Examining the potential functional significance of initially poor temporal acuity

Marin Vogelsang (ozaki@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139, USA

Lukas Vogelsang (lvogelsa@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139, USA

Pawan Sinha (psinha@mit.edu)
Department of Brain and Cognitive Sciences, MIT
Cambridge, MA 02139, USA

Abstract

The human visual system is remarkably immature at birth, exhibiting initially degraded spatial and temporal vision. While early spatial degradations have been proposed to provide important benefits to the developing visual system, less is known about the potential adaptive significance of early temporal immaturities. Here, we investigated this possibility computationally, using 3D convolutional neural networks trained on a temporally meaningful classification task. We systematically manipulated spatial and temporal blur when training on the Something-Something V2 dataset, which critically depends on temporal order. Analysis of learned receptive fields revealed that initial exposure to temporal blur led to longer-range temporal processing, persisting even after transitioning to clear temporal inputs. Such developmental trajectory commencing with initial temporal blur also significantly enhanced generalization performance compared to training with high temporal resolution input or corresponding spatial blur alone. These findings extend the concept of adaptive developmental degradations into the temporal domain, suggesting that immaturities in temporal vision may instantiate important mechanisms for robust perception later in life.

Keywords: Deep networks; spatiotemporal processing; temporal integration; visual development; early degradations

Introduction

The human visual system is remarkably immature at birth. For instance, newborns initially exhibit significantly reduced contrast sensitivity (Kiorpes, 2016), visual acuity (Dobkins, Lia, & Teller, 1997), and color sensitivity (Lenassi, Likar, Stirn-Kranjc, & Brecelj, 2008), experiencing gradual improvements throughout the first months or years. Building on past proposals (Turkewitz & Kenny, 1982; Newport, 1988; Elman, 1993), recent work suggests that these early degradations might have functional significance. Early exposure to blurred visual inputs has been proposed to yield spatially extended receptive fields (L. Vogelsang et al., 2018), better generalization in face recognition (Jang & Tong, 2021), and enhanced visual category learning (Jinsi, Henderson, & Tarr, 2023). Similarly, initially degraded color vision may induce greater resilience to color variations later in life (M. Vogelsang et al., 2024).

However, research on the role of developmental degradations has largely focused on spatial vision, neglecting temporal aspects of processing. Although temporal vision typically matures faster than its spatial counterpart, the state of initial immaturity is pronounced (Banks, 1982; Ellemberg, Lewis, Liu, & Maurer, 1999). This raises the question of whether reduced temporal acuity in the first months of life might confer adaptive benefits similar to the spatial domain. Here, we computationally explore this possibility using deep convolutional neural networks (3D-CNNs) trained on temporally meaningful classification tasks.

Methodology

We adapted the AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), modifying its first convolutional layer from 96 receptive fields (RFs) of size 11×11 to 48 RFs of size 22×22, to allow for flexibility in learned receptive field structures. We then temporally expanded the layer to include 11 time frames, yielding RFs of size $22 \times 22 \times 11$.

To select an appropriate training dataset, we evaluated both Kinetics-600 (Carreira, Noland, Banki-Horvath, Hillier, & Zisserman, 2018) and Something-Something V2 (SSv2) (Goyal et al., 2017). As shown in Table 1, temporal order was critical for SSv2 but not for Kinetics-600 (training on shuffled videos led to a drop of 16.6% in performance for the former but only 3.4% for the latter). Similarly, inspection of learned RFs when trained on SSv2 revealed markedly slower spatiotemporal movements indicative of more temporally-extended processing, relative to training on Kinetics-600. We thus selected SSv2 as our main dataset for subsequent simulations.

Training	Normal testing	Shuffled testing
Kinetics (normal)	**0.518**	0.162
Kinetics (shuffled)	0.324	**0.484**
SSv2 (normal)	**0.284**	0.017
SSv2 (shuffled)	0.092	**0.118**

Table 1: Classification accuracy for training/testing on shuffled vs. non-shuffled videos, for Kinetics-600 and Something-Something V2 (SSv2).

Videos were temporally cropped to 48 central frames, spatially resized to 128×128 pixels, scaled to a range of $[-1, 1]$, and randomly flipped horizontally. For training and testing on blur, Gaussian spatial blur ($\sigma = 4$) and Gaussian temporal blur ($\sigma = 2$) were applied. Training was carried out using SGD with a Nesterov momentum of 0.9, a learning rate of 0.001, and a batch size of 32. Models were trained while systematically varying the degree of spatial and temporal blur. We define training with blurred inputs as *low-frequency (L)*, since the applied blur leaves predominantly lower frequencies, and others maintaining low and high frequencies as *high-frequency (H)*. We thus refer to conditions as *spatial low (S_L)* or *spatial high (S_H)*, and similarly as *temporal low (T_L)* or *temporal high (T_H)*, adopting the following nomenclature:

- No temporal blur: spatial high, temporal high ($S_H T_H$) and spatial low, temporal high ($S_L T_H$).
- Temporal blur: spatial high, temporal low ($S_H T_L$) and spatial low, temporal low ($S_L T_L$).
- Shuffled temporal order, removing meaningful temporal contiguity: $S_H T_H$-shuffled as well as $S_L T_H$-shuffled.
- Staged training: $S_L T_L$-to-$S_H T_H$ and $S_H T_L$-to-$S_H T_H$, where the training input type changes after the first half of epochs.

Following training, we analyzed the temporal frequency content of individual RFs by applying an FFT (without the constant component; averaged across the two spatial dimensions) and computing a normalized amplitude-weighted average of frequency.

Results

First-layer representations

As depicted in Figure 1, temporal blur lowered the temporal frequency of learned spatiotemporal receptive fields (RFs) in the first convolutional layer. This led to RFs exhibiting slow and smooth motion across frames (depicted in Figure 2). Notably, this effect occurred irrespective of spatial blur (S_LT_L vs. S_HT_L; Figures 1 and 2). In contrast, training networks with temporally shuffled inputs resulted in receptive fields with higher temporal frequencies. Interestingly, staged training – transitioning from initially blurred temporal inputs to clear temporal inputs later – maintained RF characteristics shaped during the initial training phase. This robustness was evident regardless of whether the initial training phase involved high spatial acuity (S_HT_L-to-S_HT_H) or low spatial acuity (S_LT_L-to-S_HT_H), demonstrating persistent effects of early temporal experience across full developmental training.

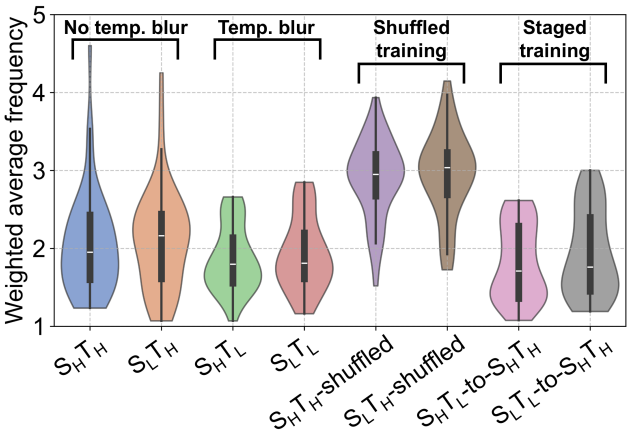

Figure 1: Distribution of temporal frequency metric scores of individual RFs as a function of training condition.

Generalization performance

Table 2 summarizes model classification accuracy across training conditions, when tested on spatial blur, temporal blur, both, or neither. Training with spatial blur alone (S_LT_H) resulted in only minor generalization improvements relative to standard (no spatial blur) training (S_HT_H). In contrast, training with temporal blur alone (S_HT_L) substantially improved performance, approaching the benefits observed when both spatial and temporal blur were combined (S_LT_L). However, optimal performance emerged from the two developmentally inspired trajectories that began training with temporal blur, subsequently transitioning to spatially and temporally clear inputs.

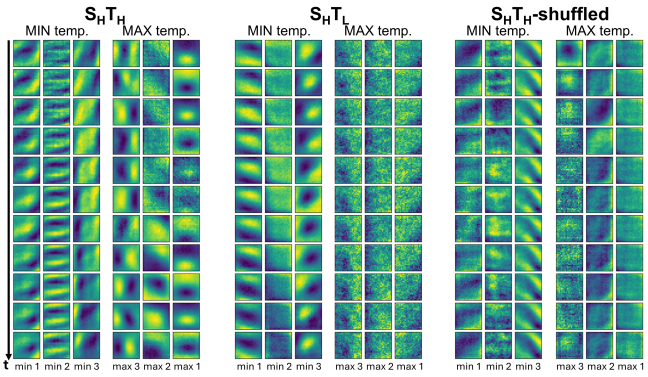

Figure 2: RFs with the 3 lowest vs. 3 highest temporal frequency metric scores, for training without blur, with temporal blur, and with shuffled temporal frames.

	Test S_HT_H	S_HT_L	S_LT_H	S_LT_L
S_HT_H	0.284	0.115	0.244	0.087
S_LT_H	0.254	0.119	0.269	0.121
S_HT_L	0.174	0.251	0.145	0.216
S_LT_L	0.176	0.232	0.170	0.244
Staged 1	0.265	0.206	0.225	0.164
Staged 2	0.258	0.191	0.243	0.178

Table 2: Classification accuracy across several training and testing conditions. *Staged 1/2* refer to S_HT_L-to-S_HT_H and S_LT_L-to-S_HT_H.

Discussion

Our findings begin to extend previous studies on the functional significance of developmental trajectories that progress from impoverished to enriched visual inputs (L. Vogelsang et al., 2018; Jang & Tong, 2021; Yoshihara, Fukiage, & Nishida, 2023) into the temporal domain. Specifically, initial experience with temporally blurred visual input yields temporally-extended receptive fields, subserving potentially important perceptual functions later on. Future research should investigate the implications for richer and more varied visual contexts as well as how findings depend on architectures. Collectively, our results highlight the importance of the temporal dimension in computational vision and illustrate the productive synergy between studies of computational modeling and human development.

Acknowledgements

This work has been supported by NIH grant R01EY020517 to Pawan Sinha. Lukas Vogelsang is supported by a grant from the Simons Foundation International to the Simons Center for the Social Brain at MIT. Marin Vogelsang is supported by the Japan Society for the Promotion of Science (JSPS), Overseas Research Fellowship and the Yamada Science Foundation.

References

Banks, M. S. (1982). The development of spatial and temporal contrast sensitivity. *Current Eye Research*, *2*(3), 191–198.

Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., & Zisserman, A. (2018). A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Dobkins, K. R., Lia, B., & Teller, D. Y. (1997). Infant color vision: Temporal contrast sensitivity functions for chromatic (red/green) stimuli in 3-month-olds. *Vision Research*, *37*(19), 2699–2716.

Ellemberg, D., Lewis, T. L., Liu, C. H., & Maurer, D. (1999). Development of spatial and temporal vision during childhood. *Vision research*, *39*(14), 2325–2333.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., ... others (2017). The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the ieee international conference on computer vision* (pp. 5842–5850).

Jang, H., & Tong, F. (2021). Convolutional neural networks trained with a developmental sequence of blurry to clear images reveal core differences between face and object processing. *Journal of vision*, *21*(12), 6–6.

Jinsi, O., Henderson, M. M., & Tarr, M. J. (2023). Early experience with low-pass filtered images facilitates visual category learning in a neural network model. *Plos one*, *18*(1), e0280145.

Kiorpes, L. (2016). The puzzle of visual development: behavior and neural limits. *Journal of Neuroscience*, *36*(45), 11384–11393.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Lenassi, E., Likar, K., Stirn-Kranjc, B., & Brecelj, J. (2008). Vep maturation and visual acuity in infants and preschool children. *Documenta Ophthalmologica*, *117*, 111–120.

Newport, E. L. (1988). Constraints on learning and their role in language acquisition: Studies of the acquisition of american sign language. *Language sciences*, *10*(1), 147–172.

Turkewitz, G., & Kenny, P. A. (1982). Limitations on input as a basis for neural organization and perceptual development: A preliminary theoretical statement. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, *15*(4), 357–368.

Vogelsang, L., Gilad-Gutnick, S., Ehrenberg, E., Yonas, A., Diamond, S., Held, R., & Sinha, P. (2018). Potential downside of high initial visual acuity. *Proceedings of the National Academy of Sciences*, *115*(44), 11333–11338.

Vogelsang, M., Vogelsang, L., Gupta, P., Gandhi, T. K., Shah, P., Swami, P., ... Sinha, P. (2024). Impact of early visual experience on later usage of color cues. *Science*, *384*(6698), 907–912.

Yoshihara, S., Fukiage, T., & Nishida, S. (2023). Does training with blurred images bring convolutional neural networks closer to humans with respect to robust object recognition and internal representations? *Frontiers in Psychology*, *14*, 1047694.