rank 5 where it has performed the worst. CGO ranked the best performing metric in 22 image pairs out of 35.

Figure 5b,c confirms further the high performance of CGO in comparison to the other four methods based on the absolute disparity estimation error and the relative disparity scores, respectively. CGO resulted in Diff95% values larger than 0 in only 10 image pairs (out of 35) with an outlier in a single image pair. The nearest performance is realized by SAD method, which results in 11 image pairs with Diff95% value larger than 0. When we consider the Ratio5% measure, both the SAD and CGO scores are clustered around 1 (which represents perfect performance). CGO has an outlier in a single image pair, though in general, CGO results are more tightly clustered around 1 as compared to SAD.

It is interesting to observe that all methods overestimated the disparity as $\frac{|d_i^E|}{|d_i^G|} > 1$ in almost all cases. This gives us confidence that the methods we applied would not miss large disparities, albeit at the risk of false alarms

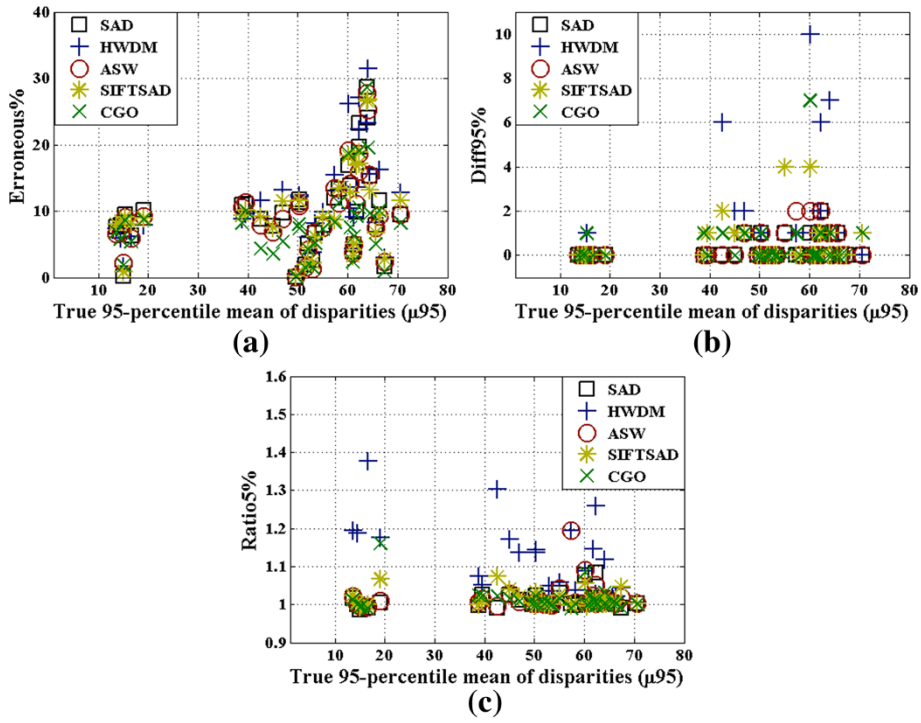

Figure 5 Comparison of results of SAD, HWDM, ASW, SADSIFT, and CGO for 35 Middlebury image pairs. Each mark in the graphs represents the result for an image pair. Image pairs take place along the horizontal axes in increasing order of true 95 percentile mean of disparities (μ95). In **(a)** overall percentage of erroneous disparities (Erroneous%), in **(b)** 95 percentile absolute error (95%), in **(c)** 95 percentile ratios (Ratio5%) are represented along the vertical axes.

by estimating the disparities larger than their actual values. It can be observed that HWDM is more prone to give false alarms since it overestimates the disparity more frequently. In this respect, CGO is a more reliable metric than the others with its Ratio5% values clustering around 1 without large outliers.

In terms of Diff95% and Ratio5% metrics, SAD and CGO may seem to have similar performances though it should be noted that the results in Figure 5 are obtained by selecting the best performing search block parameters for each similarity metric. In this sense, SAD required a larger search block size (25×25) to be taken in order to yield similar performance to CGO, where a block size of 15×15 sufficed. Obviously, larger block sizes are computationally more costly and the computational burden arising from larger block size in SAD is far greater than the simple gradient orientation map computations in CGO.

Table 1 Rank sums of the disparity estimation methods according to the percentage of erroneous disparity value (Erroneous%) of each image pair

	Rank 1	Rank 2	Rank 3	Rank 4	Rank 5
SAD	4	3	9	14	5
HWDM	3	3	4	3	22
ASW	6	11	9	6	3
SADSIFT	0	12	9	9	5
CGO	22	6	4	3	0

4.2 Subjective assessment of stereo scenes

We employed 15 subjects to assess the quality of the 3D stereo video data (as described in Section 3.2), and they reported their viewing experience in a follow-up questionnaire. This population size is suggested in the ITU-BT.500 recommendation as the lower limit of the cohort size. These subjects were chosen among students within an age range of 20 to 30. The subjects had normal visual acuity since none of them were normally wearing eye glasses. All subjects confirmed that they had previously watched stereoscopic 3D movies and that they did not experience any trouble in perceiving different levels of depth. For our calculations, we take the interpupillary distance of the subjects as 65 mm as it corresponds to the average human eye separation [3]. The display was a commercial 3D TV having $1,920 \times 1080$ pixel resolution and 89×50 cm screen dimensions. The stereoscopic display system was a two-view TV based on temporal multiplexing (shutter glasses) to create the stereoscopic depth sense. The subjects watched the videos sitting in a comfortable chair

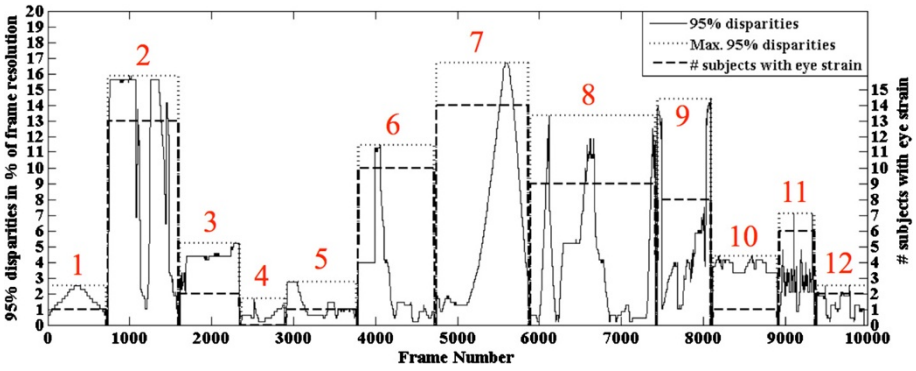

Figure 6 Results of CGO algorithm and subjective eye strain scores on video test data. Horizontal axis is the frame number, left vertical axis is the estimated 95 percentile disparity (d_v^E), and right vertical axis is the number of subjects. Dotted lines represent the max(d_v^E) scores, and the dashed lines represent the number of subjects with eye strain. The numbers on the graph represents the scene index.

at a distance of approximately 2 m in a room lit by subdued daylight. The subjects were not guided for looking at a particular point or object on the screen. The comfortable viewing zone is defined as a perceptual depth range, where the stereoscopic visual comfort is maintained [2]. Accordingly, the foreground and background distances of the comfortable viewing zone in our experiments are 0.57 and 1.33 m, respectively. After watching each scene, the subjects were asked if they felt any strain on their eyes at any part of the scene.

The d_v^E (estimated 95 percentile disparities, see Section 3.3) scores of the CGO algorithm on the whole set of video data (12 scenes resulting in 9,963 frames) are plotted in Figure 6 together with the number of subjects reporting eye strain for each scene. The ordinate on the left corresponds to the d_v^E score for each frame; the ordinate on the right corresponds to the subjective eye strain evaluation data, and the abscissa is the frame sequence number. The piecewise constant curve denotes the number of subjects reporting eye strain in the corresponding video scene (recall that we have 12 test scenes, hence 12 levels in the dashed and dotted curves). The results show a strong relation between the visual discomfort per scene with the maximum of estimated 95 percentile disparities (max(d_v^E)) per scene. Except for scenes 3 and 10, the number of subjects with eye strain is larger than 5 (33% percent of the population) whenever max(d_v^E) is above 3% of the frame resolution. The results for scenes 3 and 10 may be attributed to the small variation of d_v^E within these scenes.

Table 2 presents the results of the same experiment in alternative units, where the max(d_v^E) per scene appears in the top three rows in terms of the percentage of frame resolution, absolute minutes of arc and pixels, respectively. Maximum 95 percentile disparities are expressed in pixels and in percentage of frame resolution so as to take into account any downsampling effect. Furthermore, angular max(d_v^E) values are also given in absolute minutes of arc. The linear regression between the scene-wide max(d_v^E) values and the number of subjects with eye strain, depicted in Figure 7, confirms the above observation. We further performed single and multivariable regression between the number of subjects with eye strain per scene and estimated mean, standard deviation, and slew rate of 95 percentile disparities ($\mu(d_v^E)$, $\sigma(d_v^E)$, and $SR(d_v^E)$, respectively). The $\mu(d_v^E)$ statistic is related to the amount of excessive disparities similar to max(d_v^E), while the statistics $\sigma(d_v^E)$ and $SR(d_v^E)$ are related to the temporal change of excessive disparities within each scene. The temporal change of excessive disparities in the scenes relates to the property of HVS that causes sudden disparity changes to induce visual discomfort. The regression results are reported in mean absolute error (MAE) in Table 3. These MAE results confirm that scene-wide max(d_v^E) is the best performing single parameter for predicting viewer discomfort observed as eye strain.

Multivariable regression of max(d_v^E) and $SR(d_v^E)$ yields the lowest MAE score and hence the best prediction. In comparison to single variable regression with only

Table 2 Measures of estimated maximum 95 percentile disparity and the number of subjects reporting discomfort for the 12 scenes

Scene	1	2	3	4	5	6	7	8	9	10	11	12
max(d_v^E) (in % of the frame resolution)	2.5	15.8	5.2	1.7	2.7	11.5	16.7	13.3	14.4	4.4	7.1	2.5
max(d_v^E) (in absolute minutes of arc)	10	61	20	6	10	44	64	51	55	17	27	10
max(d_v^E) (in pixels)	12	76	25	8	13	55	80	64	69	21	34	12
Number of subjects who reported eye strain	1	13	2	0	1	10	14	9	8	1	6	2

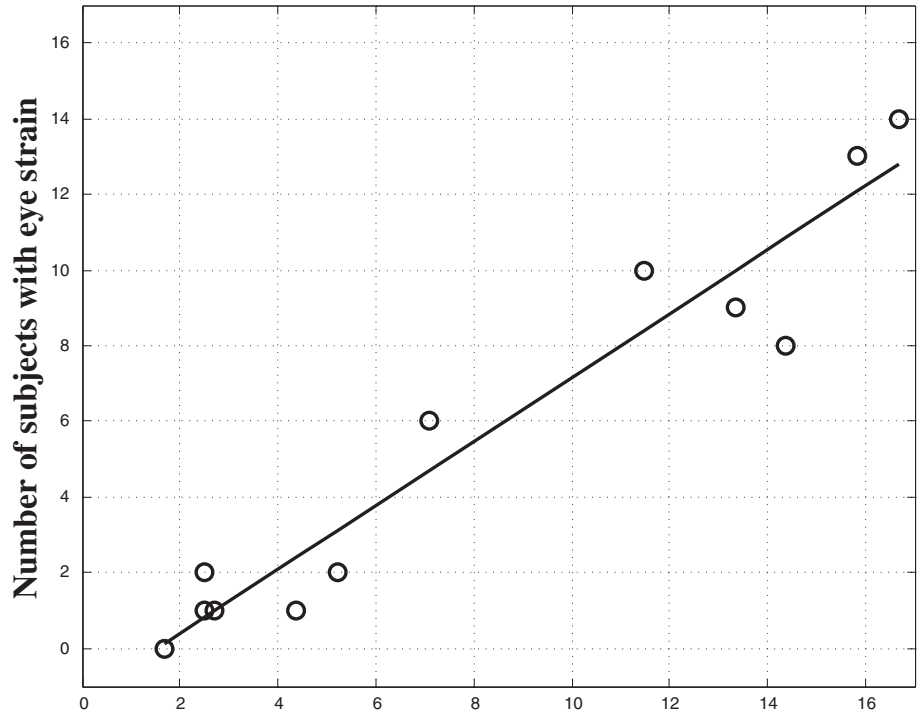

Figure 7 Maximum 95 percentile disparity ($\max(d_v^E)$) versus the number of subjects reporting eye strain. The estimated scene-wide maximum 95 percentile disparities are given in percentage of the frame resolution. The low mean absolute error (1.0725) shows that the proposed method is potentially a good estimator of eye strain, hence viewer discomfort.

Table 3 Mean absolute errors of linear regression between explanatory variables and the number of subjects reporting eye strain

Independent regression parameters	MAE
$\max(d_v^E)$	1.0725
$\mu(d_v^E)$	2.6737
$\sigma(d_v^E)$	1.1415
$SR(d_v^E)$	3.3606
$\max(d_v^E)$ and $\mu(d_v^E)$	1.0742
$\max(d_v^E)$ and $\sigma(d_v^E)$	0.9710
$\max(d_v^E)$ and $SR(d_v^E)$	0.8032
$\mu(d_v^E)$ and $\sigma(d_v^E)$	1.5752
$\mu(d_v^E)$ and $SR(d_v^E)$	2.5470
$\sigma(d_v^E)$ and $SR(d_v^E)$	1.6919
$\max(d_v^E)$, $\mu(d_v^E)$ and $\sigma(d_v^E)$	0.9330
$\max(d_v^E)$, $\mu(d_v^E)$ and $SR(d_v^E)$	0.8099
$\max(d_v^E)$, $\sigma(d_v^E)$ and $SR(d_v^E)$	0.8092
$\mu(d_v^E)$, $\sigma(d_v^E)$ and $SR(d_v^E)$	1.5163
$\max(d_v^E)$, $\mu(d_v^E)$, $\sigma(d_v^E)$ and $SR(d_v^E)$	0.8097

The variables are: maximum disparity ($\mathbf{max}(\boldsymbol{d_v^E})$), mean disparity ($\mu(d_v^E)$), standard deviation ($\sigma(d_v^E)$) and slew rate ($SR(d_v^E)$). MAE scores reflect the average prediction error over 12 scenes.

$\max(d_v^E)$, using the slew rate of 95 percentile disparities ($SR(d_v^E)$) together with $\max(d_v^E)$ improves the prediction performance by yielding 25% lower MAE score (0.8032). Although the $SR(d_v^E)$ yields the worst single parameter prediction results, it is interesting to observe that it can significantly improve the viewer discomfort prediction performance when combined with $\max(d_v^E)$. Slew rate of 95 percentile disparities gives a notion about the rate of the disparity changes within the scenes. Our experiments confirm that the sudden disparity changes, captured with $SR(d_v^E)$ statistics, are also an important factor in the assessment of stereo video quality together with maximum disparity statistics captured with $\max(d_v^E)$. In any case, it is encouraging to observe that eye strain can be predicted with an average error rate of 1 person among 15 subjects (6.7%).

We would like to discuss the limitations of our work. Recently, it has been pointed out that subjective assessment requires an evaluator cohort larger than the suggested size in ITU-BT.500. In our exploratory study, the material provided to us consisted mostly of outdoor scenes. A richer repertoire of video material including indoor scenes would be desirable. For example Lambooij et al. have shown that the visual comfort of video characteristics depends on the activity in the scene [8]. In addition to the post-session questionnaire, on-session

monitoring of the viewer's comfort, e.g., with a slider command would be valuable, for example, to monitor the onset of discomfort. Our questionnaire was binary; enabling multilevel qualified answers, allowing for example, bad, poor, fair, good, and excellent gradations would yield richer information but would possibly demand more experienced, if not expert, subjects. Finally, and perhaps most importantly, we have addressed only one aspect of visual discomfort.

5 Conclusion

In this study, we proposed a new maximum disparity estimation method and evaluated its performance in comparison with other four state-of-the-art methods, as a simple, fast, and objective 3D video quality assessment method. As the driving force for this study is to develop and validate a simple, objective, and fast method for 3D video quality assessment from the view point of viewers' comfort, to be used by broadcast companies, the video content and viewing environment dependent factors are not considered.

A combination of maximum and slew rate of 95 percentile disparity statistics per scene, estimated with the proposed CGO algorithm, was shown to predict viewer discomfort, seen as eye strain, with higher accuracy. The results of limited user tests suggest that the viewer discomfort is directly affected by even short duration of high disparities and sudden disparity changes in a scene, leading to low quality perception of the scene. This may be due to cognitive processes that drive the perception of video on the basis of scenes. This fact is especially important in decision making regarding the broadcast of 3D video content. A direct use of the proposed method would be a monitoring tool to preempt uncomfortable viewing experience especially for live 3D video shooting. In a future study, CGO algorithm can further be enhanced by using time-correlated information between consecutive frames to improve estimation reliability and smooth disparity time sequence.

Appendix

Disparities in a band

A disparity estimate at an edge pixel is chosen from among the s estimates in the band around the edge. Recall that edges were dilated horizontally (Section 2.1) by s pixels (typically, $s = 5$) via morphological operators resulting in s disparity estimates at and around each edge pixel. This many-to-one mapping helps to eliminate quite a number of *unreliable* (non-validated) disparities. As shown in Figure 8, there can be *positive*, *negative*, and unreliable disparities inside the s-wide band surrounding an edge point.

Along the depth of a scene, the disparity values are expected to range from the smallest negative values in the foreground to the largest positive values in the background (left image is taken as the reference). Accordingly, choosing the smallest signed disparities across the band means choosing the disparities that belong mostly to foreground objects. Since large disparities are typically associated with foreground objects, this processing step is consistent with our goals and is helpful in resolving some of the ambiguities. The s-wide band around each edge pixel is processed according to the following rule (Figure 8):

- If the number of unreliable estimates $< s/2$, then chose the minimum disparity.
- If the number of unreliable estimates $> s/2$ and center pixel is *reliable*, then chose the minimum disparity.
- If the number of unreliable estimates $> s/2$ and center pixel is unreliable, then leave as unreliable.

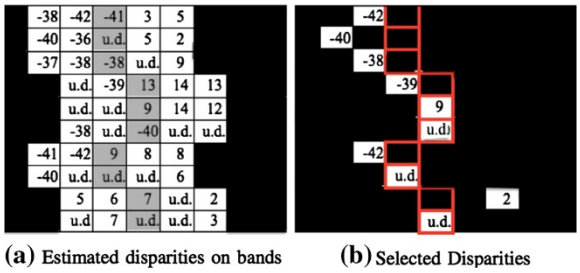

Figure 8 Example of disparities in a band post-processing step. The disparity values over the edge bands are given in **(a)** and the chosen disparities are given in **(b)**. The gray squares in **(a)** and red squares in **(b)** represent the original edge before dilation. Unreliable disparities are marked as u.d.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This research was supported by research grants from Digitürk. The authors would like to thank software development team and 3D shooting crew at Digitürk for their support in data collection and evaluation, Bernhard Moser for the fruitful discussions, and MATLAB codes for Herman Weyl's discrepancy calculations.

Author details
[1]Electrical and Electronics Engineering Department, Boğaziçi University, Bebek, Istanbul 34342, Turkey. [2]Software Development Department, Digitürk, Beşiktaş, Istanbul 34353, Turkey.

Received: 27 February 2013 Accepted: 9 September 2013
Published: 16 September 2013

References

1. A Boev, D Hollosi, A Gotchev, Classification of stereoscopic artefacts. Tech. Rep. 216503, MOBILE3DTV Project (2008). http://sp.cs.tut.fi/mobile3dtv/results/tech/D5.1_Mobile3DTV_v1.0.pdf. Accessed at 13 Sep 2013
2. W Chen, J Fournier, M Barkowsky, P Le Callet, et al., *New requirements of subjective video quality assessment methodologies for 3DTV*. (Video Process. Qual. Metrics (VPQM), Scottsdale, 2010)
3. M Lambooij, M Fortuin, I Heynderickx, W IJsselsteijn, Visual discomfort and visual fatigue of stereoscopic displays: A review. J. Imaging Sci. Technol. **53**(3), 1–14 (2009)

4. S Jolly, J Zubrzycki, O Grau, V Vinayagamoorthy, R Koch, B Bartczak, J Fournier, J Gicquel, R Tanger, B Barenbrug, M Murdoch, J Kluger, 3D Content requirements & initial acquisition work. Public Document 215075, 3D4YOU Project (2009). http://www.hitech-projects.com/euprojects/3d4you/www.3d4you.eu/images/PDFs/3D4YOU_WP1_D1.1.2_v1.0.pdf. Accessed at 13 Sep 2013
5. T Shibata, J Kim, DM Hoffman, MS Banks, The zone of comfort: Predicting visual discomfort with stereo displays. J. Vis. **11**(8) (2011)
6. K Ukai, PA Howarth, Visual fatigue caused by viewing stereoscopic motion images: background, theories, and observations. Displays **29**(2), 106–116 (2008)
7. M Lambooij, M Murdoch, W IJsselsteijn, I Heynderickx, The impact of video characteristics and subtitles on visual comfort of 3D TV. Displays **34**, 8–16 (2013)
8. M Lambooij, W IJsselsteijn, I Heynderickx, Visual discomfort of 3D TV: assessment methods and modeling. Displays **32**(4), 209–218 (2011)
9. A Benoit, P Le Callet, P Campisi, R Cousseau, in *15th IEEE International Conference on Image Processing*. Using disparity for quality assessment of stereoscopic images (San Diego, 12–15 Oct 2008), pp. 389 –392
10. T Kanade, M Okutomi, A stereo matching algorithm with an adaptive window: theory and experiment. Pattern Anal. Mach. Intell. IEEE Trans. **16**(9), 920–932 (1994)
11. KJ Yoon, IS Kweon, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Locally adaptive support-weight approach for visual correspondence search, vol. 2 (Colorado Springs, 20–25 Jun 2005), pp. 924–931
12. Y Boykov, O Veksler, R Zabih, Fast approximate energy minimization via graph cuts. Pattern Anal. Mach. Intell. IEEE Trans. **23**(11), 1222–1239 (2001)
13. P Felzenszwalb, D Huttenlocher, in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Efficient belief propagation for early vision, vol. 1 (Washington, D.C., 27 Jun–2 Jul 2004), pp. 261–268
14. D Scharstein, R Szeliski, A taxonomy evaluation of dense two-frame stereo correspondence algorithms. Int. J. Comput. Vis. **47**, 7–42 (2002)
15. H Hirschmüller, D Scharstein, in *IEEE Conference on Computer Vision and Pattern Recognition*. Evaluation of cost functions for stereo matching (Minneapolis, 17–22 Jun 2007), pp. 1–8
16. R Zabih, J Woodfill, in *Computer Vision — ECCV '94, Lecture Notes in Computer Science*, ed. by JO Eklundh. Non-parametric local transforms for computing visual correspondence, vol. 801 (Springer Berlin, 1994), pp. 151–158
17. B Moser, A similarity measure for image and volumetric data based on Hermann Weyl's discrepancy. Pattern Anal. Mach. Intell. IEEE Trans. **33**(11), 2321–2329 (2011)
18. H Hirschmüller, P Innocent, J Garibaldi, Real-time correlation-based stereo vision with reduced border errors. Int. J. Comput. Vis. **47**, 229–246 (2002)
19. C Liu, J Yuen, A Torralba, SIFT flow: dense correspondence across scenes and its applications. Pattern Anal. Mach. Intell. IEEE Trans. **33**(5), 978–994 (2011)
20. A Fitch, A Kadyrov, WJ Christmas, J Kittler, in *British Machine Vision Conference*. Orientation correlation, vol. 1 (Cardiff, 2–5 Sept 2002), pp. 133–142
21. D Scharstein, R Szeliski, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. High-accuracy stereo depth maps using structured light, vol. 1 (Wisconsin, 18–20 Jun 2003), pp. 195–202
22. Middlebury dataset. http://vision.middlebury.edu/stereo/data/. Accessed 13 Sept 2013

doi:10.1186/1687-5281-2013-53
Cite this article as: Gürol *et al.*: Efficient estimation of disparity statistics and their use as a predictor for perceived 3D video scene quality. *EURASIP Journal on Image and Video Processing* 2013 **2013**:53.