DOI: 10.2352/ISSN.2470-1173.2016.5.SDA-444
Reprinted from
Journal of Imaging Science and Technology® 59(6): 060401-1–060401-8, 2015.

Improving Visual Discomfort Prediction for Stereoscopic Images via Disparity-based Contrast

Werner Zellinger and Bernhard Moser
Software Competence Center Hagenberg, Knowledge-Based Vision Systems, Image and Information Fusion, Softwarepark 21, Hagenberg, A-4232, Austria
E-mail: werner.reisner@scch.at

Abstract. *Stereoscopic images and videos can lead to serious adverse effects on human visual perception. The phenomenon of visual discomfort depends on various influencing factors such as the arrangement of the display system, the image quality and the design of 3D effects. Real-time depth adaptations that reduce the extent of visual discomfort require computationally efficient prediction models. This article analyzes optimal combinations of image features of state-of-the-art models in terms of prediction accuracy and computational efficiency. In addition, a fast-to-compute disparity contrast feature based on Haralick contrast is introduced in this context. It turns out that the computational complexity can be reduced by restricting the number of features without loss of prediction accuracy. A Pareto-front analysis shows which features are more likely to be part of optimal combinations. It is interesting to observe that the introduced disparity contrast feature is part of combinations that are optimal in terms of both computational efficiency and accuracy. This means that state-of-the-art prediction models can be improved by means of the introduced disparity contrast feature. The analysis relies on statistical evaluations based on publicly available assessment data.* © *2015 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2015.59.6.060401]

INTRODUCTION

In the literature, the term *visual discomfort* (VD) refers to the subjective sensation of discomfort that is associated with watching stereoscopic images or video streams.[1] This article is about modeling of VD based on image analysis in order to predict this effect and to improve the overall acceptance of 3D technology. The extent of the VD experienced has been investigated by means of clinical and subjective assessments. Some viewers experience VD such as eye strain, headache and nausea.[2] Bad quality 3D content can cause permanent damage to the visual system of children.[3] This phenomenon depends on various influencing factors such as the arrangement of the display system, the image quality and the design of 3D effects such as depth grading. Visual discomfort is a key aspect for the overall quality of experience[4,5] of watching 3D content and therefore for the acceptance of 3D technology,[6] particularly in the context of 3D film production and 3D display systems.[7–10] Consequently, minimization of the viewer's discomfort is a major research activity of 3D technology production, which must take the full extent of viewer's experience into account.[11]

The literature offers various approaches for modeling the assessment of the extent of VD based on image data analysis. All of these approaches are based on the extraction of image features as input for a regression model. The various approaches differ in the chosen image features and how they are aggregated. Some authors prefer low level features such as depth range or depth distribution, mainly based on first-order statistics[11–14] that are derived from the disparity map. Recently, Sohn et al.[15] and Jin et al.[16] proposed the application of higher level image analysis techniques such as segmentation and object description. The latter approaches enable substantial improvement of the prediction accuracy for visual discomfort in comparison to the approaches that are driven by the extraction of first-order-based statistics only.

However, when VD predictions are part of computational models in the 3D content post-production process,[14,17–20] the aspect of computational efficiency becomes a major issue.

The contribution of this article is twofold: (a) we perform a Pareto optimality analysis of feature combinations in terms of balancing both prediction accuracy and time complexity of VD models, and (b) we introduce a standard feature from texture analysis, namely Haralick contrast,[21] in the context of VD prediction and adapt it to obtain a fast-to-compute disparity contrast feature (HC_D) that, in combination with other state-of-the-art features, outperforms state-of-the-art prediction models.

Our approach leads to the following claims with respect to the features used in the state-of-the-art approaches of Lambooij et al.,[11] Nojiri et al.,[12] Choi et al.,[13] Kim et al.[22] and Sohn et al.[15]

Claim 1 (Feature Selection): Taking all feature combinations into account, it suffices to concentrate on combinations with only four features without any significant loss of accuracy (see the Feature Selection section).

Claim 2 (Pareto Front): Bottom-up approaches allow one to construct substantially faster prediction models without significantly lower prediction accuracy compared with the top-down approaches under consideration (see the Pareto Front section).

Claim 3 (Haralick Contrast): The expected prediction accuracy that can be achieved by combinations of features

Received June 30, 2015; accepted for publication Sept. 14, 2015; published online Oct. 19, 2015. Associate Editor: Hang-Bong Kang.
1062-3701/2015/59(6)/060401/8/$25.00

DOI: 10.2352/ISSN.2470-1173.2016.5.SDA-444

including $HC_{\mathcal{D}}$ is significantly higher than for combinations without $HC_{\mathcal{D}}$. Moreover, the feature combination with the highest prediction accuracy contains the $HC_{\mathcal{D}}$ feature (see the Prediction Accuracy section).

As a main result these claims are underpinned with statistical significance based on state-of-the-art publicly available assessment data.[23]

The article is structured as follows. The second section recapitulates some basic facts about VD. The third section reflects related work about computational models for predicting visual discomfort. The fourth section introduces a fast-to-compute disparity contrast feature which is motivated by a co-occurrence matrix approach. The fifth section is devoted to experimental results and their statistical analysis.

VISUAL DISCOMFORT

The subjective sensation of discomfort someone experiences when watching stereoscopic images or video content is called VD.[1] This phenomenon can lead to serious adverse effects on human visual perception, as outlined in Ref. 3. The literature offers various factors that negatively affect visual discomfort.[2,12,24–32] The five most relevant types of influencing factors are[2] (1) accommodation–vergence conflict, (2) parallax distribution, (3) binocular mismatches, (4) perceptual inconsistencies and (5) cognitive inconsistencies.

Accommodation–Vergence Conflict

The stimulus of watching a 3D object triggers two physiological processes in the eye: an accommodation response and a vergence response. While the *accommodation process* refers to the adaptation of the lenses in order to focus the object on the retina, the *vergence process* refers to the relative angular constellation of the eyes' viewing direction so that both eyes are directed at the same object. However, when looking at stereoscopic displays the responses created from these processes can cause conflicts in visual perception.[2] One reason for this accommodation–vergence conflict can be excessive parallax.[12,24] To minimize the accommodation–vergence conflict, it is generally assumed that the disparities in a stereoscopic image should be limited by a "comfort zone".[2]

Parallax Distribution

The distribution of parallax that is induced by the spatial arrangement of the objects in the foreground and background in the scene is a characteristic of the scene and is related to features such as spatial frequency and disparity gradient.[2,31] Intuitively, the more scene details there are, the more there is competition for the visual attention by various potential objects of interest. Such ambiguity concerning visual attention might influence the comfort of visual perception. It has been shown that there is a high correlation between VD and parallax-distribution-based features.[15]

Binocular Mismatches and Depth Inconsistencies

Various types of binocular mismatch and their influence on VD were studied by Kooi and Toet[28] by applying distorting transformations on stereoscopic images. In contrast to blur and vertical offset effects, the results show little impact caused by transformations such as rotation, magnification and keystone distortions. It is a common technique to represent the depth information for a stereoscopic image in terms of the horizontal disparities of corresponding pixels between the left-eye and right-eye images. The resulting disparity maps are a possible source of errors in depth information. Such errors in the disparity map might be caused, e.g., by lossy compression or transmission. Corresponding depth inconsistencies also might affect VD.[2]

Perceptual and Cognitive Inconsistencies

Perceptual and cognitive inconsistencies might result from a mismatch between our cognition of the 3D appearance of real world objects and the insufficient appearance induced by the display system. For example, cognitive confusion might occur due to the border of the display system when an object in the scene (e.g., hand with five fingers) that is supposed to be in front of the screen is only partially visible (e.g., part of the hand showing three fingers).[32]

STATE OF THE ART FOR VISUAL DISCOMFORT PREDICTION

In general, computational models for predicting VD consist of two major parts: (a) choosing a subset of image features, and (b) modeling the VD as a mapping from the feature space onto a range of scales. For the mapping part, mostly linear[11] or piecewise linear[15] models are used to score the level of VD of a stereoscopic image. These supervised learning models are trained with data coming from subjective assessments. While we strive to keep the regression model as simple as possible, our major emphasis lies on the choice of appropriate features. Typical candidates for such features are statistical features that are derived from the stereoscopic input images and the rendered depth maps, or disparity maps, which result from applying stereo matching algorithms (see Table I for an overview). In particular, Choi et al.[13] rely on standard mean and variance of the disparity distribution for predicting VD. Lambooij et al.[11] suggest using the mean and range of the disparity map in order to characterize the disparity distribution; the range feature is determined by the difference of the maximum and the minimum of disparities with respect to 10% quantiles. Nojiri et al.[12] propose a weighted mean with varying weight coefficients according to different parts in the image and disparity map. Kim et al.[22] aim at modeling spatial frequency by introducing a weighted disparity map that is induced by image enhancement operations applied on the disparity map. They propose to model the effect of excessive screen disparity based on the sum of an upper percent quantile of disparities, see also Jung et al.[14] Sohn et al.[15] introduce a concept of disparity gradient as the relative disparity difference between the locations of objects. In addition, they aim at modeling the *stimulus width* of an object by proposing a disparity-based feature that takes the width of nearby objects into account; the objects result from applying a mean-shift segmentation on the disparity map.

DOI: 10.2352/ISSN.2470-1173.2016.5.SDA-444

Table I. Survey of disparity-based features.

Feature notation	Description	Literature
Mean, Var	Standard mean and variance of disparity map	Ref. 13
Range10	Difference of maximum and minimum with respect to 10% quantile	Ref. 11
Max5	Sum of 5% maximal disparity values	Ref. 22
Max5Sobel, Range10Sobel	Range10 and Max5 based on weighted disparity map induced by Sobel operator	Ref. 22
RD, OT	Mean relative disparity and width of nearby objects	Ref. 15

Note that Mean, Var, Range10 and Max5 are first-order driven image statistics, whereas RD and OT are higher level image features based on segmentation and grouping operations (i.e., based on pixel neighborhood structures).

While the outlined research mainly concentrates on prediction accuracy, in this article we also take computational efficiency into account, which becomes a critical issue from the point of view of integrating such models in workflows and processes for depth-image-based rendering of 3D video content; see, e.g., Refs. 14, 17–20. In particular, real-time-capable VD prediction models are required for real-time 2D-to-3D conversion; see, e.g., Refs. 33–36.

APPROACH BASED ON DISPARITY CONTRAST

Our approach to come up with a fast-to-compute VD prediction model with high prediction accuracy consists of analyzing the performance of single features and their combinations with respect to Pareto optimality, as outlined in the Pareto Front section. As a first step, we introduce a novel feature in this context by taking up the standard contrast feature due to Haralick,[37] which is commonly used in texture analysis. In the Fast-to-Compute Disparity Contrast Feature section, we outline how this co-occurrence-matrix-based approach due to Haralick can be adapted to obtain a fast-to-compute disparity contrast feature. This approach, which aims at modeling disparity contrast, is motivated from perception literature, as outlined in the Motivation for Disparity Contrast Feature from Psychophysics section.

Motivation for Disparity Contrast Feature from Psychophysics

In spite of intensive research effort, the nature of the neural mechanisms underlying human visual fixation behavior still remains vague.[38] One major reason is the difficulty in isolating pure physical, sensorial bottom-up mechanisms[39,40] from higher level goal-oriented top-down mechanisms that involve contextual knowledge.[41] However, a crucial bottom-up aspect refers to the *conspicuity area*, which, with single eye fixation, captures the spatial region around the center of gaze where the target can be resolved from its background.[42] The human visual target conspicuity is measured by a psychophysical procedure[43] and has been analyzed for a range of static targets in static scenes.[42,44] The investigations show that the conspicuity area is small if the target (object of interest) is surrounded by high spatial variability. This is no surprise, as in a complex scene many details compete for the observer's attention. On the other hand, targets that stand out clearly from the background induce a large conspicuity area. Therefore, distinctness of image details strongly influences human visual attention and fixation behavior. As a consequence, according to our working hypothesis, distinctness of image details is of high relevance for VD of 3D content. For rendered 3D scenes, distinctness of image details is influenced not only by natural characteristics such as the parallax distribution (scene complexity) but also by artifacts that result from excessive parallax configuration in the depth grading, errors in the disparity map, and border effects of the display system, as pointed out in second section.

It should be noted that disparity contrast features a model for the distinctness of image details by exploiting the histogram of depth gradients. Next, we look for a computationally efficient variant.

Fast-to-Compute Disparity Contrast Feature

The literature proposes various concepts for contrast based on image gray values.[37,38,45–48] These concepts rely on the co-occurrence matrix which involves quadratic computational complexity with respect to the number of gray values.

In the following we propose a disparity contrast feature that operates on the disparity map $\mathcal{D}$. We may assume that the disparity map is generated under normalized conditions, i.e., the disparities between corresponding points in the left and right images only appear horizontally. Then, the disparity map value $\mathcal{D}(x,y) = \Delta$ at pixel position (x,y) encodes the distance Δ by which the pixel (x,y) in the left image must be shifted horizontally in order to match the corresponding pixel in the right image. By summing up all squared differences $(\mathcal{D}(x,y) - \mathcal{D}(x+\delta,y))^2$ for some horizontal offset δ we come up with a measure that models distinctness of disparity map details. As shown in the Appendix, this measure can be obtained by applying the Haralick approach[37] to the disparity map $\mathcal{D}$, i.e.,

$$HC_{\mathcal{D}}(\delta) = \frac{1}{NM}\sum_{x,y}(\mathcal{D}(x,y) - \mathcal{D}(x+\delta,y))^2, \qquad (1)$$

where M, N denote the numbers of pixel rows and columns of the images. See Figure 1 for an illustration.

©2016 Society for Imaging Science and Technology
DOI: 10.2352/ISSN.2470-1173.2016.5.SDA-444

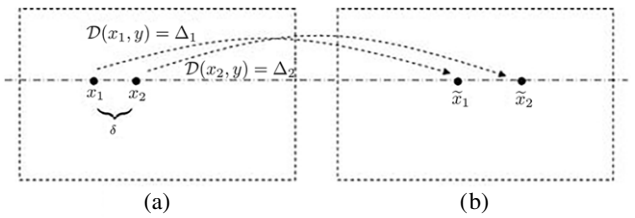


Figure 1. Illustration of the construction of the Haralick contrast feature Eq. (1) based on the disparity map $\mathcal{D}$, where x_i and $\widetilde{x}_i$ denote corresponding points in the left and right images, respectively.

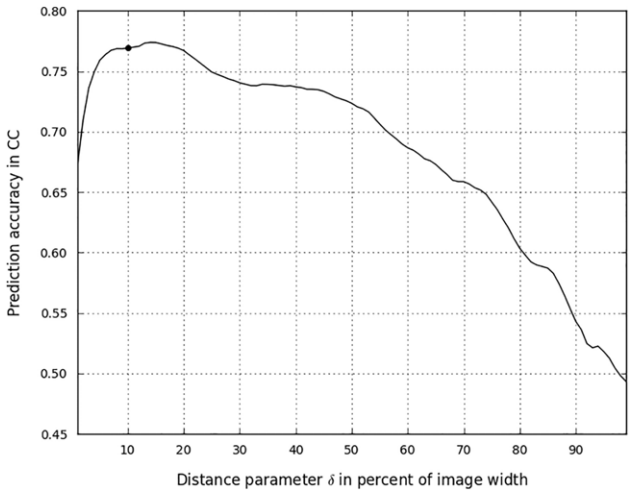


Figure 2. The graph shows the correlation between $HC_{\mathcal{D}}$ and VD scores for the database *KaistDB* as described in the fifth section. Remarkably, there is a distinct maximum around 10% of the image width.

A standard approach to also take other distances δ into account is to sum up the δ-induced disparity contrast values, i.e., $HC_{\mathcal{D}} = \sum_{d=-N}^{N} HC_{\mathcal{D}}(d)\alpha_d$, with weights α_d, where $\sum_d \alpha_d = 1$.

The question relates to what a reasonable choice for the weights α_d is. To this end, let us perform a sensitivity analysis based on linear regression analysis between $HC_{\mathcal{D}}$ and VD scores coming from ground truth data based on the database[23] (see the fifth section for details). A distance value δ in the range around 10% shows approximately maximal accuracy. Figure 2 shows the resulting correlation coefficients (CCs) for various distance values δ. For the definition of the CCs see the Experimental Setup section. This empirical study motivates us to restrict to only those distance values with high correlation, yielding a sparse disparity contrast feature. We choose for the parameter δ the value with the maximal correlation, that is approximately 10% of the image width N.

Typically, images with complex spatial arrangements of objects in the foreground and background yield higher $HC_{\mathcal{D}}$ values than images with less image details, since the gradients show higher values. For an illustrating example see Figure 3.

STATISTICAL EVALUATION

Our evaluation analysis aims to provide evidence regarding the following aspects: the Feature Selection section on feature selection, the Pareto Front section on time complexity versus prediction accuracy, and the Prediction Accuracy section on the improvement of the prediction accuracy by means of Eq. (1). The statistical analysis of the Feature Selection section is tackled by means of a sensitivity analysis of the expected accuracy of VD prediction depending on combinations of features in order to determine the optimal choice of the number of features, underpinning claim 1. In the Pareto Front section we examine the Pareto front of feature combinations when taking into account prediction accuracy on the one hand and computational time complexity on the other hand, resulting in claim 2. Finally, in the Prediction Accuracy section we check the potential for improving the overall prediction accuracy of state-of-the-art prediction models by taking various feature combinations into account. This statistical analysis underpins claim 3 that the $HC_{\mathcal{D}}$ can be used to improve state-of-the-art prediction models in terms of both prediction accuracy and time complexity.

First, in the Experimental Setup section we outline the experimental setup based on publicly available assessment data.

Experimental Setup

Our experimental analysis relies on the publicly available database *KaistDB*.[23] This database shows the results of a subjective evaluation following the guidelines given in the recommendations by the International Telecommunication Union.[49] *KaistDB* encompasses VD assessments based on a five-point grading scale of 120 images provided by 20 subjects in terms of mean opinion scores (MOSs). For details concerning the analysis based on this database see Sohn et al.[15]

The disparity maps are computed by means of the *OpenCV* implementation[50] of the semi-global block matching algorithm.[51]

For a survey of the implemented features see Table I. To model the relation between features and VD, we employ *M5P* regression trees.[52,53] *M5P* combines a conventional decision tree with the possibility of linear regression functions at the nodes and therefore generates models that are compact and relatively comprehensible.

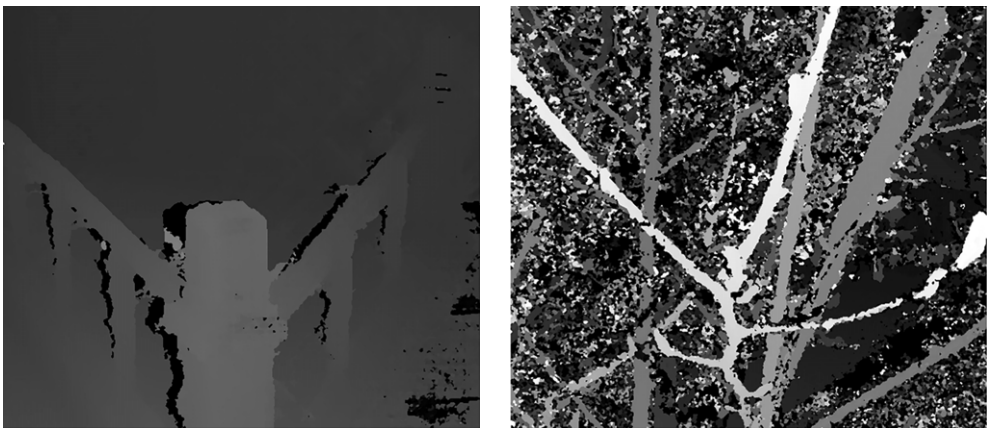
To quantify the prediction accuracy, we rely on the Pearson product-moment correlation coefficient (CC) between predicted scores and MOSs of VD.

Finally, to assess how well the model generalizes with respect to unknown data, we perform leave-one-out cross-validation.[54]

To check the statistical significance of our results, we rely on two different statistical tests: the non-parametric Mann–Whitney U test[55] for testing the mean values of sets for being different with significance level of 10^{-3} and a one-tailed Fisher-transformation test[56] for testing one sample having larger correlation coefficient than the other.

To measure the time complexity, we rely on *OpenCV* implementations[50] in Python and compute the mean evaluation time for all images. For the object-based approach of Sohn et al.,[15] we consider only the most time consuming

DOI: 10.2352/ISSN.2470-1173.2016.5.SDA-444



(a) (b)

Figure 3. Examples of disparity maps of stereoscopic images from Ref. 23: (a), a detail of a railing, shows low $HC_{\mathcal{D}}$, while (b), a crown of a tree, shows high $HC_{\mathcal{D}}$.

part, which is the pyramid-based mean-shift segmentation algorithm.[57] The evaluations were repeated 20 times on the 120 images of the database *KaistDB* on a Dell OptiPlex 990 (the images were resized to width 960 and height 540).

Feature Selection

This section analyzes the impact of the number of features on the prediction accuracy of state-of-the-art VD prediction models. In this analysis we focus on the eight features as outlined in Table I and analyze combinations together with the $HC_{\mathcal{D}}$ feature introduced in the third section. These nine features lead to a total number of $\sum_{k=1}^{9} B(9,k) = 511$ possible feature combinations, where $B(n,k)$ denotes the binomial coefficient. For this purpose, for each possible feature combination, the prediction function is trained and tested using leave-one-out cross-validation. Thus, we obtain a correlation coefficient for each prediction function.

Figure 4 shows the correlation coefficients obtained by the best combination for each number of features. The CC of the best combination increases with the number of features, and then decreases. This over-fitting phenomenon generally occurs when a model is excessively complex, such as having too many parameters relative to the number of observations. Restriction of the number of features is therefore a regularization measure that helps to improve the generalization behavior of the machine learning model (see, e.g., Ref. 58).

This analysis shows that, taking the nine features of Table I into account, a maximal number of four features is sufficient to predict VD, which underpins claim 1.

It is interesting to observe that the best combination (highest CC) with four features includes the $HC_{\mathcal{D}}$ feature.

Pareto Front

The statistical analysis of this section is devoted to the question of the best combinations considering prediction accuracy and time complexity. We are interested in feature combinations that are characterized by the property that it is impossible to improve time complexity without deteriorating prediction accuracy, and vice versa. Such combinations are called Pareto optimal solutions, lying on the Pareto front.[59] See Figure 5 for a graphical illustration of this analysis.

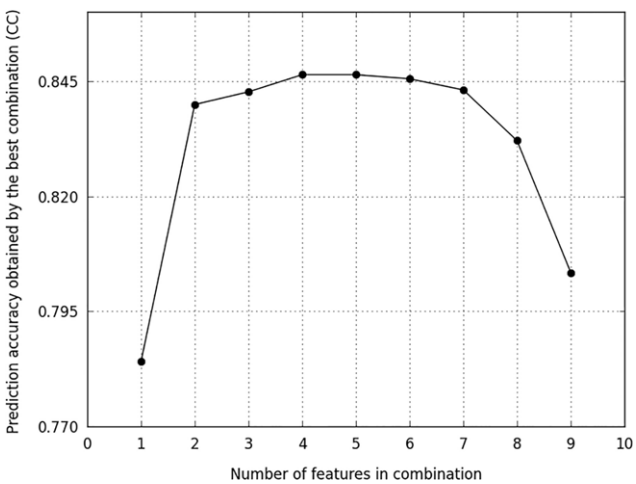


Figure 4. Analysis of prediction accuracy values of best combinations of possible features with size $n \in \{1,\ldots,9\}$. The accuracy is determined by means of the CC based on the ground truth data of Ref. 23.

The Pareto optimal solutions, with at most four features, are summarized in Table II. These combinations indicate that bottom-up approaches allow the construction of substantially faster prediction models without significantly lower prediction accuracy compared with the top-down approaches under consideration.

For example, consider the combination consisting of Var, Max5Sobel, $HC_{\mathcal{D}}$ and Mean. Evaluation of this combination is more than 58 times faster than for combinations with the object-dependent features of Ref. 15. Moreover, there is no combination (out of 511) without $HC_{\mathcal{D}}$ that shows statistically significant improvement of prediction accuracy. This hypothesis is statistically accepted by the one-tailed Fisher-transformation tests. In particular, the Z-values of this test lie in the range of $[0.0005; 0.6328]$ and the p-values

DOI: 10.2352/ISSN.2470-1173.2016.5.SDA-444

Table II. Pareto optimal solutions of combinations with at most four features of experiment on *KaistDB*.[23] Note that the $HC_{\mathcal{D}}$ feature is part of over 50% of the combinations, including the one with the highest CC.

Combination on Pareto front	Evaluation time (s)	CC
Mean	0.0004	0.300
Var	0.002	0.745
$HC_{\mathcal{D}}$	0.006	0.759
$HC_{\mathcal{D}}$, Mean	0.007	0.760
Var, $HC_{\mathcal{D}}$	0.008	0.770
Max5	0.016	0.771
Range10	0.016	0.781
Max5, Mean	0.017	0.796
Var, Max5, Mean	0.018	0.803
Max5, $HC_{\mathcal{D}}$	0.022	0.805
Max5, $HC_{\mathcal{D}}$, Mean	0.023	0.814
Var, Max5, $HC_{\mathcal{D}}$, Mean	0.024	0.816
Max5Sobel, $HC_{\mathcal{D}}$, Mean	0.057	0.817
Var, Max5Sobel, $HC_{\mathcal{D}}$, Mean	0.059	0.817
Var, OT, $HC_{\mathcal{D}}$, Mean	3.425	0.824
Max5, OT	3.432	0.838
Max5, OT, Mean	3.433	0.839
Var, Max5, OT	3.434	0.842
Var, OT, Max5, $HC_{\mathcal{D}}$	3.440	0.846
OT, Range10, Max5, $HC_{\mathcal{D}}$	3.455	0.846

lie in the range of [0.2634; 0.4998]. This statistical analysis underpins claim 1, stating that bottom-up approaches allow the construction of substantially faster prediction models without significantly lower prediction accuracy compared with the top-down approaches under consideration.

Table II and Fig. 5 also indicate that $HC_{\mathcal{D}}$ offers a reasonable trade-off between prediction accuracy and time complexity.

Prediction Accuracy

This section is devoted to the question of whether the $HC_{\mathcal{D}}$ feature can be used to improve the prediction accuracy of VD models using different combinations of features described in the third section.

Table III shows that feature combinations of two features including the $HC_{\mathcal{D}}$ feature substantially improve the prediction accuracy compared with the accuracy achieved by the single features only (improvement in CC of between 0.0092 and 0.5994).

For the analysis of all 511 possible combinations we separated the set of these combinations into two sets: one of combinations including $HC_{\mathcal{D}}$ and one of combinations without $HC_{\mathcal{D}}$. The results show that the mean CC (0.81) of the set of combinations with $HC_{\mathcal{D}}$ is higher than the mean CC (0.79) of the combinations without $HC_{\mathcal{D}}$. The Mann–Whitney U test shows that this result is significant ($U = 25823.0$, p-value $= 4.419 \times 10^{-5}$). Together with the combination with the highest prediction accuracy (see the Feature Selection section), this statistical analysis, of all possible combinations, underpins claim 3, stating that the expected prediction accuracy that can be achieved by combinations of features including $HC_{\mathcal{D}}$ is significantly higher than for combinations without $HC_{\mathcal{D}}$. Moreover, the feature combination with the highest prediction accuracy contains the $HC_{\mathcal{D}}$ feature.

Table III. Prediction performance of single features and the corresponding combination with $HC_{\mathcal{D}}$, based on the ground truth data given by the database *KaistDB*.[23] The italic numbers in the first two columns mark the best prediction accuracy of features listed in Table I for the database *KaistDB*.[23] The accuracy is measured based on the Pearson product-moment correlation coefficient (CC). Column 2 shows the accuracy results for combinations with the $HC_{\mathcal{D}}$ feature. The performance gain is shown in column 3.

Single feature	CC of single feature	CC of single feature+ $HC_{\mathcal{D}}$	Improvement in CC
RD	0.2465	0.7478	0.5013
OT	0.3807	0.7627	0.3820
Mean	0.2995	0.7603	0.4608
Var	0.7449	0.7698	0.0249
Range10	0.7806	0.7939	0.0133
Max5	0.7707	*0.8047*	0.0340
Max5Sobel	0.7761	0.8036	0.0275
Range10Sobel	*0.7841*	0.7933	0.0092
$HC_{\mathcal{D}}$	0.7593	0.7593	0.0000

CONCLUSION AND OUTLOOK

This article addresses state-of-the-art computational models for predicting VD. Starting with the standard second-order statistical approach based on co-occurrence matrices, as commonly used in texture analysis, we evolved the approach with a computationally efficient contrast feature based on the disparity map, the Haralick disparity contrast. Finally, experiment analysis shows that this feature improves prediction accuracy in combination with other features and, above all, offers a reasonable trade-off between prediction accuracy and time complexity. It remains for future research to apply the proposed method to real-time VD prediction for stereoscopic videos by also taking motion features into account.

APPENDIX. DISPARITY CONTRAST AS HARALICK FEATURE

Let us start with the co-occurrence matrix $h_{\delta,\theta} : \{1, \ldots, n\} \times \{1, \ldots, n\} \to [0, 1]$, where n denotes the number of gray levels. An entry $h_{\delta,\theta}(i, j)$ in the co-occurrence matrix represents the joint probability that a pair of pixels at distance δ and angle θ show the gray values i and j. The contrast feature proposed by Haralick[37] is defined by the four values $HC_\theta = \sum_{i,j=0}^{n-1} |i - j|^2 h_{1,\theta}(i, j)$ for the angles $\theta = 0, \pi/4, \pi/2, \pi$ and the pixel distance $\delta = 1$.

Instead of gray values let us consider disparity values that induce the co-occurrence matrix $h_{\delta,\theta}$. For horizontal

DOI: 10.2352/ISSN.2470-1173.2016.5.SDA-444

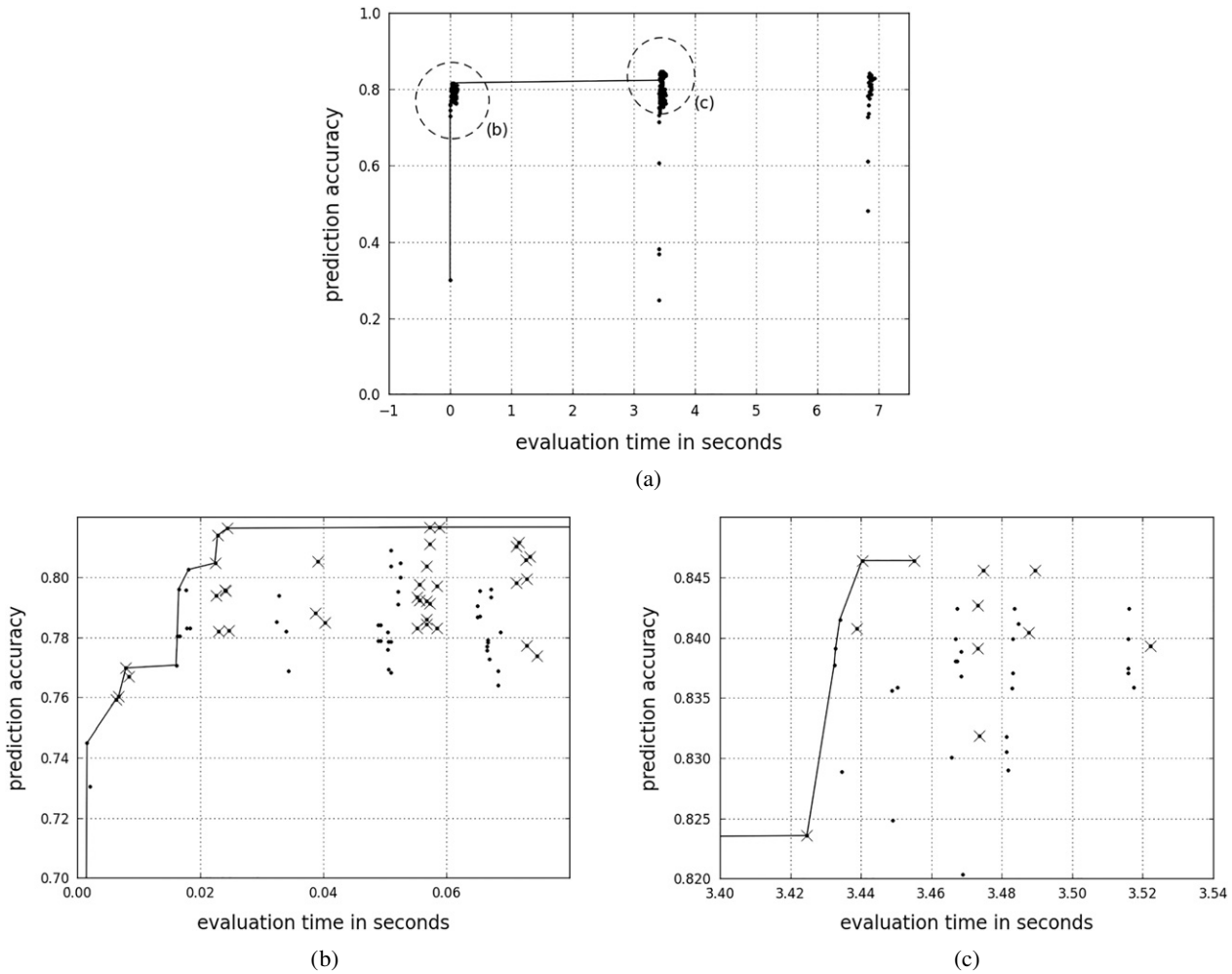


Figure 5. Pareto front of combinations with at most four features considering prediction accuracy and time complexity with evaluations based on *KaistDB*.[23] (b) and (c) scale up the details indicated by the encircled regions of graph (a). The solid lines represent the *Pareto front*, which includes the best solutions considering time complexity and prediction accuracy. The combinations marked by 'X' include the $HC_{\mathcal{D}}$ feature. Note that all the combinations in (b) are without object-dependent features *OT* and *RD*, while detail (c) shows only combinations including object-dependent features. Comparing the performance of feature combinations of (b) with (c), we observe a drastic reduction of time complexity by a speed up of about 100, while the prediction accuracy does not differ significantly.

shifts, i.e., $\theta = 0$, we obtain $HC_{\mathcal{D}}(\delta) = \sum_{\Delta_1,\Delta_2=-N}^{N}(\Delta_1 - \Delta_2)^2 h_{\delta,0}(\Delta_1, \Delta_2)$, where Δ_1 and Δ_2 encode the disparity values of the δ-neighboring pixels (x_1, y) and (x_2, y), i.e., $|x_2 - x_1| = \delta$. The notation $HC_{\mathcal{D}}$ emphasizes that the underlying co-occurrence matrix relies on disparity values. The entry in the co-occurrence matrix $h_{\delta,0}(\Delta_1, \Delta_2)$ is given by the probability that the disparities of two randomly chosen δ-neighboring pixels assume the values Δ_1 and Δ_2, respectively, i.e., $h_{\delta,0}(\Delta_1, \Delta_2) = \frac{\#_\delta[\Delta_1,\Delta_2]}{\sum_{\Delta_1,\Delta_2}\#_\delta[\Delta_1,\Delta_2]}$, where $\#_\delta[\Delta_1, \Delta_2] = \#\{((x_1, y), (x_2, y)) \in (\{1, \ldots, N\} \times \{1, \ldots, M\})^2 \mid \mathcal{D}(x_1, y) = \Delta_1, \mathcal{D}(x_2, y) = \Delta_2, |x_2 - x_1| = \delta\}$ and M, N denote the numbers of pixel rows and columns of the images and $\#$ the cardinality. Putting all together, we obtain $HC_{\mathcal{D}}(\delta) = \frac{1}{NM}\sum_{\Delta_1,\Delta_2}(\Delta_1 - \Delta_2)^2 \#_\delta[\Delta_1, \Delta_2] = \frac{1}{NM}\sum_{x,y}(\mathcal{D}(x, y) - \mathcal{D}(x + \delta, y))^2$, which can be interpreted as the expected squared disparity gradient with basis length δ.

ACKNOWLEDGMENT

This work was supported by the *Austrian Research Promotion Agency* (FFG) under the project *Hyperion3D*.

REFERENCES

[1] M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn, "Visual discomfort and visual fatigue of stereoscopic displays: A review," J. Imaging Sci. Technol. **53**, 030201–1 (2009).
[2] W. J. Tam, F. Speranza, S. Yano, K. Shimono, and H. Ono, "Stereoscopic 3D-TV: Visual comfort," IEEE Trans. Broadcast. **57**, 335–346 (2011).
[3] D. A. Goss and H. Zhai, "Clinical and laboratory investigations of the relationship of accommodation and convergence function with refractive error," Doc. Ophthalmol. **86**, 349–380 (1994).
[4] V. Kulyk, S. Tavakoli, M. Folkesson, K. Brunnstrom, K. Wang, and N. Garcia, "3D video quality assessment with multi-scale subjective method," *2013 Fifth Int'l. Workshop on Quality of Multimedia Experience (QoMEX)* (2013), pp. 106–111.
[5] U. Matthieu, B. Marcus, and L. C. Patrick, "How visual fatigue and discomfort impact 3D-TV quality of experience: a comprehensive review of technological, psychophysical, and psychological factors," Ann. Telecommun. **68**, 641–655 (2013).

[6] L. Onural, "3D video technologies: an overview of research trends," Proc. SPIE **7864B**, (2011).
[7] S. Yano, "Experimental stereoscopic high-definition television," Displays **12**, 58–64 (1991).
[8] S. Yano and I. Yuyama, "Stereoscopic HDTV: Experimental system and psychological effects," SMPTE J. **100**, 14–18 (1991).
[9] W. IJsselsteijn, H. de Ridder, R. Hamberg, D. Bouwhuis, and J. Freeman, "Perceived depth and the feeling of presence in 3DTV," Displays **18**, 207–214 (1998).
[10] R. G. Kaptein, A. Kuijsters, M. T. Lambooij, W. A. IJsselsteijn, and I. Heynderickx, "Performance evaluation of 3D-TV systems," Proc. SPIE **6808**, (2008).
[11] M. Lambooij, W. A. IJsselsteijn, and I. Heynderickx, "Visual discomfort of 3D TV: Assessment methods and modeling," Displays **32**, 209–218 (2011).
[12] Y. Nojiri, H. Yamanoue, S. Ide, S. Yano, and F. Okana, "Parallax distribution and visual comfort on stereoscopic HDTV," *Proc. IBC* (2006), pp. 373–380.
[13] J. Choi, D. Kim, B. Ham, S. Choi, and K. Sohn, "Visual fatigue evaluation and enhancement for 2D-plus-depth video," *IEEE 17th Int'l. Conf. on Image Processing (ICIP)* (IEEE, Piscataway, NJ, 2010), pp. 2981–2984, doi: 10.1109/ICIP.2010.5653389.
[14] Y. J. Jung, H. Sohn, S.-i. Lee, and Y. M. Ro, "Visual comfort improvement in stereoscopic 3d displays using perceptually plausible assessment metric of visual comfort," IEEE Trans. Consum. Electron. **60**, 1–9 (2014).
[15] H. Sohn, Y. J. Jung, S. il Lee, and Y. M. Ro, "Predicting visual discomfort using object size and disparity information in stereoscopic images," IEEE Trans. Broadcast. **59**, 28–37 (2013).
[16] X. Jin, G. Jiang, H. Ying, M. Yu, S. Ding, Z. Peng, and F. Shao, "A foreground object features-based stereoscopic image visual comfort assessment model," Proc. SPIE **9273**, 92730Q (2014).
[17] B. Michel, "Digital stereoscopy, scene to screen 3D production workflow," (Stereoscopic News, Belgium, 2013), ISBN 978-1-48015709-5.
[18] H. Sohn, Y. J. Jung, S.-i. Lee, F. Speranza, and Y. M. Ro, "Visual comfort amelioration technique for stereoscopic images: Disparity remapping to mitigate global and local discomfort causes," IEEE Trans. Circuits Syst. Video Technol. **24**, 745–758 (2014).
[19] C. Thébault, D. Doyen, P. Routhier, and T. Borel, "Automatic depth grading tool to successfully adapt stereoscopic 3D content to digital cinema and home viewing environments," Proc. SPIE **8648**, (2013), 86480X.
[20] F. Battisti, M. Carli, A. Stramacci, A. Boev, and A. Gotchev, "A perceptual quality metric for high-definition stereoscopic 3d video," Proc. SPIE **9399**, (2015).
[21] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," IEEE Trans. Syst. Man Cybern. **SMC-3**, 610–621 (1973).
[22] D. Kim and K. Sohn, "Visual fatigue prediction for stereoscopic image," IEEE Trans. Circuits Syst. Video Technol. **21**, 231–236 (2011).
[23] Y. Jung, H. Sohn, S.-i. Lee, H. Park, and Y. Ro, "Predicting visual discomfort of stereoscopic images using human attention model," IEEE Trans. Circuits Syst. Video Techn. **23**, 2077–2082 (2013).
[24] M. Wöpking, "Viewing comfort with stereoscopic pictures: An experimental study on the subjective effects of disparity magnitude and depth of focus," J. Soc. Inf. Disp. **3**, 101–103 (1995).
[25] S. Yano, M. Emoto, and T. Mitsuhashi, "Two factors in visual fatigue caused by stereoscopic HDTV images," Displays **25**, 141–150 (2004).
[26] Y. Nojiri, H. Yamanoue, A. Hanazato, M. Emoto, and F. Okano, "Visual comfort/discomfort and visual fatigue caused by stereoscopic HDTV viewing," Proc. SPIE **5291**, 303–313 (2004).
[27] A. Woods, "Understanding crosstalk in stereoscopic displays," *Keynote Presentation at the Three-Dimensional Systems and Applications Conference* (Tokyo, Japan) (2010), pp. 19–21.
[28] F. L. Kooi and A. Toet, "Visual comfort of binocular and 3D displays," Displays **25**, 99–108 (2004).
[29] F. Speranza, W. J. Tam, R. Renaud, and N. Hur, "Effect of disparity and motion on visual comfort of stereoscopic images," Proc. SPIE **6055**, (2006).
[30] A. J. Woods, T. Docherty, and R. Koch, "Image distortions in stereoscopic video systems," Proc. SPIE **1915**, 36–48 (1993).
[31] S. Ide, H. Yamanoue, M. Okui, F. Okano, M. Bitou, and N. Terashima, "Parallax distribution for ease of viewing in stereoscopic HDTV," Proc. SPIE **4660**, 38–45 (2002).
[32] R. Patterson and A. Silzars, "Immersive stereo displays, intuitive reasoning, and cognitive engineering," J. Soc. Inf. Disp. **17**, 443–448 (2009).
[33] M. T. M. Lambooij, W. A. Ijsselsteijn, and I. Heynderickx, "Visual discomfort in stereoscopic displays: a review," Proc. SPIE **6490**, (2007).
[34] S. Winkler, "Efficient measurement of stereoscopic 3d video content issues," Proc. SPIE **9016**, (2014).
[35] C. Zhu, Y. Zhao, L. Yu, and M. Tanimoto, *3D-TV System with Depth-Image-Based Rendering* (Springer, 2014).
[36] C. Bowers, B. R. Cowan, C. Creed, and G. Hakvoort, "Challenges of using stereoscopic displays in a touch interaction context," *Proc. BCS-HCI'14, Proc. 28th Int'l. BCS Human Computer Interaction Conference on HCI 2014 - Sand, Sea and Sky - Holiday HCI* (2014), pp. 276–280.
[37] R. M. Haralick, "Statistical and structural approaches to texture," Proc. IEEE **67**, 786–804 (1979).
[38] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," IEEE Trans. Pattern Anal. Mach. Intell. **33**, 2131–2146 (2011).
[39] S. J. Dickinson, H. I. Christensen, J. K. Tsotsos, and G. Olofsson, "Active object recognition integrating attention and viewpoint control," Comput. Vis. Image Underst. **67**, 239–260 (1997).
[40] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Matters of Intelligence* (Springer, 1987), pp. 115–141.
[41] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," Perception **28**, 1311–1328 (1999).
[42] F. L. Engel, "Visual conspicuity, visual search and fixation tendencies of the eye," Vis. Res. **17**, 95–108 (1977).
[43] A. Toet, F. L. Kooi, P. Bijl, and J. M. Valeton, "Visual conspicuity determines human target acquisition performance," Opt. Eng. **37**, 1969–1975 (1998).
[44] W. S. Geisler and K.-L. Chou, "Separation of low-level and high-level factors in complex tasks: visual search," Psychol. Rev. **102**, 356 (1995).
[45] E. Peli, "Contrast in complex images," J. Opt. Soc. Am. A **7**, 2032–2040 (1990).
[46] A. Hyvärinen, J. Hurri, and P. O. Hoyer, *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision*, Computational Imaging and Vision (Springer-Verlag London Limited, 2009), Vol. 39, ISBN 978-1-84882-490-4.
[47] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," Vis. Res. **42**, 107–123 (2002).
[48] R. Carmi and L. Itti, "Visual causes versus correlates of attentional selection in dynamic scenes," Vis. Res. **46**, 4333–4345 (2006).
[49] ITU-R, "Methodology for the subjective assessment of the quality of television pictures," Technical Report BT.500-11 (2002).
[50] G. Bradski, (2000) *"Opencv-library", Dr. Dobb's Journal of Software Tools.*
[51] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," IEEE Trans. Pattern Anal. Mach. Intell. **30**, 328–341 (2008).
[52] J. R. Quinlan, "Learning with continuous classes," *Proceedings of the 5th Australian joint Conference on Artificial Intelligence, Vol. 92* (Singapore, 1992), pp. 343–348.
[53] D. Culibrk, M. Mirkovic, V. Zlokolica, M. Pokric, V. Crnojevic, and D. Kukolj, "Salient motion features for video quality assessment," IEEE Trans. Image Process. **20**, 948–958 (2011).
[54] S. Geisser, *Predictive Inference, Vol. 55* (CRC Press, 1993).
[55] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," Ann. Math. Stat. **18**, 50–60 (1947).
[56] R. A. Fisher, "On the "probable error" of a coefficient of correlation deduced from a small sample," Metron **1**, 3–32 (1921).
[57] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Trans. Pattern Anal. Mach. Intell. **24**, 603–619 (2002).
[58] V. N. Vapnik, *Statistical Learning Theory, Vol. 2* (Wiley, New York, 1998).
[59] M. Ehrgott, *Multicriteria Optimization, Vol. 2*, Lecture Notes in Economics and Mathematical Systems, Vol. 491 (Springer, Berlin, Heidelberg, 2005).