Unsupervised learning of manifold models for neural coding of physical transformations in the ventral visual pathway

Marissa C. Connor (marissa.connor@gatech.edu)
School of Electrical and Computer Engineering, Georgia Institute of Technology, 777 Atlantic Dr. NW
Atlanta, GA 30332-0250, USA

Christopher J. Rozell (crozell@gatech.edu)
School of Electrical and Computer Engineering, Georgia Institute of Technology, 777 Atlantic Dr. NW
Atlanta, GA 30332-0250, USA

Abstract

Biological vision is envied for its ability to learn to recognize objects in the 3D world undergoing physical transformations. A recent hypothesis is that the ventral visual pathway exploits the manifold nature of these transformations to form neural codes that are efficient for discrimination, but there is no compelling model for how this representation is learned from data. We propose a computational model that performs unsupervised learning on received retinal imagery to infer identity-preserving transformations. We show that such a model can successfully learn useful representations on a subset of objects that can be transferred to new objects. We also demonstrate that this model can be used to infer 3D transformations from 2D imagery despite the ill-conditioned nature of the problem. This model for 3D inference can account for psychophysical experiments such as 3D shape perception from random-dot kinematograms.

Keywords: manifolds;transfer learning;mental rotation

Introduction

The central task of the human visual system is to use the received retinal images to make inferences about the environmental causes of those stimuli. This is a challenging problem because the retinal images change dramatically when objects in the world undergo simple physical transformations such as rotation and scaling. Recent theories of neural coding in the ventral visual pathway propose that the system takes advantage of the fact that identity-preserving physical transformations (e.g., rotation) induce a manifold in any given neural code. Specifically, the hypothesis is that the stages of the ventral pathway serve to flatten the manifold of the representation at each stage, resulting in high-level areas (e.g., Inferior Temporal (IT) Cortex) that have responses that are more robust for object recognition (DiCarlo & Cox, 2007). Understanding the representations and algorithms underlying invariant visual recognition in biology would be of immense value in computer vision systems.

The role of manifold models in neural coding has preliminary support from electrophysiology data (DiCarlo, Zoccolan, & Rust, 2012). However, we lack a computational model of how a neural system could learn to represent and exploit the manifold nature of transformations in the 3D world from the received retinal (2D) images of transforming objects. The contribution of our work is to propose a computational model that uses unsupervised learning to create an analytic representation of these manifold structures. We use this model to demonstrate the transfer of identity-preserving transformation and to explain non-trivial perceptual experiments in the psychophysics literature.

We base our approach on proposed techniques for learning Lie group operators (called manifold transport operators) that capture motion along a manifold (Culpepper & Olshausen, 2009; Sohl-Dickstein, Wang, & Olshausen, 2010). The main idea is to learn a dictionary of operators that describe transformations along a manifold and then infer, for any given pair of points, which operators are active. In detail, the model assumes that two nearby points in a state space $\mathbf{x}_0, \mathbf{x}_1 \in \mathbb{R}^N$, are related through the following equation:

$$\mathbf{x}_1 = \exp(\mathbf{A})\mathbf{x}_0 + \mathbf{n}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ is an operator that represents the transformation dynamics and $\mathbf{n}$ is the noise. The matrix $\mathbf{A}$ can be represented as a weighted sum of M dictionary elements ($\mathbf{\Psi}_m \in \mathbb{R}^{N \times N}$):

$$\mathbf{A} = \sum_{m=1}^{M} \mathbf{\Psi}_m c_m. \tag{2}$$

The set of dictionary elements $\{\mathbf{\Psi}_m\}$ are learned from pairs of data points using gradient descent. As one example, we could use keypoint locations (e.g., corners) from different frames of rotating objects in $\mathbb{R}^3$ as our state vectors and the operators will learn to transport objects in 3D along paths consistent with rigid body rotation.

Transferring Manifold Transformations

These transport operators represent identity-preserving transformations that can map out the low-dimensional

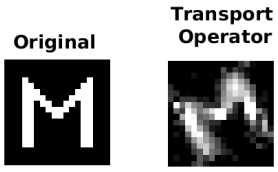

Figure 1: An example of manifold transformations trained on '8' digits being applied to the letter 'M'.

manifolds that objects exist on. In many cases, the physical processes producing variations in one class of objects can also impact many other classes. Therefore, the model for manifold transformations in one class may be transferred to other classes to induce similar transformations. Unlike most other manifold learning techniques, the transport operator manifold representation, once trained, is not reliant on the training points. This allows the manifold model to be transferred between classes. We utilize the USPS handwritten digit image dataset (Hull, 1994) to demonstrate our ability to use transport operator representations to perform transfer learning.

For training, we create a dataset that consists of 1000 examples of the digit '8' paired with that same image rotated 2.5 degrees and transport operators are learned between those point pairs. In other words, without telling the algorithms about the concept of rotation, we seek to have them learn the general transformation manifold from examples of only slightly rotated '8' digits. To highlight the performance when information is transferred between manifolds, we apply the transformation that was learned only on rotated '8' digits to the letter 'M' which is a new object class. Fig. 1 shows the original 'M' as well as the result after applying the learned transport operator transformation. Despite being trained only on slightly rotated '8' digits, the transport operator can rotate the 'M' by nearly 45 degrees while maintaining the shape of the letter and without inducing much distortion. This provides an example of how the transport operator model can transfer manifold transformations to new classes and extrapolate transformations beyond the original training samples.

3D Inference

As we showed in the previous section, this manifold transformation model can be powerful when images are input in their original pixel format. However, we want to learn about the physics of transformations in the 3D world and the data available to the visual system are the 2D points, $\mathbf{y}_0, \mathbf{y}_1 \in \mathbb{R}^2$, the projections of $\mathbf{x}_0$ and $\mathbf{x}_1$.

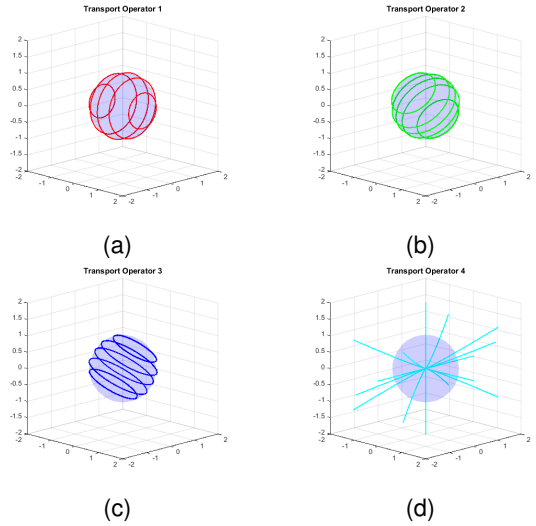

Figure 2: Trajectories for each of the four transport operators learned from 2D projections of points rotated in 3D. (a-c) show three transport operators that rotate in different orientations around the sphere and (d) shows a transport operator that scales inputs.

Because $\mathbf{y}_1$ is a transformed view of $\mathbf{y}_0$, we need only the depth of $\mathbf{y}_0$ and the transformation between the two points to estimate $\mathbf{y}_1$. Therefore, $\mathbf{y}_0$ and $\mathbf{y}_1$ are related through the following equation:

$$\mathbf{y}_1 = \mathbf{K} \exp\left(\sum_{m=1}^{M} \mathbf{\Psi}_m c_m\right) \begin{bmatrix} \mathbf{y}_0 \\ \lambda \end{bmatrix} + \mathbf{n} \qquad (3)$$

where $\mathbf{K}$ is a projection matrix (assumed to be orthographic projection) and λ is the depth associated with $\mathbf{y}_0$. Using this observation model and an unsupervised learning approach that includes the orthographic projection model in the inverse problem, we have used 2D stimuli of rotating objects to learn a manifold representation of 3D physical transformations such as rotation and scaling in a completely unsupervised model (illustrated in Fig. 2).

Random-dot kinematograms provide an example of how humans are able to infer a 3D structure from simple moving points in a 2D image without additional depth cues. Random-dot kinematograms are sequences of images with randomly placed dots that move along a transparent 3D shape. For instance, the dots in the left plot in Fig. 3a show one frame of a random dot kinematogram where the shape is a cylinder (perceived when the dots are in motion). We can apply the model described above that has been trained on rotating objects to jointly estimate the transformations and depths of the dots in the kinematogram stimulus. The left plot from Fig 3a shows one frame of the random-dot kinematogram from which it is impossible to determine the 3D structure of the points in isolation. The model

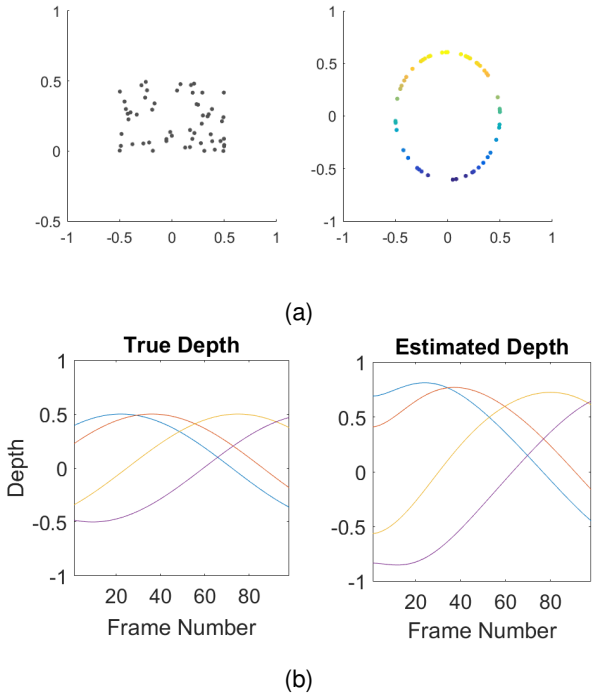

Figure 3: (a) Left: View of one frame of the random-dot kinematogram of a cylinder. Right: Top-down view of the depth inferred for each point in the kinematogram using the rotational transport operators. The points are colored by the true depth of each point. (b) Left: True depth for four points over a succession of frames. Right: Inferred depth for the same four points.

inference of depth from the moving video (shown in top-down view) is shown in the right plot from Fig 3a. Fig 3b provides a quantitative comparison of the true depth and the inferred depth for four points in the kinematogram. The estimated depths are very close to the true depths within a scale factor, despite the model never being told explicitly about the concept of rotation and never having access to data directly from a 3D model. Preliminary simulations show that this model accounts for some aspects of other classic psychophysical experiments such as reaction time in mental rotation tasks in the classic work of Shepard and Metzler (Shepard & Metzler, 1971).

Conclusion

We presented a model for how identity-preserving transformations are learned from 2D imagery. This model may be able to explain how the human visual system is able to learn natural transformations from a limited number of samples and use that knowledge to effectively recognize new objects in the future. Additionally, this model provides an example for how 3D transformations can be represented in the brain and applied to 2D visual stimuli.

Acknowledgments

This work was partially supported by NSF Graduate Research Fellowship grant number DGE-1148903, NSF grant number CCF-1409422, NSF CAREER grant CCF-1350954, James S.McDonnell Foundation grant number 220020399, and ONR grant N00014-15-1-2619.

References

Culpepper, B. J., & Olshausen, B. A. (2009). Learning transport operators for image manifolds. In *Nips* (pp. 423–431).

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

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434.

Hull, J. J. (1994). A database for handwritten text recognition research. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *16*(5), 550–554.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703. doi: 10.1126/science.171.3972.701

Sohl-Dickstein, J., Wang, C. M., & Olshausen, B. A. (2010). An unsupervised algorithm for learning lie group transformations. *arXiv preprint arXiv:1001.1027*.