SC-VQM – A Visual Quality Metric for Synthetic Contents

Christopher Haccius and Thorsten Herfet
Telecommunications Lab, Saarland University, Saarbrücken, Germany
Email: {haccius, herfet}@nt.uni-saarland.de

***Abstract*—Metrics approximating the perceived visual quality of image or video content play an important role in the development and usage of manifold processing and compression algorithms. Several visual quality metrics have been proposed in the past and produce good results. However, visual content is more and more influenced by synthetic content generation. Fully synthetic scenes or augmented reality contents present a new challenge to visual quality metrics. We present a Visual Quality Metric for Synthetic Contents (SC-VQM), which considers visual errors common to synthetic sources. We have tested our metric on an image quality database. Comparisons of the correlation between predicted Mean Opinion Score and subjective Mean Opinion Score show that our proposed Visual Quality Metric discriminates perceived visual quality significantly better than known standard quality metrics (PSNR, SSIM, HDR-VDP 2).**

***Index Terms*—subjective quality metric, visual quality, augmented reality, synthetic content, perceived quality**

I. Introduction

The quality of visual content is important whenever images or videos are presented to human observers. Quality here means subjective perception and the human judgment whether content has a high level of realism, contains few artifacts, little noise, or other noticeable degradations.

The question of image quality has been around for decades, and there have been several answers to this question. The best answer to this question is the Mean Opinion Score (MOS), as it is obtained by querying a large number of humans to evaluate the quality of given data. The MOS has first been used to evaluate the quality of communication channels, and therefore its specification can be found in a recommendation by the International Telecommunication Union. Here a communication channel is classified in five categories from excellent to poor by asking the users to rate the "difficulty in talking or hearing over the connection".[1]

While collecting user opinions gives the most credible results for the perceived quality, conducting an evaluation to obtain a MOS is expensive in both time and money, which is often infeasible for real world applications. Algorithmic approaches which evaluate the quality of a given content are necessary. These algorithms have the sole requirement of coming up with the same score that humans would assign to content; thus allowing to predict the MOS without having to conduct expensive surveys.

Existing metrics already offer decent solutions to approximate MOS algorithmically for "classical" image errors. Classical errors include forms of distortion introduced by either capture, coding or transmission, like random noise, illumination change or blocking artifacts. However, today more and more visual content is generated purely or partially from synthetic sources. Rendering synthetic content that is often algorithmically combined generates novel sources of image errors. Such errors include transformations of objects in a scene or texture changes. An example of the effect such an error can have is shown in Fig. 1. The lady in Figu. 1(a) and 1(b) looks very much the same. It requires close observation to note that the order of the stripes on her shirt has changed. Even if a human observer does notice, this neither alters the realism nor the perceived quality of the image, thus both images would receive a good MOS. A metric, however, would notice the large difference in image content, as shown in Fig. 1(c), thus assign a poor quality score to the tested image.

(a) Original (b) Wrong Shirt (c) Error Map

Figure 1. Perceived quality does often not correspond to image statistics © by Collage Vintage, Bette Franke

We propose a Visual Quality Metric for Synthetic Contents (SC-VQM) that not only considers the image statistics and the human visual system, but also cares for the human cognitive systems ability to neglect and auto-correct objects which have been misplaced by 3D transformations. This approach outperforms existing metrics on SID2015, a database with synthetic image distortions [1].

II. Related Work

Algorithmic image quality metrics are generally classified in three groups. Algorithms evaluating the image quality based purely on a test image are called no-

Manuscript received November 18, 2016; revised March 24, 2017.
[1] ITU-T Recommendation P.800: Methods for Subjective Determination of Transmission Quality.

doi: 10.18178/ijsps.5.1.1-6

reference metrics. If certain image features are known to the metric, but no full reference is available, such metrics are called partial reference metrics. Third, if a full reference image is available, metrics operating on this information are known as full reference metrics. While no-reference metrics can be used most comprehensively, they are also the most challenging ones. Most research advances have been made in the area of full-reference metrics, leading to an extensive coverage of such metrics in the literature (e.g. [2]-[4]). In the following subsections we will concentrate on metrics that have been designed for quality levels and usage scenarios comparable to our proposed metric. Full reference metrics can again be classified in three groups. We distinguish metrics based on image statistics, metrics considering the Human Visual System (HVS) and - with our proposed metric - metrics considering the Human Cognitive System (HCS).

A. Metrics Based on Image Statistics

For several decades image quality metrics were purely based on image statistics. These methods do not consider the human viewer at all: purely deviations of the content are considered. For a reference image R and a test image T with dimensions $x \times y$an average value expressing the overall statistical image error is the Mean Squared Error (MSE) calculated as

$$MSE = \frac{1}{x \cdot y} \sum_{i=0}^{x} \sum_{j=0}^{y} (R(i,j) - T(i,j))^2 \qquad (1)$$

To evaluate the impact of this difference, the MSE is often related to the original signal. The larger the amplitude of the original signal, the smaller the impact of a certain error. Such a relation is created by the Peak Signal to Noise Ratio (PSNR), which puts the noise measured by the MSE in relation to the peak signal power

$$PSNR = \frac{max_{i \in [0,x]} \left(max_{j \in [0,y]} \left(r(i,j)^2 \right) \right)}{MSE} \qquad (2)$$

The PSNR is still a very common metric for image quality analysis. It can be easily implemented and has a very low computational complexity, which is an important criterion for real-time applications. For many fields requiring information analysis, e.g. signal analysis in communication systems, PSNR is a tool of sufficient quality. For image quality assessment, however, PSNR relates poorly to subjective image quality findings. According to Wang *et al.* the correlation coefficient between PSNR and MOS is only at 0.3267 [5], which shows that PSNR is, even though widely used, not very useful for image quality assessment.

B. Metrics Considering the HVS

In 1997 Sarnoff *et al.* designed a vision model to algorithmically determine and rate the "Just Noticeable Difference" (JND) in images and videos, which became known as the Sarnoff Metric [6]. This metric was widely used, and JND became a quasi-standard for visual differences. However, until today, the metric remains under commercial license. In 2004 an alternative metric was developed by Wang *et al.* [7]. The Structural Similarity Index (SSIM), designed to match DMOS from objective evaluations, compares three different image components: luminance, contrast and structure. Structural similarity between a test and a reference image $SIMM(T,R)$ is calculated as the weighted product of luminance l, contrast c and structure s:

$$SSIM(T,R) = l(T,R)^{\alpha} \cdot c(T,R)^{\beta} \cdot s(T,R)^{\gamma} \qquad (3)$$

with $0 < \alpha, \beta, \gamma$.

Based on the ideas developed by Wang *et al.* in their works on structural similarity, Mantiuk *et al.* have extended this visual model to evaluate image qualities in more complex scenarios. A high dynamic range visible difference predictor (HDR-VDP) was introduced in 2005 [8] and completely overhauled in 2011, forming HDR-VDP 2 [9]. According to Mantiuk *et al.* the vision model presented in [9] is applicable to a wide range of viewing conditions, especially luminance changes.

III. Proposed Metric Based on the HCS

After light information has been captured by the eye it is processed by the human cognitive system. Here content is segmented and sorted into known classes. Sorting is often successful, even if the underlying information itself is insufficient for correctly assigning semantic meanings to image content. Optical illusions often employ these correction capabilities of the human brain, exemplary in the Hidden Dalmatian[2] or Kanizsa's Triangle[3] shown in Fig. 2.

(a) Hidden Dalmatian (b) Kanizsa's Triangle

Figure 2. Optical Illusions employing correction capabilities of the human brain (human cognitive system)

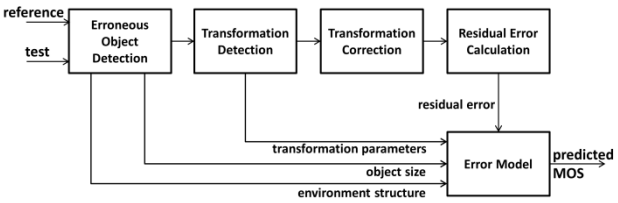


Figure 3. Structure of our proposed SC-VQM

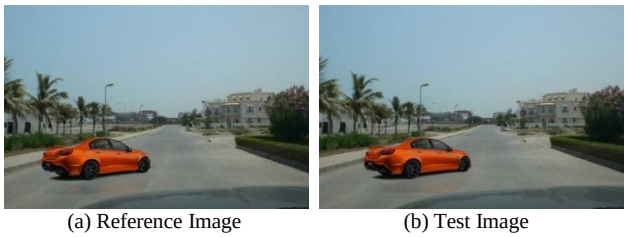
The SC-VQM is designed to detect changes in visual content that affect the perceived realism different to noise or compression artifacts, as they are corrected (partially) by the human brain. The idea to achieve this goal is straight forward: We detect object changes and correct those before calculating a residual error. This idea is outlined in the block diagram in Fig. 3 and described by the following six steps.

[2] http://www.moillusions.com/hidden-dalmation/
[3] https://en.wikipedia.org/wiki/Illusory_contours

1. *Erroneous object detection:* Distorted objects in a scene composition are detected
2. *Erroneous object matching:* Objects in test image are matched with objects in reference image
3. *Object size calculation:* The portion of the image affected by the distorted object is calculated
4. *Environment structure analysis:* The environment of the distorted objects is analyzed for the amount of structures contained
5. *Object correction:* The object in the test image is corrected according to the reference object, transformation parameters are recorded
6. *Residual error calculation:* The residual error between corrected object and reference image is calculated
7. *Approximate MOS by detected parameters:* All parameters from the previous analysis steps are combined in an error model to predict a MOS

A. Implementation

The SC-VQM is implemented along the idea outlined above. The following paragraphs explain the implementation details of the six steps and are visualized using the images shown in Fig. 4, a synthetic scene of a sports car on a street. In the test image shown in Fig. 4(b) the car is transposed with respect to the car shown in the reference image, Fig. 4(a).



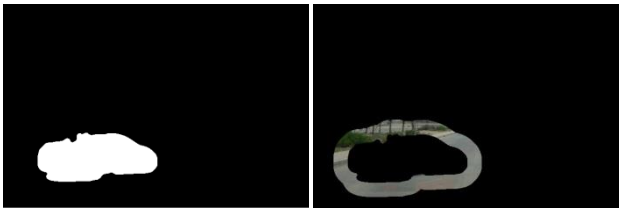
(a) Reference Image (b) Test Image

Figure 4. Optical illusions employing correction capabilities of the human brain (human cognitive system)

1. Erroneous object detection: A characteristic of erroneous objects is that image errors accumulate in the areas of these objects. We use this characteristic and in a first step compute the average image error as the MSE introduced in (1). If objects are misplaced the image error in these areas is above the average image error, while the error is below in other areas. By filtering the error areas with a disk shaped stencil, object areas can be distinguished. Two things are important to note: First, the object outline is only rough, but covers the whole area in which an object is misplaced with respect to the original. Second, the averaging disk size depends on image size and viewing conditions, to differentiate between noise and relevant objects.

The result of this detection step is a mask with outlined areas. If multiple objects in an image are moved, all of these areas are marked and noted. For the sample images shown in Fig. 4 the object detection mask is given in Fig. 5(a).

2. Erroneous object matching: To match objects between test and reference images there are two possible cases: a transformed object may be overlapping in reference and test image (only one erroneous region detected) or they may be spatially distinct (two erroneous regions). With the additional possibility to have several wrong objects in an image, we need to match each region with itself and with all other error regions. For region matching we employ Scale Invariant Features (SIFT) as proposed by Lowe [10]. For each area detected in the previous step we record the closest match between reference and test image. Fig. 6 shows detected features between reference (top) and test image (bottom). The translation of the car between test and reference image can already clearly be seen by the feature lines (white) running slightly tilted between both images.



(a) Mask outlining Erronous Object (b) Environment of Erroneous Object

Figure 5. Mask and environment of erroneous object

Figure 6. SIFT matching between test and reference image

3. Object size calculation: The size of a distorted object was observed to be critical for the perceived visual quality. We therefore calculate the size of an erroneous object, by considering the object masks calculated in the *Object Detection* step of two matching areas, as determined in the *Object Matching* step. The object size is given by the average pixel count of both matching object masks.

4. Environment structure analysis: A critical factor in the perceived quality degradation of object transformations is the amount of background structure. Unstructured backgrounds tend to 'hide' object transformations from human perception. To determine the amount of structure that is found in the environment of an object, we define an environment region that is proportional to the object size calculated in the previous step. The object environment is given by a boundary of this proportional size around the object mask, as returned from the object detection step. The environment for the sample images from Fig. 3 is given in Fig. 5(b).

To determine a single environment structure parameter we employ the edge detector proposed by Canny [11]. The amount of edge pixels found by this edge detector is normalized by the size of the structure environment,

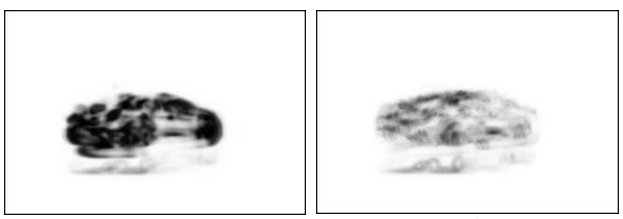
which is determined analog to the object size calculation above. This allows to have comparable structure parameters across distorted objects of different sizes.

5. *Object correction:* Reallocating the distorted object from the test image to its original position in the reference image is an important task to calculate the visual disturbance of the picture irrespective of any transformations. Initially, we remove the misplaced object from the test image and fill the created hole with a hole filling algorithm. Second, we use the SIFT feature correspondences to get a rough registration of the object in the test image [10]. As SIFT feature matching leaves inaccuracies in the order of single pixels we employ a Levenberg-Marquardt least-square optimization with a Fourier-Mellin transform module to achieve an image registration with sub-pixel precision for exact object placement [12]. The order of applying the SIFT registration before the Fourier-Mellin transform based registration is advantageous, as the SIFT registration works robustly, but with a certain inaccuracy, while the Fourier-Mellin transform becomes unstable for images that are too different from each other but works with a high precision when images are closely aligned already. Our implemented concatenation is both robust and precise. Finally, the registered object is fitted onto the filled background image. Filled background image and test image after object registration are shown in Fig. 7. Next to the registered image this step retrieves the scaling, translation and rotation values between reference and test object.

(a) Filled Background (b) Registered Object

Figure 7. Filled background and registered object in front of filled background

6. *Residual error calculation:* We calculate the residual error of the registered object in the filled image using SSIM, as introduced above. Here any other metric (SNR, PSNR, MS-SSIM, HDR-VDP2) could fit in, but SSIM is a widely used and well established metric, better conforming to the human visual system than purely statistical metrics like PSNR. For visualization purpose we show the SSIM maps of the original and of the corrected test image in Fig. 8. For the image quality assessment we consider the mean SSIM of the map shown in Fig. 8(b).



(a) Original SSIM Error Map (b) SSIM Map after Registration

Figure 8. SSIM before and after object registration

B. Calibration

The implementation as described in the previous section returns six parameters describing the test image, which are the residual error e, object size g, environment structure k, object rotation r, scaling s and translation t. We define a general relation between these error parameters and the predicted MOS (MOS_p) with the following properties:

- $MOS_p \propto e \; and \; MOS_p \propto r \; and \; MOS_p \propto s \; and \; MOS_p \propto t$
- $MOS_p \propto g \; and \; MOS_p \propto \frac{1}{k}$
- $If \; r = s = t = 0 \; then \; MOS_p \propto e$
- $If \; g = 0 \; then \; MOS_p \propto e$
- $If \; e = 0 \; then \; MOS_p \propto r, s, t$

Without any knowledge about weighting or gradient of the different factors, we propose a general model in which each parameter p may be tuned by a constant factor c_f and a constant exponent c_x, such that $c_f \cdot p^{c_x}$. This leads to a general definition of the underlying model

$$\begin{aligned} MOS_p = \quad & c_{f_e} \cdot e^{c_{x_e}} + \\ & c_{f_r} \cdot r^{c_{x_r}} \frac{g^{c_{x_{g1}}}}{k^{c_{x_{k1}}} + 1} + \\ & c_{f_s} \cdot s^{c_{x_s}} \frac{g^{c_{x_{g2}}}}{k^{c_{x_{k2}}} + 1} + \\ & c_{f_t} \cdot t^{c_{x_t}} \frac{g^{c_{x_{g3}}}}{k^{c_{x_{k3}}} + 1} \end{aligned} \tag{4}$$

This model has 14 free parameters, which need to be set in a calibration step. For calibration we have designed a dataset containing a scaled, rotated and translated cube (shown in Fig. 9) which has been evaluated by a small group of assessors. Based on the obtained MOS values we have obtained values for the 14 free parameters given above by fitting the error model in a least square sense. The obtained parameters are, in the order of their appearance in (4): $c_{f_e} = 7.2$, $c_{x_e} = 3.5$, $c_{f_r} = 10.3$, $c_{x_r} = 0.2$, $c_{x_{g1}} = 2.0$, $c_{x_{k1}} = 14.6$, $c_{f_s} = 4.5$, $c_{x_s} = 6.5$, $c_{x_{g2}} = 16.9$, $c_{x_{k2}} = 15.5$, $c_{f_t} = 12.1$, $c_{x_t} = 3.2$, $c_{x_{g3}} = 12.6$ and $c_{x_{k3}} = 17.8$. For fixed object size and environment structure the relation between measured error descriptors and MOS based on these parameters is plotted in Fig. 10. Noteworthy is that small residual errors, translation and scaling remain unnoticed, while small rotations directly lead to a decreased perceived quality. These parameters have then been used to calculate predicted MOS values on the SID2015 database.

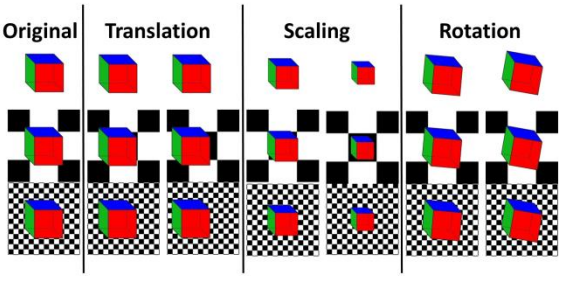


Figure 9. Test images for calibration

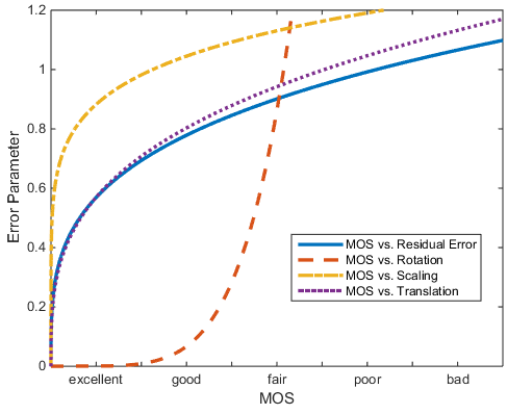


Figure 10. Visualization of free parameters from calibration

IV. Results

The correlation between the predicted MOS values MOS_p according to our proposed SC-VRM and the MOS values given for the SID2015 database are given in Table I as Spearmans ρ [13]. We also calculate the image qualities according to the metrics introduced in Section II, which are PSNR, SSIM, and HDR-VDP2, and calculate their correlation to the subjective MOS. It is important to note that our proposed metric employs the SSIM metric to calculate the residual error, since - out of the three metrics - SSIM results in the best correlation to the MOS values of SID2015. In the cases of classical image errors our metric falls back to a pure residual error calculation, which is therefore identical to the values achieved by SSIM.

TABLE I. Spearman – Correlation between MOS and PSNR, SSIM, HDR-VDP2 and SC-VQM

	JPEG	Noise	Transformation	Classical	All
PSNR	$\rho = \mathbf{0.72}$	$\rho = 0.59$	$\rho = 0.31$	$\rho = \mathbf{0.69}$	$\rho = 0.42$
SSIM	$\rho = 0.69$	$\rho = \mathbf{0.64}$	$\rho = 0.36$	$\rho = 0.67$	$\rho = 0.60$
HDR-VDP 2	$\rho = 0.51$	$\rho = 0.56$	$\rho = 0.24$	$\rho = 0.52$	$\rho = 0.37$
SC-VQM	$\rho = 0.69$	$\rho = \mathbf{0.64}$	$\rho = \mathbf{0.46}$	$\rho = 0.67$	$\rho = \mathbf{0.62}$
p-Test	—	—	$p = 0.38$	—	—

Adding 14 free parameters to an error model might allow to adjust a model fairly well even to uncorrelated data, thus falsely suggesting an improved correlation. We therefore perform hypothesis testing by calculating the p-value. In our case this is done by employing uncorrelated data (randomly distributed) for the five error descriptors that are returned from our analysis step (translation, rotation, scaling, structure and size). We then try to again calculate a best fit in a least-square sense for the 14 free parameters. If it is possible to generate an equally good or better correlation to the MOS values based on randomized input, the approach is necessarily wrong. The correlations for the hypothesis test as described are included in the last line of Table I. It is important to note that the best fit is achieved with $c_{f_r} = c_{f_s} = c_{f_t} = 0$: having 14 free parameters does not per se result in a better correlation between MOS and MOS_p.

V. Conclusion and Future Work

In this paper we have proposed a novel visual quality metric, which is especially potent for image distortions due to content manipulations enabled in synthetic contents. The proposed metric analyzes scene objects for transformations, and employs detected transformation parameters as well as the object size and its environment structure for visual quality prediction. The quality model is calibrated with a small set of test images, and our proposed metric is tested on the SID2015 database [1]. A comparison of correlations between the different metrics shows that our proposed metric increases the correlation between MOS and predicted MOS for transformation errors by 28%. The significance of this result is undermined by the ρ-test, which shows that the increased correlation is indeed a result of a correlation between the detected translation, rotation, scaling, size and environment structure, as well as the optimized residual error value.

For further validation and plausibility check we evaluated Fig. 1 with our proposed metric. While SSIM assigns a MOS score of "Fair" to the image ($MOS_p = 3$), our metric evaluated the test image close to "Excellent" ($MOS_p = 4.6$)

The presented approach is - to our knowledge - the only approach using computer vision algorithms to model parts of the human cognitive system for enhanced image quality assessment. As such, this initial work calls for a whole line-up of further questions and tasks. First of all, the correlation can probably be improved with a better calibration data-set, and more calibration data. In an optimal case a second data-set like SID2015 would be desirable, of which one can be used for metric calibration, the second for metric validation.

The error model we employ is based on a set of assumptions that are formulated in III.B. This error model results in an improved correlation between MOS and predicted MOS, but is not necessarily the best or the correct error model. Medical and psychological studies might be necessary to learn the correlation between perceived image distortion and object motion, based on object size and background structure.

Finally, geometric transformations cover an important part but not the full range of pre-rendering distortions that are possible in synthetic scenes. Lighting and texture, material properties or vertex normals are other parameters that influence visual results. Evaluating their influence on the perceived reality will be an important task for visual

quality predictors suitable for augmented and virtual reality scenarios.

REFERENCES

[1] C. Haccius and T. Herfet, "An image database for design and evaluation of visual quality metrics in synthetic scenarios," in *Proc. International Conference on Image Analysis and Recognition*, 2016.

[2] D. J. Bermejo, "High definition video quality assessment metric built upon full reference ratios," PhD thesis, Telecomunicaci ón, Polyt écnica Madrid, 2012.

[3] L. Jin, A. Gotchev, A. Boev, and K. Egiazarian, "Validation of a new full reference metric for quality assessment of mobile 3DTV content," in *Proc. 19th European Signal Processing Conference*, 2011, pp. 1894-1898.

[4] K. Okarma, "Combined full-reference image quality metric linearly correlated with subjective assessment," in *Artificial Intelligence and Soft Computing*, Springer, 2010, pp. 539–546.

[5] Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE International Conference on Image Processing*, 2002, p. I–477.

[6] J. Lubin and D. Fibush, *Sarnoff JND Vision Model*, 1997.

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[8] R. K. Mantiuk, S. J. Daly, K. Myszkowski, and H. P. Seidel, "Predicting visible differences in high dynamic range images: Model and its calibration," in *Electronic Imaging*, International Society for Optics and Photonics, 2005, pp. 204–214.

[9] R. K. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Transactions on Graphics (TOG)*, vol. 30, p. 40, 2011.

[10] D. G. Lowe, "Object recognition from local scale invariant features," in *Proc. Seventh IEEE International Conference on Computer Vision*, 1999, pp. 1150–1157.

[11] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 679–698, 1986.

[12] G. Wolberg and S. Zokai, "Robust image registration using log-polar transform," in *Proc. IEEE International Conference on Image Processing*, 2000, pp. 493-496.

[13] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

Prof. Thorsten Herfet was born in Bochum, Germany, on April 26th, 1963. Thorsten received a Diploma in Electrical Engineering in 1988 and a Ph.D. in telecommunication in 1992. Having been a PostDoc for 4 years he joined industry in 1996, finally being appointed Director of Research & Innovation of GRUNDIG. In 2004 he rejoined academia and became Full University Professor at Saarland University. His fields of research are cyber-physical networking, low latency streaming, computational videography and high mobility.

Prof. Herfet 2006-2008 served as the Dean for Informatics and Mathematics, in 2009 has been appointed Director of Research and Operations of the Intel Visual Computing Institute at Saarland University and since 2014 is Saarland University's Vice President for Research and Technology-Transfer. Thorsten published more than 100 papers, holds 15+ patents and has initiated and led several multi-million € collaborative research projects. He is a Senior Member of the IEEE, member of ACM SIGGRAPH, member of the German VDI/FKTG and serves as Steering Board member and Curator for various consortia and institutes.

Christopher Haccius was born in Lünen, Germany, on December 7, 1986. After his university-entrance diploma he received a BSc in Computer Science from the International University in Germany, Bruchsal, in 2009 and a MSc in Computer Science with focus on computer vision and telecommunications from Saarland University, Saarbrücken, Germany, in 2013. As a researcher his major fields of study are computer vision and telecommunications.

Mr. Haccius has work experience as a software developer with Fluid Operations and as a researcher at the telecommunications lab of Saarland University. Currently he is working for Continental Automotive in R&D of vehicle electronics in Wetzlar, Germany.