Accelerated Texforms: Alternative Methods for Generating Unrecognizable Object Images with Preserved Mid-Level Features

Arturo Deza (arturo_deza@fas.harvard.edu)
Department of Psychology, Harvard University
Cambridge, MA. USA

Yi-Chia Chen (yi-chia.chen@fas.harvard.edu)
Department of Psychology, Harvard University
Cambridge, MA. USA

Bria Long (bria@stanford.edu)
Department of Psychology, Stanford University
Stanford, CA. USA

Talia Konkle (talia_konkle@fas.harvard.edu)
Department of Psychology, Harvard University
Cambridge, MA. USA

Abstract

Texforms are images that preserve the coarse shape and texture information of objects, while rendering them unrecognizable at the basic level (Long, Konkle, Cohen, & Alvarez, 2016). These stimuli have been valuable to test whether cognitive and neural processes depend on explicit recognition of the objects. However, to generate these images, the current implementation and computational complexity of the model requires approximately 4-6 hours per object – thus preventing data-hungry experiments that may require generating thousands of texforms. Our contribution in this work includes the introduction of 2 new texform generation methods that accelerate the rendering time from hours to minutes or seconds respectively. The first we call Fast-FS-Texform where we accelerate the rendering time of the Freeman and Simoncelli (2011) model and increase the output resolution by placing a simulated point of fixation outside of the visual field. The second, which we call NeuroFovea-Texform, is an adaptation of the newly proposed metamer model of Deza, Jonnalagadda, and Eckstein (2019) which leverages a VGGNet and foveated style transfer. We show qualitative and quantitative results of both new methods opening the door to data-intensive texform experiments.

Keywords: texforms, object representation, visual perception

Introduction

The human visual system transforms retinal input into high-level representations of the physical world. To do so, early processing stages detect low-level feature information such as edges, extending into mid-level features that are more object-centered such as textures and form information, and ultimately give rise to high-level object categorization (DiCarlo & Cox, 2007; Mishkin, Ungerleider, & Macko, 1983).

A major challenge in visual cognitive neuroscience is to understand the interface between mid-level representational stages–which are more perceptual, and high-level recognition processes–which are more semantic in nature. However, dissociating these levels of representation has been challenging when using pictures of recognizable real-world objects as stimuli. Recently, a new stimulus class called 'texforms' was developed in order to separate high-level recognition processes from mid-level shape and texture processes (Long et al., 2016; Long, Störmer, & Alvarez, 2017; Long, Yu, & Konkle, 2018).

Specifically, texforms are images that preserve the coarse shape and texture information, while rendering them unrecognizable at the basic-level. Using texform stimuli, Long et al. (2016) showed that there are mid-level perceptual differences between big and small inanimate objects – a distinction which had often been thought as being purely 'semantic' because of the variety in each class of objects (see also Long et al., 2017). Further, these mid-level feature differences preserved in texforms are sufficient to drive the large-scale organization of neural responses by animacy and object size in visual cortex, highlighting an extensive role for mid-level feature computations in ventral stream organization and broadly in the visual system (Long et al., 2018). Most recently, this stimulus class is proving to be a useful image manipulation to probe the necessity of semantic processing across a wide variety of visual tasks (*e.g.* long term memory: Lam, Schurgin, & Brady, 2019, curvature processing: Magri, Long, Chiou, & Konkle, 2019, visual curiosity and exploration: Gottlieb & Oudeyer, 2018).

However, there are two main limitations to using texforms in future experiments: rendering time and image resolution. To create these stimuli, Long et al. (2016) used the scene metamer model of Freeman and Simoncelli (2011) which coerces noise to have the same intermediate-level image texture statistics as the input image, given a point of fixation. The rendering process is iterative in nature and takes about 4-6 hours

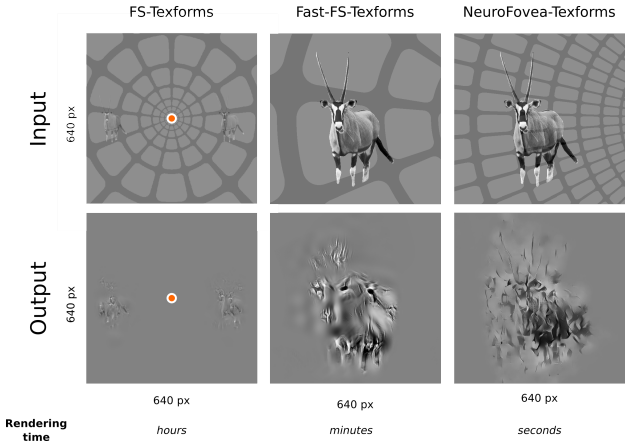

Figure 1: The collection of current and proposed texform rendering models. Left: The original FS-Texform rendering pipeline. Center: Fast rendering variant with the fixation point is outside the frame of the image. Right: Alternative texform method using foveated style transfer for visual scene metamers.

per image to synthesize one texform. Further, to make a texform of an object with this model, the target object is placed on a gray background in the visual periphery of the model, leveraging image summary statistics in relatively coarse pooling regions. As a consequence of placing the object away from center fixation, the resolution of the resulting synthesized texform is relatively low (200×200 px; see Figure 1).

In this paper, we introduce and compare two alternative methods for generating texform images. The first method uses the same algorithm as Long et al. (2016), but more efficiently leverages peripheral pooling windows to make both faster and higher resolution images, by simulating a point of fixation *outside* of the rendering window. The second uses a different algorithm developed by Deza et al. (2019) for visual scene metamerism that capitalizes the representations of a VGG19-Net (Simonyan & Zisserman, 2014) and feed-forward style transfer (Huang & Belongie, 2017; Gatys, Ecker, & Bethge, 2016) to rapidly generate texform-like stimuli. These two approaches also enable different ways of parameterizing the texform rendering process. Our overall goal is to make these faster methods publicly available to speed research that investigates mid-level feature processing of objects.

Methods

Algorithm 1: Fast-FS-TexForm. The original texform model by Long et al., 2016 places an object in the visual periphery and renders a texform through the metamer model of Freeman and Simoncelli (2011). During this procedure, an image is synthesized from noise in order to match the texture statistics of the object in the periphery for every overlapping receptive field in addition to roughly matching the structure given a low-pass residual of the input image. Given the parameters selected by Long et al. (2016), there are roughly $1-4$ receptive fields that partially overlap with the input object of size 200×200 px in a 640×640 px window frame (see Figure 1). Henceforth, we will refer to this original method for creating texforms as FS-texform method, named after the Freeman and Simoncelli synthesis method.

The accelerated texform model operates with the same rendering pipeline of Freeman and Simoncelli (2011) and parameter settings of Long et al. (2016), but critically places a simulated point of fixation *outside* of the image, at an equivalent eccentricity of the previous settings. This has two major consequences. First, this method enables the construction of a higher-resolution texform, preserving the resolution of the input image at 640×640. Additionally, the computational complexity of the model is reduced in an order of about $\times 25$ to $\times 100$, given that original algorithm contained many pooling regions that had to be synthesized slowly even though they only covered the uniform gray background and were ultimately cropped.

Algorithm 2: NeuroFovea-Texform:

Recently Deza et al. (2019) developed a fast way to generate visual scene metamers by capitalizing on Peripheral Representations (Deza & Eckstein, 2016) and localized Style Transfer (Gatys, Ecker, Bethge, Hertzmann, & Shechtman, 2017) creating the notion of *Foveated Style Transfer*. The idea behind their model is to gently perturb the original image in the direction of its texture representation, for each receptive field in the field of view of a simulated human observer. Each receptive field of the input image is encoded with a spatially masked *relu4_1* (the rectified 4th convolutional block) activation of a VGG19, and they are each 'texturized' by applying Adaptive Instance Normalization (AdaIN) (Huang & Belongie, 2017) on noise and transferring the style of the receptive field content onto the noise. They then find the maximum perturbation coefficients (α) in an image given the size of each receptive field. The main difference with the model of Freeman and Simoncelli (2011) is that rather than coerce noise to match the same local texture statistics of each receptive field, Deza et al. (2019) use noise to perturb the image into the direction of its intrinsic texture representation for each receptive field, allowing a purely feed-forward and stochastic metamer generating pipeline.

Given the hyperparametric nature of the Deza et al. (2019) model, we decided to simplify it's computation and restrict the perturbation coefficients ($\gamma_s(\circ) = \alpha_0$) to a single value. We made this simplification given that the image stimuli is not a scene (and rather a small object in the gray background), and we are computing distortions over a small amount of pooling regions. We thus performed a perceptual optimization procedure (Deza et al., 2019) over 240 images with MS-SSIM (Wang, Simoncelli, & Bovik, 2003) and found that the optimal coefficient that matched the distortions of the texforms was $\alpha_0 = 0.80$, for a scaling factor of $s = 0.25$ and pooling window aspect ratio of 2.0. We did not use a refinement module in this paper, however our final outputs are normalized via contrast adjustment with respect to the input image.

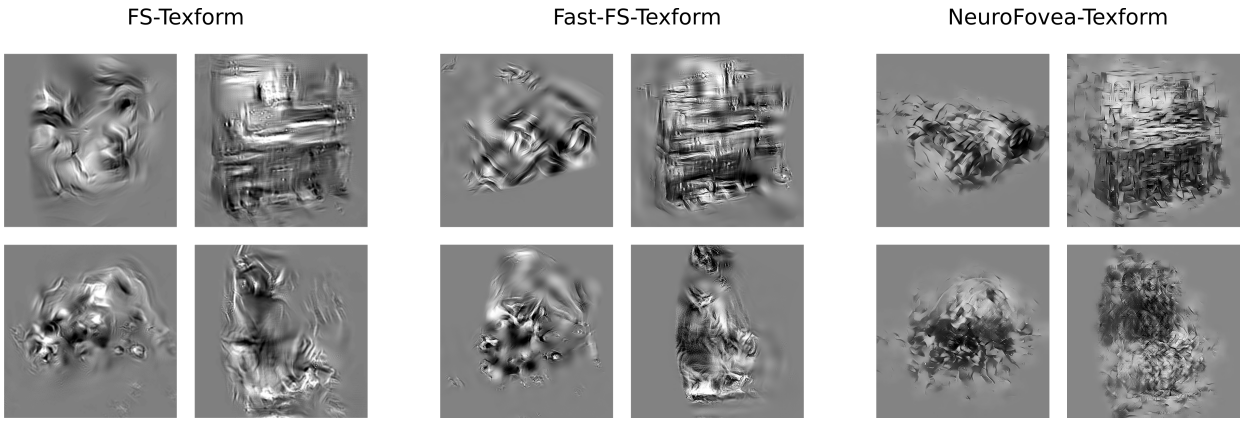

Figure 2: A collection of texform samples for the FS-Texform (Left) originally rendered at 200×200 px, and the two new proposed models (Middle: Fast-FS-Texform, Right: NeuroFovea-texform) rendered at high resolution 640×640 px. The true labels are as follows: Top Left: Binoculars, Top Right: Upright Piano, Bottom Left: Lady Bug, Bottom Right: Gorilla.

Results

General Model Comparison: Figure 1 shows a schematic comparing the original model with these two alternatives. In the figure all models render a texform stimuli from a hoofed gemsbok (an animal). The Fast-FS-Texform model noticeably renders the image an order of magnitude faster than the original Texform model, and in addition can render a higher-resolution object stimuli at 640×640 px that can later be down-sampled if necessary. In addition, the distortion type is identical of the Long et al. (2016) texforms as the same texture statistics of Portilla and Simoncelli (2000) are used to compute local texture statistics, although a structural constraint is added in the Freeman and Simoncelli (2011) to roughly preserve the object shape (Wallis, Bethge, & Wichmann, 2016).

The model of Deza et al. (2019) creates a different *flavor* of texforms, given that the model perturbs the object in the direction of its texture-like representation vs performing texture-matching. Further, this method a achieves an even faster rendering speed.

Figure 2 shows texform exemplars across object size and animacy dimensions, rendered across all models (FS-Texforms, Fast-FS-Texforms and NeuroFovea-Texforms), where only the later two are able to receive input from high-resolution images. Notice that all models roughly preserve global shape, while distorting the image locally via their respective texture model.

Quantitative Assessment: It is clear that the two methods preserve texture and coarse form in different ways. Is the NeuroFovea method creating more distorted texforms than the Fast-FS-Texform algorithm? To examine this, we computed the image-level similarity metric as a proxy for a human observer between the texform and original recognizable image for a collection of $N = 180$ input images $I^{(j)}$ using MS-SSIM scores between the original images and their outputs. MS-SSIM computes a multi-resolution perceptual similarity score that takes into account factors such as contrast, luminance and structure – giving an intuition of how a model observer would perceive equal distortions that are different in flavor for both the Fast-FS-Texforms T_F and NeuroFovea-Texforms T_N. Thus we compute:

$$\mathbb{E}(\Delta\text{-MS-SSIM})^2 = \frac{1}{N}\sum_{j=1}^{N}(\text{MS-SSIM}(T_F^{(j)}, I^{(j)}) - \text{MS-SSIM}(T_N^{(j)}(\gamma_s), I^{(j)}))^2 \quad (1)$$

We find that $\mathbb{E}(\Delta\text{-MS-SSIM})^2$ is 0.03 ± 0.03, which indicates similar perceptual losses across both families of texforms with respect to their references despite their difference in texture distortions. Thus, this analysis suggests that these two methods of generating texforms equally preserve basic-low level features, and call for future work that examines both how exactly they differ in mid-level features (*e.g.*, specific kinds of texture/shape properties)

Discussion

Here we developed two methods for generating texforms that address the key limitations of the original method. Specifically, we introduced a minor modification to the original method to generate higher resolution texforms at a significant fraction of a speed: 15 *minutes per texform*. Additionally, we adapted the newly proposed metamer model of Deza et al. (2019) for a deep-net like texform that is rendered at 1 *second per texform* independent of stimuli size.

It is worth noting that this is only the first stage of generating texforms; critically, the next step following *any* of these rendering methods is to perform a behavioural recognition task to exclude any model outputs that remain recognizable as performed in Long et al. (2016, 2017). In our experience, there are some target stimuli that even under heavy distortions are still recognizable at the basic-level such as zebras and giraffes, potentially due to their unique textures.

To our knowledge this is the first effort that leverages other metamer models to accelerate texform rendering. Indeed, recent metamer models such as the CNN-Synthesis model (Wallis et al., 2019), and the SideEye model (Fridman et al., 2017) may also potentially be able to render a different flavor of texforms. In parallel, we have also found that the work of Roberts, Kingstone, and Todd (2019) has also tried similar strategies with the Freeman and Simoncelli (2011) model to accelerate texform rendering as they too have highlighted the problem of the computational intractability given the current state of the art.

Finally, future work with these texform rendering algorithms will explore the ways in which texform generation can be parameterized and varied. For example, in the Fast-FS-Texform algorithm, we can vary the degree of visual eccentricity, pooling region aspect ratio, point of fixation, and receptive field rate of growth (scale). In the NeuroFovea method, we can also vary the previous parameters in addition to the deepnet layer over which we perform style transfer, potentially allowing for texforms that contain more or less recognizable shape parts and texture variations. By creating different variants of texforms, we can start to probe the degree to which different kinds of mid-level visual features are necessary for semantic processing. More broadly, these new methods that quickly generate different variants of texforms will permit us to test more specific hypothesis about how meaning is ultimately derived from the statistics of the visual input.

References

Deza, A., & Eckstein, M. (2016). Can peripheral representations improve clutter metrics on complex scenes? In *Advances in neural information processing systems (nips)*.

Deza, A., Jonnalagadda, A., & Eckstein, M. P. (2019). Towards metamerism via foveated style transfer. In *International conference on learning representations*.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, *11*(8), 333–341.

Freeman, J., & Simoncelli, E. P. (2011). Metamers of the ventral stream. *Nature neuroscience*, *14*(9), 1195.

Fridman, L., Jenik, B., Keshvari, S., Reimer, B., Zetzsche, C., & Rosenholtz, R. (2017). Sideeye: A generative neural network based simulator of human peripheral vision. *arXiv preprint arXiv:1706.04568*.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2414–2423).

Gatys, L. A., Ecker, A. S., Bethge, M., Hertzmann, A., & Shechtman, E. (2017). Controlling perceptual factors in neural style transfer. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 3985–3993).

Gottlieb, J., & Oudeyer, P.-Y. (2018). Towards a neuroscience of active sampling and curiosity. *Nature Reviews Neuroscience*, 1.

Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the ieee international conference on computer vision* (pp. 1501–1510).

Lam, K., Schurgin, M. W., & Brady, T. F. (2019). The contributions of visual details vs semantic information to visual long-term memory. In *Vision sciences society (vss) annual meeting.*

Long, B., Konkle, T., Cohen, M. A., & Alvarez, G. A. (2016). Mid-level perceptual features distinguish objects of different real-world sizes. *Journal of Experimental Psychology: General*, *145*(1), 95.

Long, B., Störmer, V. S., & Alvarez, G. A. (2017). Mid-level perceptual features contain early cues to animacy. *Journal of vision*, *17*(6), 20–20.

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015–E9024.

Magri, C., Long, B., Chiou, R., & Konkle, T. (2019). Behavioral and neural associations between object size and curvature. In *Vision sciences society (vss) annual meeting.*

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in neurosciences*, *6*, 414–417.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International journal of computer vision*, *40*(1), 49–70.

Roberts, K. H., Kingstone, A., & Todd, R. M. (2019). Generating visual stimuli that vary in recognisability. In *Vision sciences society (vss) annual meeting.*

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Wallis, T. S., Bethge, M., & Wichmann, F. A. (2016). Testing models of peripheral encoding using metamerism in an oddity paradigm. *Journal of vision*, *16*(2), 4–4.

Wallis, T. S., Funke, C. M., Ecker, A. S., Gatys, L. A., Wichmann, F. A., & Bethge, M. (2019). Image content is more important than boumas law for scene metamers. *eLife*, *8*, e42512.

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale structural similarity for image quality assessment. In *The thrity-seventh asilomar conference on signals, systems & computers, 2003* (Vol. 2, pp. 1398–1402).