https://doi.org/10.2352/issn.2694-118X.2021.LIM-58

Spatial recall index for machine learning algorithms

Patrick Müller[1], Mattis Brummel[1], Alexander Braun
[1]Hochschule Düsseldorf, Düsseldorf, Germany

Abstract

We present a novel metric Spatial Recall Index *to assess the performance of machine-learning (ML) algorithms for automotive applications, focusing on* where *in the image which performance occurs. Typical metrics like intersection-over-union (IoU), precision-recall-curves or average precision (AP) quantify the performance over a whole database of images, neglecting spatial performance variations. But as the optics of camera systems are spatially variable over the field of view, the performance of ML-based algorithms is also a function of space, which we show in simulation: A realistic objective lens based on a Cooke-triplet that exhibits typical optical aberrations like astigmatism and chromatic aberration, all variable over field, is modeled. The model is then applied to a subset of the BDD100k dataset with spatially-varying kernels. We then quantify local changes in the performance of the pre-trained Mask R-CNN algorithm. Our examples demonstrate the spatial dependence of the performance of ML-based algorithms from the optical quality over field, highlighting the need to take the spatial dimension into account when training ML-based algorithms, especially when looking forward to autonomous driving applications.*

Introduction

The performance of vision-based machine learning algorithms for automotive applications strongly depends on the image quality of the input images used during training and validation. In our working group we therefore research the influence of optical systems on ML-based computer vision algorithms, trying to develop a process that enables that link between image quality and performance. Nonetheless, the topic of linking image quality to algorithmic performance has only started to gain academic and industrial traction within the last few years: In [12] Saad and Schneider consider the influence of vignetting on the performance of object detectors on the KITTI database. They additionally train a Deep Neural Network (DNN) on the augmented VKITTI dataset including a vignetting model and compare the distinct trained algorithms on real images from the KITTI database using mean Average Precision (mAP). They show a correlation between vignetting and the percentage of locally detected vehicles relative to all detected vehicles with respect to the image width. To describe the positional dependency they use bounding box center of mass. Pezzementi et al. examine in [11] the robustness of different person detectors to several image modifications assessed with common metrics. They investigate the influence on person detection for "Simple Mutators" such as invariant Gaussian blur and "Contextual Mutators" haze and defocus with no specific lens model applied.

However, any optical system is spatially variant, i.e. the optical quality in terms of aberrations varies over the field of view. Therefore, as a first step to gauge the spatial influence of the optical quality on the performance of ML-based algorithms we present a novel metric that spatially resolves the algorithmic performance. Furthermore, this metric can be used for any situation in which the content of the training images exhibits any sort of spatial dependence. To demonstrate the use of this new metric a large number of images from the BDD100k database [15] is degraded with a physical-realistic optical model based on Zernike-polynomials, which can be parameterised. Using different established detection algorithms [2, 6] the performance of these algorithms is measured, both traditionally and with our new approach. Finally, the spatially resolved metric can then be compared to the optical performance of the simulated lens, given as Full Width Half Maximum (FWHM) map.

This article is structured as follows. First we introduce the novel metric, followed by the application of the physical-realistic lens model of a Cooke-triplet. For this, we quantify the performance of the lens model in terms of different optical metrics. A defocus parameter is introduced in the model, as a technical parameter against which the algorithmic performance degradation may be compared.

Spatial recall index

The *spatial recall index* (SRI) quantifies the spatially resolved performance of an object detector, such as a pedestrian or car detection system. Similar to the

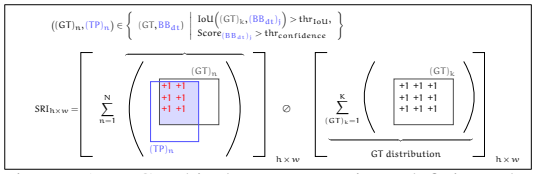


Figure 1: Graphical representation defining the $\mathrm{SRI}_{h\times w}$ distribution evaluated for a detection system on a database with images of size $h \times w$. Detected bounding boxes BB_{dt} matching a GT_k passing the IoU threshold and whose score exceed a predefined confidence threshold are considered to be True Positive TP bounding boxes. These tuples $((\mathrm{GT})_n, (\mathrm{TP})_n)$ are passed to the SRI evaluation as in Eq. 2, where $\oslash$ denotes the element-wise division of the matrices.

definition of the recall value [3] defined as

$$\mathrm{Recall} = \frac{\sum_n^N \mathrm{TP}_n}{\sum_k \mathrm{TP}_k + \mathrm{FN}_k} = \frac{\sum_n^N \mathrm{TP}_n}{\sum_k^K \mathrm{GT}_k} \tag{1}$$

where TP and FN are the True Positives and False Negatives for a particular database and the denominator represents the number of all ground truth objects GT, we present a *local* performance index, the SRI:

$$\mathrm{SRI}(x,y) = \frac{\sum_n^N \begin{cases} 1 & (x,y) \in [\mathrm{TP}_n \cap \mathrm{GT}_n] \\ 0 & \text{else} \end{cases}}{\left[\sum_k^K \begin{cases} 1 & (x,y) \in \mathrm{GT}_k \\ 0 & \text{else} \end{cases}\right]_{(x,y)\in \mathrm{GT}}} \tag{2}$$

In Eq. 2 the pixel location in the image is (x,y), $\mathrm{GT}_n, \mathrm{TP}_n$ represent the n-th bounding box of the ground truth and as *True Positive* labeled predicted bounding box, respectively. To apply the spatial recall index bounding boxes below a certain IoU threshold and score based on miss rate vs. False Positives per Image (FPPI) are eliminated to get True Positives TP from the set of all predicted bounding boxes. The index can now be evaluated from the remaining subset of predicted bounding boxes labeled as True Positives as defined in Eq. 2: To get the SRI at pixel (x,y) for each True Positive TP_n, the intersection of the bounding box with the corresponding ground truth bounding box GT_n is evaluated. Note that we take the intersection instead of the full bounding box to define a valid set of pixels. If the pixel (x,y) is inside the intersection it is counted to the index, otherwise nothing is added. The so acquired nominator value is weighted by the total number of ground truth bounding boxes overlapping at the pixel. This yields the spatial recall index at location (x,y) , the "probability" to locate an object correctly in a certain

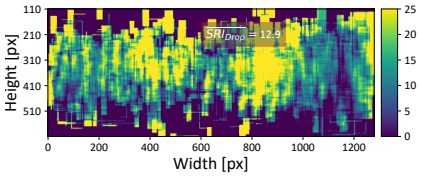


Figure 2: SRI performance drop as defined in Eq. 3 for the pre-trained Hybrid Task Cascade (HTC) [2] object detector for the class "person", where FPPI $= 0.1$, IoU $= 0.5$. A subset of the BDD100k validation dataset and the optical model $Z_\Delta = +1.25$ were used.

image region. The index fulfills some basic properties: An ideal detector would have the index $\mathrm{SRI}(x,y) = 1$ at each pixel and $\mathrm{SRI}(x,y) = 0$, if nothing has been detected. Further, to compare different SRI-distributions we define the SRI performance drop as:

$$\mathrm{SRI}_{Drop}(x,y) = \mathrm{SRI}_{Base}(x,y) - \mathrm{SRI}_{Z_\Delta}(x,y) \tag{3}$$

Note that the SRI distribution for a particular dataset and detector is defined only, if the database consists of images with same size, and if at least one ground truth bounding box is at the corresponding pixel (x,y).

Spatially variant image degradation and optical model

The effect of defocussing of a simple objective lens is gained by the application of a parameterised space-variant optical model. The model is based on the Zernike polynomials [13], from which we calculate a l_1-normed point-spread function (PSF) for each image pixel and color channel. The output is then used to degrade an image by a pixel-wise varying filter kernel.

We use a simple three-element lens configuration, a Cooke-Triplet, with basic ability to correct for chromatic aberration and astigmatism. The lens is simulated in the commercial software OpticStudio by Zemax [16], from which we export the first 20 Fringe-Zernike coefficients [13] to allow for parameterisation of the model. This is done for multiple positions over the imaging field and three wavelengths. For simplicity, we consider a rotationally symmetric system. We assume a setup with f#2.8, focal length 12.5mm, resolution $[1280 \times 720]$, pixel size 4.46µm and diagonal FoV $= 50°$. To get a pixel-resolved PSF over the imaging field, we linearly interpolate the imported coefficients in Zernike space between all samples for a particular wavelength and output a physical intensity PSF as outlined in [1, 5, 13]. The PSF is then rotated with respect to the corresponding off-axis position, scaled and cropped to match the physical pixel size of the assumed imager. Note that the model does not include a color filter array and thus no (de-) mosaicing effects are visible. The PSFs are validated by comparison with the PSF

export from Zemax. With these PSFs available in a self-developed Python framework, images are degraded by a pixel-wise, per-color varying convolution kernel. This is similar to superposition with limited support and the limiting case for isoplanar patches as in [10, 9]. Since, the presented performance metric for object detectors is based on IoU and thus sensitive to pixel shifts, Tilt should be taken into account. For the current optical model the Tilt_y is less than 0.57λ and Tilt_x is zero.

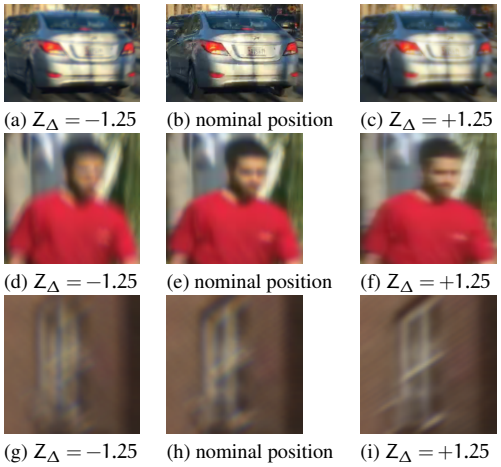
Defocus study

The defocussing parameter Z_Δ repositions the Cooke Triplet model between two extreme positions and serves as exemplary parameter to test the SRI. Defocussing of a objective happens in real situations, when e.g. materials expand due to heat. A constant offset Z_Δ is added to the defocus coefficient Z_2^0 in Zernike space, such that $\tilde{Z}_2^0 = Z_2^0 + Z_\Delta$, where $Z_\Delta \in [-1.25, -0.75, -0.5, 0, +0.5, +0.75, +1.25]\lambda$. From this parameterisation different sets of PSFs are derived as described above. Note that the nominal position at $Z_\Delta = 0$ already contains a spatially dependent defocus Z_2^0 according to the field curvature of the exported objective lens. Thus, the offset Z_Δ adds to or cancels the original contribution to the wavefront error W from the field curvature of the lens, and the positive and negative values of Z_Δ yield different results, as in any real lens. The impact of varying defocus is visualized in Fig. 3 displaying details from images of the BDD100k dataset: The left rear light from a *central* ROI in 3(a-c) appears most blurred for $Z_\Delta = +1.25$ followed by $Z_\Delta = -1.25$ and the nominal position. Contrary, the person taken from an *edge* ROI in Fig. 3(d-f) appears to be more blurred for $Z_\Delta = -1.25$, followed by the nominal position and $Z_\Delta = +1.25$ with noticeable astigmatism, visible at the person's face. Fig. 3(g-i) show different levels of chromatic aberration for another edge ROI.

Results

In this article, we examine the influence of the optical model on a car detection task. For this an image database is degraded with the optical model at different defocus offsets Z_Δ, resulting in seven additional databases. Subsequently, the results of the detectors for the unmodified database - the baseline - are compared with the results for the seven degraded databases, similar to [11].

We choose a subset from the diverse and huge Berkeley Deep Drive database (BDD100k) [15] with the tag "day". All images have the same resolution of 1280x720x3 pixels[15]. Table 1 lists the statistics of the chosen subset. To investigate a possible correlation between the performance of object detection



(a) $Z_\Delta = -1.25$ (b) nominal position (c) $Z_\Delta = +1.25$

(d) $Z_\Delta = -1.25$ (e) nominal position (f) $Z_\Delta = +1.25$

(g) $Z_\Delta = -1.25$ (h) nominal position (i) $Z_\Delta = +1.25$

Figure 3: Example scenes from [15] with varying defocus from a **central** (a-c) and **edge** (d-i) locations.

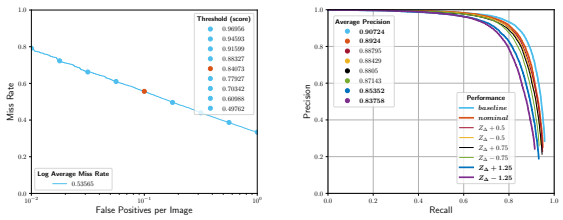
BDD100k subset	Selection	# Images	# Persons	#Cars
Training	Day, all bbox sizes	36728	25450 (fully visible)	402222
Validation	Day, all bbox sizes	5258	3651 (fully visible)	58283
Validation	Day, medium size	5258	1908 (fully visible)	21041

Table 1: Number of bounding boxes (bbox) for different classes and selections from the BDD100k database using the tags from [15], where $\mathrm{Medium}[32^2, 96^2]\mathrm{px}^2$.

algorithms and the optical performance, two different pre-trained detectors for the classes "person" and "car" on the BDD100k database are considered. In this article however, we discuss only results for the "car" detector, which we take from the Detectron2 project [14] the Mask R-CNN [6] algorithm, pre-trained on the COCO dataset [8] with state-of-the-art performance at the COCO dataset.

Standard performance evaluation

The performance of object detectors is often judged by their corresponding Miss Rate vs. FPPI or Precision-Recall curves.[3, 4] Fig. 4a displays the Miss Rate vs. FPPI curve for the car detector, where the log average miss rate for the "car" detector is 0.56. Note, that the Miss Rate depends on bounding box size and is higher for smaller bounding boxes. From this, we count bounding boxes above the confidence score threshold as listed in Fig. 4a, and IoU > 0.5 as True Positive. Fig. 4b shows the precision-recall curve for the car detector. The AP drops by 1.5% for the medium bounding boxes and the optical model in nominal position. It continues to fall, if we increase the defocus offset Z_Δ: The AP drops by 7.0% for $Z_\Delta = -1.25$ and by 5.4% for $Z_\Delta = +1.25$ as visualized in Fig. 4b. Note that the AP for *all* bounding boxes drops by 2.8% as we apply the optical model. If we add different defocus offsets Z_Δ, the AP decreases by 10.5% and 11.1% for $Z_\Delta = \pm 1.25$. Thus, the AP is a function of bounding box size.



(a) (b)

Figure 4: Standard performance metrics for car detection [6]. (a) Miss Rate vs. FPPI. (b) Precision vs. Recall for medium bounding boxes and different defocus parameter settings. These values correlate with the APs.

Spatially variant performance evaluation

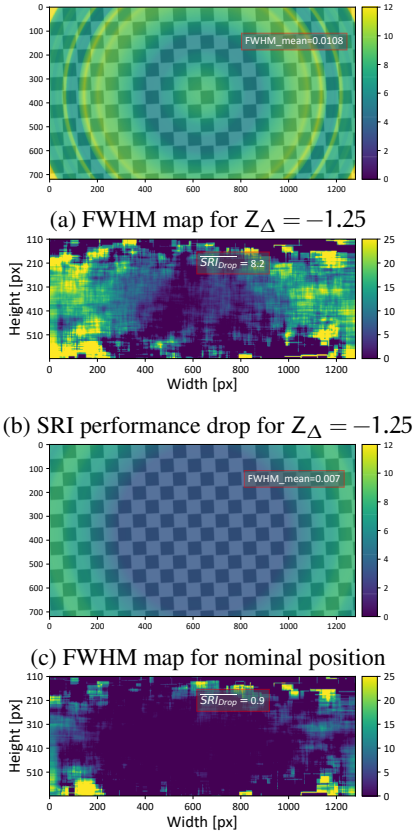
We examine the SRI distribution for different defocus parameter settings on the validation subset of the BDD100k at daytime for the class "car". For each detector we choose in the linear part of the Miss Rate vs. FPPI curve the central point FPPI $= 0.1$ in Fig. 4a.

First the SRI, as defined in Eq. 2, is evaluated at each available pixel (x, y) for the particular class and detector system, which yields a baseline distribution for the undisturbed dataset. Second, SRI-distributions are evaluated for the different defocus degradations of the dataset. The optical performance is given by maps of the Full Width Half Maximum (FWHM) evaluated for each PSF, where the two-dimensional distribution is reduced to $FWHM_{total} = \sqrt{FWHM_x^2 + FWHM_y^2}$.

Comparing optical and detector performance metrics

Fig. 5(a,c) visualize the FWHM maps for different defocus offset values and Fig. 5(b,d) show the SRI performance drop for the "car" detector and medium bounding box sizes. For this subset the resulting ground truth distribution is more equal than for small bounding boxes, which are mainly located at the center as they represent distant cars - cf. the optical flow. Large bounding boxes refer to parking cars or cars close to the observer and are easily detected. Because of their large bounding box size, local effects are masked and local effects are smoothed.

Fig. 5d shows good detection results for the optical model applied at nominal position, recap the 1.5% points loss in AP. However, the performance drop is spatially dependend and increases towards the edge. This spatial dependence is readily visible for the FWHM map in 5c. Introducing a strong defocus offset $Z_\Delta = -1.25$, Fig. 5a, results in a drop of 7.0% points in AP. Now, Fig. 5b indicates a clear performance drop close to the edge, while the central region is less affected. Also, the performance drop follows the rotational symmetric field dependence as in 5a.



(a) FWHM map for $Z_\Delta = -1.25$

(b) SRI performance drop for $Z_\Delta = -1.25$

(c) FWHM map for nominal position

(d) SRI performance drop for nominal position

Figure 5: FWHM maps and SRI performance drop for medium sized car bounding boxes, FPPI $= 0.1$, IoU $= 0.5$ and selected defocus offsets $Z_\Delta \in [-1.25, 0]$. Higher values indicate lower performance.

Conclusion

In this article, we present the Spatial Recall Index (SRI), which can be evaluated for an object detector system and a sufficient large and properly distributed image database. We applied a parameterised optical model based on Zernike Polynomials to a subset of the BDD100k database. We have shown a local correlation between the optical performance and the object detector's performance, measured as local SRI distribution.

Although, SRI is highly sensitive to pixel shifts and bounding box size, for which future work should account for, necessary properties of the examined database, such as required scene versatility and its influence on the metric need to be investigated, spatial performance metrics for ML may be crucial for industry pipelines and improve the development process of advanced driver assistance systems (ADAS). Moreover, future work extends the research to instance and panoptic segmentation and other optical models such as [7].

References

[1] Max Born and Emil Wolf. *Principles of optics: electromagnetic theory of propagation, interference and diffraction of light*. 7th expanded ed. Cambridge ; New York: Cambridge University Press, 1999.

[2] Kai Chen et al. "Hybrid Task Cascade for Instance Segmentation". en. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, June 2019, pp. 4969–4978.

[3] Jesse Davis and Mark Goadrich. "The relationship between Precision-Recall and ROC curves". en. In: *Proceedings of the 23rd international conference on Machine learning - ICML '06*. Pittsburgh, Pennsylvania: ACM Press, 2006, pp. 233–240.

[4] P. Dollar et al. "Pedestrian Detection: An Evaluation of the State of the Art". en. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.4 (Apr. 2012), pp. 743–761.

[5] Joseph W. Goodman. *Introduction to Fourier optics*. Fourth edition. New York: W.H. Freeman, Macmillan Learning, 2017.

[6] K. He et al. "Mask R-CNN". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. ISSN: 2380-7504. Oct. 2017, pp. 2980–2988.

[7] Matthias Lehmann et al. "Resolution and accuracy of nonlinear regression of point spread function with artificial neural networks". In: *Optical Engineering* 58.04 (Apr. 2019), p. 1.

[8] Tsung-Yi Lin et al. "Microsoft COCO: Common Objects in Context". en. In: *Computer Vision – ECCV 2014*. Ed. by David Fleet et al. Vol. 8693. Series Title: Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014, pp. 740–755.

[9] Patrick Müller, Matthias Lehmann, and Alexander Braun. "Optical quality metrics for image restoration". en. In: *Digital Optical Technologies 2019*. Ed. by Bernard C. Kress and Peter Schelkens. Munich, Germany: SPIE, June 2019, p. 37.

[10] James G. Nagy and Dianne P. O'Leary. "Fast iterative image restoration with a spatially varying PSF". en. In: ed. by Franklin T. Luk. San Diego, CA, United States, Oct. 1997, p. 388.

[11] Zachary Pezzementi et al. "Putting Image Manipulations in Context: Robustness Testing for Safe Perception". en. In: *2018 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. Philadelphia, PA: IEEE, Aug. 2018, pp. 1–8.

[12] Kmeid Saad and Stefan-Alexander Schneider. "Camera Vignetting Model and its Effects on Deep Neural Networks for Object Detection". en. In: *2019 IEEE International Conference on Connected Vehicles and Expo (ICCVE)*. Graz, Austria: IEEE, Nov. 2019, pp. 1–5.

[13] Jim Schwiegerling. *Optical specification, fabrication, and testing*. Bellingham, Washington: SPIE Press, 2014.

[14] Yuxin Wu et al. *Detectron2*. 2019.

[15] F. Yu et al. "BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. ISSN: 2575-7075. June 2020, pp. 2633–2642.

[16] Zemax. *OpticStudio — Optical, Illumination & Laser System Design Software*. 2021.

Author Biography

Patrick Müller received his B.Eng. in 2016 and his M.Sc. in 2018. His Master's thesis examined the influence of a Point Spread Function Model to Digital Image Processing algorithms. He is currently pursuing his PhD with a focus on the application of optical models to digital images, their validation, performance and correlation with the performance of Computer Vision algorithms.

Mattis Brummel is a Master student at Hochschule Düsseldorf in Electrical Engineering and Information Technology. He received his Bachelor's degree from FH Bielefeld. His research interests cover deep learning and computer vision.

Alexander Braun received his diploma in physics with a focus on laser fluorescent spectroscopy from the University of Göttingen in 2001. His PhD research in quantum optics was carried out at the University of Hamburg, resulting in a Doctorate from the University of Siegen in 2007. He started working as an optical designer for camera-based ADAS with the company Kostal, and became a Professor of Physics at the University of Applied Sciences in Düsseldorf in 2013, where he now researches optical metrology and optical models for simulation in the context of autonomous driving. He's member of DPG, SPIE and IS&T, participating in norming efforts at IEEE (P2020) and VDI (FA 8.13), and currently serves on the advisory board for the AutoSens conference.