

Machine Learning based Predictions of Subjective Refractive Errors of the Human Eye

Alexander Leube^{1,2,*}, Christian Leibig^{1,*}, Arne Ohlendorf^{1,2} and Siegfried Wahl^{1,2}

¹*Institute for Ophthalmic Research, Eberhard Karls University, Tübingen, Germany*

²*Carl Zeiss Vision International GmbH, Technology and Innovation, Aalen Germany*

Keywords: Big Data, Machine Learning, Subjective Refraction.

Abstract: The aim of this research was to demonstrate the suitability of a data-driven approach to identify the spherocylindrical subjective refraction. An artificial deep learning network with two hidden layers was trained to predict power vector refraction (M, J₀ and J₄₅) from 37 dimensional feature vectors (36 Zernike coefficients + pupil diameter) from a large database of 50,000 eyes. A smaller database of 460 eyes containing subjective and objective refraction from controlled experiment conditions was used to test for prediction power. Bland-Altman analysis was performed, calculating the mean difference (eg ΔM) and the 95% confidence interval (CI) between predictions and subjective refraction. Using the machine learning approach, the accuracy ($\Delta M = +0.08D$) and precision (CI for $\Delta M = \pm 0.78D$) for the prediction of refractive error corrections was comparable to a conventional metric ($\Delta M = +0.11D \pm 0.89D$) as well as the inter-examiner agreement between optometrists ($\Delta M = -0.05D \pm 0.63D$). To conclude, the proposed deep learning network for the prediction of refractive error corrections showed its suitability to reliably predict subjective power vectors of refraction from objective wavefront data.

1 INTRODUCTION

The current gold standard for the measurement of refractive errors of the eye is the subjective refraction that aims to correct the lower aberrations of the eye in order to provide the best retinal image quality (Goss and Grosvenor, 1996). The intra-examiner agreement which is defined as the agreement between different measurements of the refractive error by the same examiner, as well as the agreement between the subjective correction from multiple examiners (inter-examiner agreement) are in the range between $\pm 0.25D$ (80%) and $\pm 0.50D$ (95%) (Goss and Grosvenor, 1996). The subjective ability to judge the level of focus depends on the depth of focus of the eye, which is known to be around $\pm 0.30D$ (Leube et al. 2016) and higher order aberrations whose influence is modulated by the size of the pupil (Wang et al. 2003). Other factors that can affect the successful computation of subjective corrections is the ability of the visual system to easily adapt to blur and contrast. For example, it is well known that the eye is adapted to its own aberrations and can easily

adapt to spherical as well as astigmatic defocus (Artal et al., 2004; Ohlendorf and Schaeffel, 2009; Ohlendorf et al., 2011).

Pioneered by Applegate, Williams, Thibos and others, computational attempts have been made in order to predict the refractive correction of sphere, cylinder and its axis from the monochromatic lower and higher order aberrations of the eye, using image quality metrics (Applegate et al. 2003; Guirao and Williams 2003; Thibos et al. 2002; Thibos et al. 2004a). In general, one can distinguish between pupil-plane metrics that are used to optimize the wavefront aberrations of the eye and image-plane metrics, which are used to maximize the quality of the retinal image in relation to the subjective best image quality (Guirao and Williams, 2003). While comparing pupil-plane as well as image-plane metrics to the subjectively best correction of refractive errors in a cohort of 147 eyes, Guirao and Williams (2003) concluded that image-plane metrics are the better choice to predict the subjective refraction. In support to this findings, Thibos et al. (2004) extended the use of pupil-plane metrics as well as image-plane metrics

* Both authors contributed equally to this work.

and presented 33 different metrics. Their accuracy (defined by the mean error of the prediction) as well as precision (equivalent to the 95% limits of agreement) to predict the subjective refraction in a population of 200 eyes was shown previously (Thibos et al. 2004). The range of the mean difference for the estimated spherical defocus was reported to be between -0.50D and +0.25D for all metrics, while the limits of agreement (LoA) ranged from $\pm 0.50D$ to $\pm 1.00D$.

The mapping from detailed, high-dimensional measurements of low- to high-order aberration errors to the best possible subjective refraction correction includes both, optical and neural factors that are difficult to assess with detailed physical models. Using a training set of paired objective and subjective measurements, the task of predicting the optimal subjective correction from an objective measurement is essentially a (non-linear) regression problem. While using for example a deep learning network in order to allow enough space for a flexible function, the prediction of refractive corrections should be learnable from data. Further, it should perform at least as good as detailed physical models, while the former does not require the detailed mechanistic knowledge of the latter. Additionally and by construction, such a regressor inherently captures optical and neural factors, while it is assumed that these are consistent across subjects. The aim of the current research was to explore the applicability of machine learning approaches for the prediction of the spherocylindrical correction of refractive errors and compare their performance (in terms of accuracy and precision) against currently used objective metrics and subjective measurements.

2 METHODS

2.1 Datasets

Three datasets were used: (1) An excerpt from a data base of spectacle lens orders (provided by Carl Zeiss Vision GmbH, Aalen, Germany) that included monocular data from 50,496 eyes, measured by various professional optometrists. Each eye in this database was characterized by its aberrometry data and pupil size measured with a wavefront aberrometer (i.Profiler 1, Carl Zeiss Vision GmbH, Aalen, Germany) and a subjective measurement of its refractive errors. Aberrometry data were assessed up to the 7th radial order and included refractive data such as the spherical equivalent M from -16.00D to +9.00D with a mean of $-0.85 \pm 2.67D$. This large data

set was divided into a training set (32,316 samples, coined *ZV train*), a developing set (8,080 samples, coined *ZV develop*) and a testing set (10,100 samples, coined *ZV test*). (2) A second set of monocular data was collected independently for research purposes with a Subjective Refraction Unit (SRU) and the wavefront aberrometer (i.Profiler plus, Carl Zeiss Vision GmbH, Aalen, Germany) from 460 eyes (Ohlendorf, Leube and Wahl 2016). The SRU included a digital phoropter (ZEISS Visuphor 500, Carl Zeiss Vision GmbH, Aalen, Germany) and a LCD-screen to display optotypes (ZEISS Visuscreen 500, Carl Zeiss Vision GmbH, Aalen, Germany) with a minimum luminance of $L = 250\text{cd/m}^2$. SLOAN letters were used as optotypes and were shown in an EDTRS layout (Sloan, 1959; National Eye Institute, 1991). This dataset was used exclusively for testing purposes (simply called *EagleEye*) to discriminate against *ZV test*. (3) A third dataset was used to examine the inter-examiner agreement of monocular subjective refractions in a group of 54 eyes, measured by two of the authors both using the standardized procedure.

2.2 Machine Learning Methodology

An artificial deep learning network was used to compute the best correction of the power vectors of the refractive errors for a given set of low- and high-order aberrations of an eye. Therefore, the target values t drawn out of the power vector entries $\{M, J_0, J_{45}\}$ were predicted by using the set of Zernike coefficients together with the pupil diameter as input features for any given eye. Every eye was hence objectively characterized by concatenating these features into a vector $x \in \mathbb{R}^{37}$. The non-linear function $y = f(x, \theta)$ to predict the subjectively assigned value t was learned by using the respective target values from the training data set (*ZV train*, 32,316 samples): The networks' trainable parameters θ were learned by backpropagating the errors between predictions and target values. The separate development set of data (*ZV develop*, 8,080 samples) was used to determine all necessary hyperparameters such as the network architecture described further below. The two separate and independent test data sets were then used to report final performance for the sub data set *ZV test* that constituted a non-overlapping split from the same database used for training that was comprised of 10,100 samples.

2.3 Model Development

In principle, any (non)-linear regression model that fits the training data and generalizes to unseen test data is conceivable. Here, we chose deep learning networks because their effective capacity can be tuned flexibly to the problem at hand. Initial experiments (data not shown) ruled out a simple linear model and Bayesian hyperparameter optimization using hyperopt, suggested a similar model to the one developed here by using current best principles (Bergstra et al. 2013; Goodfellow, et al. 2016; Shahriari et al., 2016). More specifically, development converged to using multi-layered feedforward networks with a single linear output unit to jointly learn the basis functions and predictions by minimizing the mean squared error $1/N \sum^N (y-t)^2$ between predicted and subjectively assigned spherocylindrical corrections. We trained three networks in total, one each for M, J_0 and J_{45} respectively. Every network had two hidden layers, with 64 hidden units each and parametric rectifier nonlinearities (He et al., 2015). Initial weights were drawn from a zero-mean Gaussian with a standard deviation of $\sqrt{2/n}$ with n the number of incoming connections (He et al., 2015). All data were preprocessed by subtracting the mean of each feature dimension calculated across the training set. The network was trained using the Adam optimizer with a learning rate of 0.001 which was automatically reduced by a factor of 0.2 whenever the loss on the validation set did not improve for 10 epochs (passes through the training data) (Kingma and Ba, 2015). To avoid overfitting, small network weights were encouraged via global $L2$ regularization ($\lambda = 0.02$) and randomly dropping 0.2 of the hidden units (Bishop, 2006; Hinton et al., 2012; Srivastava et al., 2014). Training was performed with a batch size of 128 and stopped, when the loss evaluated on the validation set did not improve for 50 epochs ("early stopping").

2.4 Implementation

The machine learning models were developed in Python, using the scientific computing stack and in particular the deep learning library keras together with tensorflow as backend (Francois, 2016; Girija, 2016). Network related computations were performed on the graphics processing unit (GeForce GTX 1080 GPU, NVIDIA, Santa Clara, USA).

Code and models will be publicly available upon publication under <https://github.com/chleibig/airefraction>.

2.5 Validation and Comparison with Alternative Methods

To compare the performance of the machine learning approach to (1) existing subjective measurements as well as (2) alternative computational approaches and (3) to the inter-examiner agreement of subjective refractions, the following analysis was performed: accuracy and precision in terms of Bland-Altman analysis (Bland and Altman, 1986) of the machine learning approach was compared to power vectors of refraction (Thibos et al. 1997) from (1) subjective refractions and (2) computations based corrections using the visual Strehl of the optical transfer function (VSOTF) (Thibos et al., 2004a), both for the described dataset *EagleEye*. Inter-examiner agreement was established for a smaller set of eyes (dataset 3, $n = 54$ eyes) in order to be able to better rank the obtained results in the framework of the subjective assessment of refractive errors.

3 RESULTS

3.1 Testing of the Deep Learning Network

Using the sub dataset *ZV train*, the deep neuronal network was trained and further tested, while using the additional sub datasets *ZV develop*. In a first step, the performance in terms of accuracy and precision of the deep learning network was assessed, while using the data-set *ZV test*. The results revealed mean differences (accuracy) that were around zero for the three power vectors (M, J_0 and J_{45}), while the 95% limits of agreement (precision) were twice as big for M as for the cylindrical vector components J_0 and J_{45} (see Table 1). In a second step, the true generalization of the deep learning network was analyzed, while using the independent dataset *EagleEye* and again, mean differences were around zero for the three power vectors M, J_0 and J_{45} , while the 95% limits of agreement was higher for M, when compared to J_0 and J_{45} . Since the developed deep learning network performed similar on both datasets (*ZV test* and *EagleEye*), it will be also applicable to data from potentially different distributions.

Table 1: Mean differences and 95% limits of agreement between computationally and subjectively assessed refraction for the power vector components (M, J0 and J45).

	Refractive component	Mean difference, D (95% CI)	95% limits of agreement, lower, D (95% CI)	95% limits of agreement, upper, D (95% CI)
Deep network (ZV test, n = 10100)	M	0.01 (0.02, 0.01)	-0.64 (-0.66, -0.63)	0.67 (0.66, 0.68)
	J ₀	0.01 (0.01, 0.00)	-0.31 (-0.32, -0.30)	0.32 (0.31, 0.33)
	J ₄₅	0.00 (0.00, 0.00)	-0.28 (-0.29, -0.27)	0.28 (0.27, 0.29)
Deep network (EagleEye, n = 460)	M	0.08 (0.12, 0.05)	-0.70 (-0.76, -0.63)	0.86 (0.80, 0.92)
	J ₀	-0.01 (0.00, 0.00)	-0.31 (-0.37, -0.25)	0.28 (0.22, 0.35)
	J ₄₅	0.01 (0.02, 0.00)	-0.23 (-0.29, -0.17)	0.25 (0.19, 0.31)
Visual Strehl OTF (EagleEye, n = 460)	M	0.11 (0.15, 0.06)	-0.78 (-0.86, -0.71)	1.00 (0.92, 1.07)
	J ₀	-0.02 (0.00, -0.04)	-0.40 (-0.47, -0.33)	0.36 (0.29, 0.43)
	J ₄₅	0.01 (-0.01, 0.02)	-0.27 (-0.34, -0.20)	0.28 (0.21, 0.35)
Human1 vs. Human2 (EagleEye, n = 53)	M	-0.05 (-0.13, 0.04)	-0.68 (-0.84, -0.53)	0.59 (0.43, 0.74)
	J ₀	0.01 (-0.02, 0.04)	-0.19 (-0.34, -0.04)	0.21 (0.06, 0.36)
	J ₄₅	-0.01 (-0.04, 0.01)	-0.17 (-0.32, -0.02)	0.14 (0.01, 0.29)

3.2 Computation of Refractive Vector Components

To compare the proposed deep learning network approach to other computational methods, the three vector components of refraction of dataset *EagleEye* were computed using the deep learning network and the visual metric VSOTF and judged in respect of agreement and precision with the available subjective data (Thibos et al. 2004). The deep learning network performed similar in terms of agreement and precision for all three vector components in comparison to the traditional metric (see Table 1). In terms of inter-examiner agreement between professional optometrists (see Table 1) that was calculated from double ratings of the same participants from two optometrists in 54 eyes from dataset 3, the artificial neural network covers a similar range of agreement and precision for all three power vectors of refraction.

4 DISCUSSION

In the current study, an artificial deep learning network approach was used in order to learn the complicated transfer from objective to subjective data and highly non-linear mapping from objective aberrometry data to the subjectively optimal sphero-

cylindrical correction of refractive errors was applied. Given the results, the detailed forward modeling of optical, neuronal and perceptual contributions to the transfer function using a deep learning network is comparable to earlier introduced methods (such as visual Strehl metrics) and the current gold standard, the subjective refraction.

4.1 Comparison with Conventional Computational Metrics

Compared to various pupil-plane and image quality metrics that were developed and applied for equal objectives, our results showed similar mean differences and comparable 95% limits of agreement, when compared to the VSOTF metric (Thibos et al. 2004). For the prediction of M from 200 eyes, different used metrics resulted in a mean difference between 0.24D to -0.04D, but were not the best ranked regarding their 95% limits of agreement. Best results in the precision of the prediction of M was shown from two pupil fraction metrics: PFWc and PFS_c; two image quality metrics for grating objects: AreaOTF, the visual Strehl of the optical transfer function (VSOTF) and one contrast metric: light-in-the-bucket (LIB) that all ranged between $\pm 0.49D$ to $\pm 0.59D$ (Thibos et al. 2004). Using pupil plane and image plane metrics, Guirao et al. (2003) reported a mean difference for M of 0.40D, when predicted by

the metrics and compared to subjective data of 146 eyes (95% limit of agreement: $\pm 0.98D$). With respect to the 95% limits of agreement, the results of the current investigation revealed similar or better values, when compared to the best ranked metrics regarding the prediction of M by Thibos et al. (2004) as well as Guirao et al. (2003). The average mean differences for the spherical equivalent refractive error M is much lower compared to the typical steps in subjective refraction, where errors are corrected in steps of 0.25D. Therefore, the performance for the prediction of the power vector M, J_0 , J_{45} of the developed artificial deep learning network can be classified as excellent, with the advantage over existing computational methods that computational time is significantly reduced.

4.2 Agreement with Subjective Refraction

The 95% limits of agreement for the computation of M were similar compared to studies that subjectively assessed the refractive to test the inter- as well as intra-observer agreement. Zadnik et al. (1992) reported 95% limit of agreement range from $\pm 0.94D$ for cycloplegic subjective refraction to $\pm 0.63D$ for the non-cycloplegic assessment of the spherical equivalent refractive error, when refraction was assessed by the same examiner multiple times (Zadnik et al. 1992). In case of retinoscopy, the 95% limits of agreement were reported to be $\pm 0.95D$ for cycloplegic retinoscopy and $\pm 0.78D$ for non-cycloplegic retinoscopy (Zadnik et al. 1992). Bullimore et al. (1993) found a 95% limit of agreement regarding the repeatability of the spherical equivalent refractive error, measured by two optometrists of $\pm 0.78D$ with a mean difference of -0.12D (Elliott and Bullimore, 1993). Values reported by Rosenfield and Chiu (1995) for the 95% limit of agreement of the inter-observer variability are $\pm 0.29D$, when the spherical equivalent refractive error was subjectively assessed by the same examiner for five times. When findings for the spherical equivalent error from the current study are compared to the previously reported subjective measures, it can be concluded that the presented method is as precise and accurate as the current gold standard, the subjective refraction.

4.3 Retinal and Neural Factors Affecting Perception

The question arises, whether a subjective refraction, either under monocular or binocular conditions,

would lead to the optical best correction of the aberrations, or if it represents a correction of aberrations that is most accepted by the wearer. In order to minimize any possible bias, we have only analyzed monocular measurements of refractive errors and additionally, the subjective refraction followed the rule "best visual acuity with maximum plus power". Compared to previous studies that were conducted in order to predict the best spherocylindrical correction of refractive errors using either pupil plane or image plane metrics, our approach, an artificial deep learning network that is able to incorporate perceptual and neural processes, assuming that they are similar over a cohort. Given enough capacity from the data, deep learning networks are able to learn any function, while alternative methods provide design transfer functions with possibly limited capacity. These approaches use hypotheses and limited domain knowledge about optics and the visual system in order to determine the best correction of an individual's aberration, with- or without the incorporation of for instance the contrast sensitivity of the eye. Since the deep learning network that was applied, is able to incorporate such processes, the network learned the mapping from Zernike coefficients to subjectively optimal refractive error corrections.

4.4 Monochromatic Aberrometry vs. Polychromatic Subjective Refraction

The prediction of refractive errors based on aberrometry data has some shortcomings that have to be taken into account. When measuring the refractive errors subjectively, a polychromatic test chart (white produced out of the red, green and blue LEDs of the monitor) is used, while wavefront aberrations are measured under monochromatic conditions. Following, one has to decide, which wavelength to use to compute the objective autorefraction as well as power vectors from aberrometry measurements and the results are only valid for the specific single wavelength. In the current analysis, the data were analyzed for a wavelength of $\lambda = 550nm$, based on the maximal spectral sensitivity of the eye and this is in contradiction to for example Thibos et al., where they have used a reference wavelength of $\lambda = 570nm$ (Thibos et al. 2004). Since the obtained results are comparable to the data of the subjective refractions, we conclude that the chosen wavelength only minimally effects the results obtained.

5 CONCLUSIONS

The aim of the current research was to use artificial deep learning networks for the prediction of subjective refractive corrections, in order to allow for an effective description of perceptual and neural processes that occur during the subjective assessment of such errors. The obtained results have shown that the presented methods lead to exact values of the power vectors of refraction, when compared to the subjective measurement and to a conventional metric. Additionally, aberrations need not necessarily be described by Zernike coefficients, neither is a detailed description more powerful to predict the refractive errors.

ACKNOWLEDGEMENTS

The authors would like to thank Steve Spratt from the Carl Zeiss Vision International GmbH to provide and support with the objective and subjective refraction database.

REFERENCES

- Applegate, R. A. *et al.* (2003) Interaction between aberrations to improve or reduce visual performance, *Journal of Cataract and Refractive Surgery*, 29(8), 1487–1495.
- Applegate, R. A. *et al.* (2003) Visual acuity as a function of Zernike mode and level of root mean square error, *Optometry and vision science*, 80(2), 97–105.
- Artal, P. *et al.* (2004) Neural compensation for the eye's optical aberrations, *Journal of Vision*, 4(4), 281–287.
- Bergstra, J., Yamins, D. L. K. and Cox, D. D. (2013) Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures, *ICML*, 115–123. doi: 1209.5111v1.
- Bishop, C. M. C. C. M. (2006) Pattern Recognition and Machine Learning, *Pattern Recognition*. doi: 10.1117/1.2819119.
- Bland, J. M. and Altman, D. G. (1986) Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet*, 1(8476), 307–310.
- Elliott, D. B. and Bullimore, M. A. (1993) Assessing the reliability, discriminative ability, and validity of disability glare tests, *Investigative Ophthalmology & Visual Science*, 34(1), 108–119.
- Francois, C. (2016) *Keras, GitHub repository*. Available at: <https://github.com/fchollet/keras>.
- Girija, S. S. (2016) Tensorflow: Large-scale machine learning on heterogeneous distributed systems.
- Goodfellow, Ian, Bengio, Yoshua, Courville, A. and Goodfellow, I. (2016) Deep Learning, *MIT Press*.
- Goss, D. A. and Grosvenor, T. (1996) Reliability of refraction--a literature review, *Journal of the American Optometric Association*, 67(10), 619–630.
- Guirao, A. and Williams, D. R. (2003) A method to predict refractive errors from wave aberration data, *Optometry and Vision Science*, 80(1), 36–42.
- He, K. *et al.* (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034. doi: 10.1109/ICCV.2015.123.
- Hinton, G. E. *et al.* (2012) Improving neural networks by preventing co-adaptation of feature detectors, *arXiv preprint arXiv:1207.0580*.
- Kingma, D. P. and Ba, J. L. (2015) Adam: a Method for Stochastic Optimization, *International Conference on Learning Representations 2015*, 1–15. doi: h10.1145/1830483.1830503.
- Leube, A., Ohlendorf, A. and Wahl, S. (2016) The influence of induced astigmatism on the depth of focus, *Optometry and Vision Science*, 93(10). doi: 10.1097/OPX.0000000000000961.
- National Eye Institute (1991) Early Treatment Diabetic Retinopathy Study design and baseline patient characteristics. ETDRS report number 7, *Ophthalmology*.
- Ohlendorf, A., Leube, A. and Wahl, S. (2016) Steps towards Smarter Solutions in Optometry and Ophthalmology-Inter-Device Agreement of Subjective Methods to Assess the Refractive Errors of the Eye, *Healthcare (Basel)*, 4(3). doi: 10.3390/healthcare4030041.
- Ohlendorf, A. and Schaeffel, F. (2009) Contrast adaptation induced by defocus - a possible error signal for emmetropization?, *Vision Research*, 49(2), 249–256. doi: 10.1016/j.visres.2008.10.016.
- Ohlendorf, A., Taberner, J. and Schaeffel, F. (2011) Neuronal adaptation to simulated and optically-induced astigmatic defocus, *Vision Research*, 51(6), 529–534. doi: 10.1016/j.visres.2011.01.010.
- Rosenfield, M. and Chiu, N. N. (1995) Repeatability of subjective and objective refraction', *Optometry and Vision Science*, 72(8), 577–579.
- Shahriari, B. *et al.* (2016) Taking the human out of the loop: A review of Bayesian optimization, *Proceedings of the IEEE*, 148–175. doi: 10.1109/JPROC.2015.2494218.
- Sloan, L. L. (1959) New test charts for the measurement of visual acuity at far and near distances, *American Journal of Ophthalmology*, 48(6), 807–813.
- Srivastava, N. *et al.* (2014) Dropout: A Simple Way to Prevent Neural Networks from Overfitting, *Journal of Machine Learning Research*, 15, 1929–1958. doi: 10.1214/12-AOS1000.
- Thibos, L. N. *et al.* (2004a) Accuracy and precision of objective refraction from wavefront aberrations, *Journal of Vision*, 4(4), pp. 329–351. doi: 10.1167/4.4.9.
- Thibos, L. N., Bradley, A. and Hong, X. (2002) A statistical model of the aberration structure of normal, well-

corrected eyes, *Ophthalmic Physiol Opt*, 22(5), 427–433.

Thibos, L. N., Wheeler, W. and Horner, D. (1997) Power vectors: an application of Fourier analysis to the description and statistical analysis of refractive error, *Optom Vis Sci*. 1997/06/01, 74(6), 367–375.

Wang, Y. *et al.* (2003) Changes of higher order aberration with various pupil sizes in the myopic eye, *Journal of refractive surgery*, 19(2), 270-274. doi: 10.3928/1081-597X-20030302-21.

Zadnik, K., Mutti, D. O. and Adams, A. J. (1992) The repeatability of measurement of the ocular components, *Invest Ophthalmol Vis Sci*, 33(7), 2325–2333.

