Deep Learning–Based Perceptual Stimulus Encoder for Bionic Vision

Lucas Relic
University of California, Santa Barbara
Santa Barbara, CA, USA
lucasrelic@ucsb.edu

Bowen Zhang
University of California, Santa Barbara
Santa Barbara, CA, USA
bowen68@ucsb.edu

Yi-Lin Tuan
University of California, Santa Barbara
Santa Barbara, CA, USA
ytuan@ucsb.edu

Michael Beyeler
University of California, Santa Barbara
Santa Barbara, CA, USA
mbeyeler@ucsb.edu

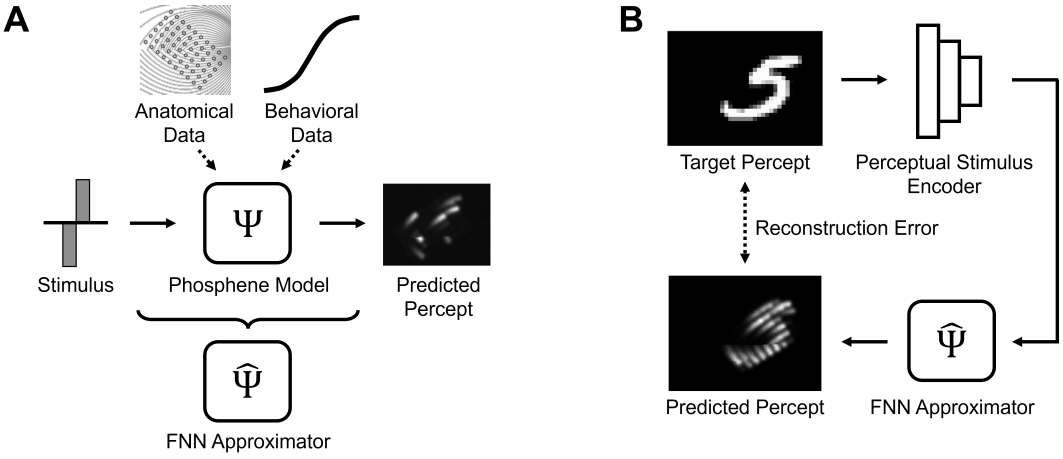

Figure 1: A) Simulated prosthetic vision (SPV). Constrained by neuroanatomical and/or psychophysical data, a phosphene model Ψ (e.g., [2]) predicts what a retinal implant user should "see" for any given input stimulus. The predicted percept is typically a nonlinear continuous function of the input stimulus, thus Ψ can be approximated by a generic feedforward neural network (FNN), $\hat{\Psi}$, which is amenable to differentiation. B) End-to-end optimization of bionic vision. For a given target percept, a stimulus encoder based on a convolutional neural network (CNN) is trained to predict the combination of active electrodes that generates the percept with the smallest possible reconstruction loss. The FNN approximator is fixed during encoder training.

ABSTRACT

Retinal implants have the potential to treat incurable blindness, yet the quality of the artificial vision they produce is still rudimentary. An outstanding challenge is identifying electrode activation patterns that lead to intelligible visual percepts (phosphenes). Here we propose a perceptual stimulus encoder (PSE) based on convolutional neural networks (CNNs) that is trained in an end-to-end fashion to predict the electrode activation patterns required to produce a desired visual percept. We demonstrate the effectiveness of the encoder on MNIST using a psychophysically validated phosphene model tailored to individual retinal implant users. The present work constitutes an essential first step towards improving the quality of the artificial vision provided by retinal implants.

AHs 2022, March 13–15, 2022, Kashiwa, Chiba, Japan

ACM ISBN 978-1-4503-9632-5/22/03.
https://doi.org/10.1145/3519391.3524034

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Modeling and simulation**.

KEYWORDS

deep learning, retinal implants, stimulus optimization

ACM Reference Format:
Lucas Relic, Bowen Zhang, Yi-Lin Tuan, and Michael Beyeler. 2022. Deep Learning–Based Perceptual Stimulus Encoder for Bionic Vision. In *Augmented Humans 2022 (AHs 2022), March 13–15, 2022, Kashiwa, Chiba, Japan.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3519391.3524034

1 INTRODUCTION

Hereditary retinal diseases such as retinitis pigmentosa are among the leading causes of incurable blindness in the world. Retinal implants (though elementary) provide an improved ability to localize high-contrast objects as well as perform basic orientation & mobility tasks [1]. These devices electrically stimulate surviving cells in the visual pathway to evoke visual percepts (phosphenes).

However, the quality of this artificial vision is still rudimentary, as the visual percepts elicited by current implants are often unrecognizable [4]. A major outstanding challenge is identifying electrode activation patterns that lead to perceptually intelligible phosphenes. One approach is to consider this an end-to-end optimization problem, where a deep neural network (encoder) is trained to predict the electrical stimulus needed to produce a desired percept (Fig. 1).

To this end, we make the following contributions:

- We propose a perceptual stimulus encoder (PSE) based on convolutional neural networks (CNNs) that is trained in an end-to-end fashion to predict the electrode activation patterns required to produce a desired visual percept. Importantly, the encoder is based on a psychophysically validated computational model of bionic vision [2].
- We demonstrate the effectiveness of the PSE on the MNIST dataset for three different users of the Argus II Retinal Prosthesis System (Second Sight Medical Products) [8].

2 RELATED WORK

2.1 Simulated Prosthetic Vision (SPV)

The goal of simulated prosthetic vision (SPV) is to predict what bionic eye users "see" in response to electrical stimulation. To date, most SPV studies rely on the scoreboard method, which assumes that each phosphene acts as a small independent light source, analogous to the images projected on the light bulb arrays of some sports stadium scoreboards [3]. However, evidence suggests that phosphenes often appear distorted (e.g., as simple geometric shapes such as lines, wedges, and blobs) and vary drastically across subjects and electrodes [4]. More recently, SPV models have therefore aimed to explain phosphene appearance as a function of neuroanatomical and psychophysical data [2, 5] (Fig. 1A). Open-source implementations for many of these models are provided by *pulse2percept*, a Python-based bionic vision simulator [9].

2.2 End-to-End Optimization of Bionic Vision

Although deep learning has previously been combined with SPV to perform image processing on the predicted percept (e.g., [6, 7, 11]), only few studies have considered the quest to improve the quality of artificial vision as an end-to-end optimization problem (Fig. 1B). Most notably, van Steveninck et al. [10] used a similar approach to ours by training an encoder-decoder deep neural network with a scoreboard model in the loop. However, their approach differs in three crucial ways: i) their phosphene model could not account for empirical data from current retinal implant users [2, 4], ii) their loss function did not consider the performance of either encoder or phosphene model, and most importantly iii) their decoder was in itself a deep neural network that could potentially learn to compensate for any deficiencies in the encoder or the phosphene model.

3 METHODS

Our model is illustrated in Fig. 1 and described in more detail below. We first approximated a psychophysically validated phosphene model (Ψ, [2]) with a generic feedforward neural network (FNN) ($\hat{\Psi}$, Fig. 1A). Once trained, the weights of the FNN Approximator were frozen and used to train a perceptual stimulus encoder (PSE) to minimize the reconstruction error between predicted and target percepts (Fig. 1B). The PSE took a target image as input and returned a combination of active electrodes as output, which were then fed into the FNN Approximator to predict a visual percept. The pixel-wise mean squared error between predicted and target percept served as the reconstruction error. The error was then backpropagated via the differentiable FNN Approximator to update the weights of the PSE. As a proof of concept, we considered the model's ability to predict handwritten digits from the popular MNIST dataset.

3.1 FNN Approximator

Our group [2] demonstrated through computational modeling that phosphene shape in epiretinal implants primarily depends not just on stimulus parameters but also on the retinal location of the stimulating electrode. This model (Ψ) can be fit to individual patients; however, it is not differentiable. We therefore approximated Ψ with a single-layer FNN ($\hat{\Psi}$) that took a 1×60 vector of current amplitudes as input (one amplitude for each electrode in the Argus II Retinal Prosthesis System) and returned a 121×161 image as output (i.e., the predicted percept). The FNN Approximator was trained on a synthetic dataset (50,000 samples, 80-20 train-test split) generated with Ψ: each sample was generated by first randomly selecting a number $N \in [1, 30]$ to stimulate, then randomly assigning a stimulation current between 1 and 10 microamps to each electrode. After training, the weights of the FNN Approximator were frozen.

3.2 Perceptual Stimulus Encoder (PSE)

The PSE was a convolutional neural network (CNN) consisting of two 3×3 convolutional layers (stride 1) followed by a max pooling layer after each convolutional layer, and a fully connected layer at the end.

Rather than optimizing the PSE with a pre-trained FNN Approximator in the loop, one might also consider training an inverse phosphene model ($\hat{\Psi^{-1}}$) to directly predict the required stimulus for a desired percept. We therefore trained an inverse model of equal depth to the PSE as a baseline model.

3.3 Simulated Bionic Eye Users

Since phosphene appearance varies drastically across patients, we followed Beyeler et al. [2] to tailor the phosphene model Ψ to the individual implant setup of three Argus II patients: 12-005, 51-009, and 52-001. The setup for these patients mainly differed in the implant location and in the values of model parameters ρ (describing phosphene size) and λ (describing phosphene elongation) [2]. Consequently, we had to train three different FNN Approximators and three different PSEs.

	Subject 12-005	Subject 51-009	Subject 52-001
Perceptual Stimulus Encoder (PSE)	**0.0317 ± 0.014**	**0.0547 ± 0.019**	**0.0311 ± 0.012**
Inverse Phosphene Model ($\hat{\Psi^{-1}}$)	0.0371 ± 0.016	0.0718 ± 0.025	0.0456 ± 0.017

Table 1: Reconstruction error (pixel-wise mean squared error) achieved on MNIST, reported as mean ± standard deviation across samples in the test set.

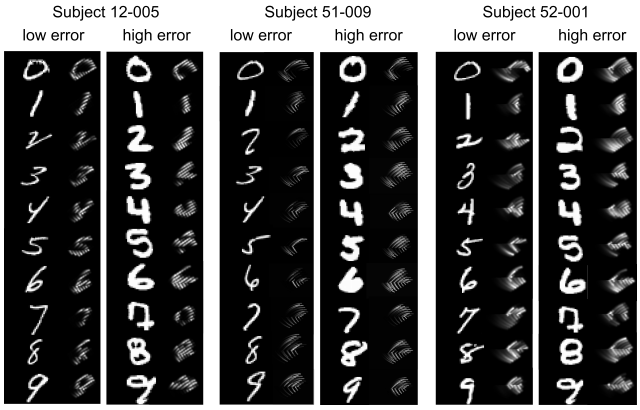

Figure 2: Representative example predictions achieved by the PSE trained on three different phosphene models that represent three different Argus II subjects. For each digit (left column in each panel), the corresponding predicted percept (right column) is shown.

4 RESULTS

Results are summarized in Table 1. The best reconstruction error was achieved for Subject 52-001 (0.0311), followed by 12-005 (0.0317) and 51-009 (0.0547). The relatively poor performance of the model for Subject 51-009 may be due to the fact that this subject sees very thin, elongated phosphenes that may not be easily combined to form an MNIST digit. A deeper encoding model may help to overcome this issue.

The PSE outperformed the inverse phosphene model in all three instances. While and end-to-end trained inverse phosphene model might seem advantageous at first, it has the notable drawback that the mapping from desired percept to required stimulus may be one-to-many, which may not lend itself well to gradient-based optimization.

Fig. 2 shows representative examples of digits with low (good) and high (bad) reconstruction errors for all three subjects. The PSE was able to utilize the inherent streakiness of the phosphene model to produce recognizable digits with a small number of active electrodes. However, it is worth noting that the digits with the lowest reconstruction error did not always look the best. The model tended to yield poor results for thick digits, which required the activation of a large number of electrodes and subsequently produced large indistinct blobs. Since all phosphenes have an inherent orientation, the model also struggled with digits whose edges ran orthogonal to that orientation.

5 CONCLUSION

The present work constitutes an essential first step towards improving the quality of artificial vision provided by current retinal implants. Future work should focus on more naturalistic datasets and developing a better perceptual error metric. The most important future contribution, however, will be to demonstrate that the proposed approach is a viable strategy to improve the quality of artificial vision in real bionic eye users.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (NIH R00 EY-029329 to MB).

REFERENCES

[1] Lauren N. Ayton, Nick Barnes, Gislin Dagnelie, Takashi Fujikado, Georges Goetz, Ralf Hornig, Bryan W. Jones, Mahiul M. K. Muqit, Daniel L. Rathbun, Katarina Stingl, James D. Weiland, and Matthew A. Petoe. 2020. An update on retinal prostheses. *Clinical Neurophysiology* 131, 6 (June 2020), 1383–1398. https://doi.org/10.1016/j.clinph.2019.11.029

[2] Michael Beyeler, Devyani Nanduri, James D. Weiland, Ariel Rokem, Geoffrey M. Boynton, and Ione Fine. 2019. A model of ganglion axon pathways accounts for percepts elicited by retinal implants. *Scientific Reports* 9, 1 (2019). https://doi.org/10.1038/s41598-019-45416-4

[3] Wm H. Dobelle. 2000. Artificial Vision for the Blind by Connecting a Television Camera to the Visual Cortex. *ASAIO Journal* 46, 1 (Feb. 2000), 3–9.

[4] Cordelia Erickson-Davis and Helma Korzybska. 2021. What do blind people "see" with retinal prostheses? Observations and qualitative reports of epiretinal implant users. *PLOS ONE* 16, 2 (Feb. 2021), e0229189. https://doi.org/10.1371/journal.pone.0229189 Publisher: Public Library of Science.

[5] Jacob Granley and Michael Beyeler. 2021. A Computational Model of Phosphene Appearance for Epiretinal Prostheses. 4477–4481. https://doi.org/10.1109/EMBC46164.2021.9629663 ISSN: 2694-0604.

[6] Nicole Han, Sudhanshu Srivastava, Aiwen Xu, Devi Klein, and Michael Beyeler. 2021. Deep Learning–Based Scene Simplification for Bionic Vision. In *Augmented Humans Conference 2021 (AHs'21)*. Association for Computing Machinery, New York, NY, USA, 45–54. https://doi.org/10.1145/3458709.3458982

[7] Lachlan Horne, Jose Alvarez, Chris McCarthy, Mathieu Salzmann, and Nick Barnes. 2016. Semantic labeling for prosthetic vision. *Computer Vision and Image Understanding* 149 (2016), 113–125.

[8] Y. H. Luo and L. da Cruz. 2016. The Argus((R)) II Retinal Prosthesis System. *Prog Retin Eye Res* 50 (Jan. 2016), 89–107. https://doi.org/10.1016/j.preteyeres.2015.09.003

[9] Michael Beyeler, Geoffrey M. Boynton, Ione Fine, and Ariel Rokem. 2017. pulse2percept: A Python-based simulation framework for bionic vision. In *Proceedings of the 16th Python in Science Conference*, Katy Huff, David Lippa, Dillon Niederhut, and M Pacer (Eds.). 81 – 88. https://doi.org/10.25080/shinma-7f4c6e7-00c

[10] Jaap de Ruyter van Steveninck, Umut Guclu, Richard JA van Wezel, and Marcel AJ van Gerven. 2020. End-to-end optimization of prosthetic vision. *bioRxiv* (2020).

[11] Ying Zhao, Qi Li, Donghui Wang, and Aiping Yu. 2018. Image Processing Strategies Based on Deep Neural Network for Simulated Prosthetic Vision. In *2018 11th International Symposium on Computational Intelligence and Design (ISCID)*, Vol. 01. 200–203. https://doi.org/10.1109/ISCID.2018.00052