Learning Internal Representations of 3D Transformations From 2D Projected Inputs

Marissa Connor
marissa.c.connor@gmail.com
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

Bruno Olshausen
baolshausen@berkeley.edu
Helen Wills Neuroscience Institute and School of Optometry, University of California, Berkeley, CA 94720, U.S.A.

Christopher Rozell
crozell@gatech.edu
School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, U.S.A.

We describe a computational model for inferring 3D structure from the motion of projected 2D points in an image, with the aim of understanding how biological vision systems learn and internally represent 3D transformations from the statistics of their input. The model uses manifold transport operators to describe the action of 3D points in a scene as they undergo transformation. We show that the model can learn the generator of the Lie group for these transformations from purely 2D input, providing a proof-of-concept demonstration for how biological systems could adapt their internal representations based on sensory input. Focusing on a rotational model, we evaluate the ability of the model to infer depth from moving 2D projected points and to learn rotational transformations from 2D training stimuli. Finally, we compare the model performance to psychophysical performance on structure-from-motion tasks.

1 Introduction

When interacting in a three-dimensional world, humans must estimate 3D structure from visual inputs projected down to two-dimensional retinal images. This problem of recovering 3D structure from 2D projections is generally underconstrained, as there are infinite numbers of possible depths for any given 2D point. To resolve this challenge, humans rely on a variety of cues when inferring depth, including motion parallax, binocular disparity, texture, occlusions, shadows, size, blur, and shading (Reichelt et al., 2010).

Neural Computation **36, 2505–2539** (2024)
https://doi.org/10.1162/neco_a_01695
© 2024 Massachusetts Institute of Technology

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

Specifically, the persistence of object shape over motion-induced transformations provides a powerful cue, even in the absence of other cues (Petersik, 1979; Braunstein et al., 1987; Todd et al., 1988; Dosher et al., 1989; Sperling et al., 1989; Braunstein, 2014), that can be used to resolve the depth ambiguity for points on an object's surface and improve accuracy of depth perception.

Mental transformation experiments (Shepard & Cooper, 1986; Lamm et al., 2007), as well as qualitative descriptions from subjects performing mental transformation tasks (Zacks & Michelon, 2005), suggest that humans internally imagine 3D spatial transformations when performing tasks such as identifying rotated reference objects (Shepard & Metzler, 1971; Cooper & Shepard, 1973; Just & Carpenter, 1985). However the mechanism in the brain for representing internal transformations is not well understood. The aim of this work is to present a model for how biological vision systems may internally represent 3D transformations, how they could learn or adapt to the statistics of these 3D transformations with minimal supervision, and how this knowledge could be used to aid in discerning structure from motion.

Motivated by the manifold hypothesis, which states that natural variations in high-dimensional data lie on or near a low-dimensional, nonlinear manifold (Fefferman et al., 2016), we introduce generative manifold models as a possible mechanism for learning and representing internal models of natural motion-induced transformations. These models represent manifolds through continuous, nonlinear transformation operators that traverse the geometric structure of the manifold (Culpepper & Olshausen, 2009; Sohl-Dickstein et al., 2010; Connor & Rozell, 2020; Connor, Canal, & Rozell, 2021; Connor, Fallah, & Rozell, 2021). The transformation operators can be used to infer relationships among different object views and to interpolate or extrapolate views of transformed objects. While transformation operators have previously been used to learn 3D transformations from 3D stimuli, we introduce a novel approach that adapts to the setting of 2D stimuli undergoing 3D motion. Our approach enables simultaneous inference of depth and transformational motion from 2D stimuli, and our model can be used to learn 3D transformations with limited supervision. Neuroscience research has suggested that the brain explicitly exploits the manifold structure of object variations by using hierarchical processing stages to flatten the manifolds produced by different objects undergoing the same physical transformations (e.g., changes in pose and position; DiCarlo & Cox, 2007; DiCarlo et al., 2012), but to our knowledge, no detailed model has been proposed for how a biological system could learn or represent the manifolds of such natural variations from data.

Notably there are precise definitions of motion-induced geometric transformations such as rotations and translations that can be employed to successfully compute point depths from multiple viewpoints (Longuet-Higgins, 1981; Fischler & Bolles, 1981; Tomasi & Kanade, 1992; Nistér, 2005;

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

Pollefeys et al., 2008; Snavely et al., 2006) or frames (Godard et al., 2017; Garg et al., 2016; Xie et al., 2016; Zhou et al., 2017), and there are many neural network–based models for successfully inferring point depth (Eigen et al., 2014; Ladicky et al., 2014; Liu et al., 2015; Godard et al., 2017; Garg et al., 2016; Xie et al., 2016; Zhou et al., 2017). However, these techniques are focused on achieving computer vision performance objectives, such as high computation speed and low depth reconstruction error instead of being focused on the biological plausibility of their algorithms. In contrast, our work presents a neurally plausible model, and our analysis is intentionally restricted to simple 2D stimuli of 3D motion that can provide a proof of concept for the viability of learning and adapting 3D transformation representations from 2D projected inputs. Focusing on the rotational motion that is used in many structure-from-motion tasks (Petersik, 1979; Dosher et al., 1989; Braunstein, 2014), we develop a manifold-based method for inferring depth from moving 2D projected points and learning 3D rotational transformation models from 2D training stimuli. Finally, we apply the learned transformation model to structure-from-motion tasks and compare to human performance on psychophysical experiments.

2 Background

In this work, we focus on the development of a model that can learn transformations that may be used to model internal mental rotation. While there have been computational models introduced for mental rotation (Funt, 1983; Fukumi et al., 1997; Inui & Ashizawa, 2011; Seepanomwan et al., 2015), they have assumed prior knowledge of the rotational transformations and been focused on modeling specific brain areas that are involved in this process. In contrast, we focus on the representation of the transformation model itself, including the learning and inference process within such a model.

We use our model of 3D transformations to infer point depths from 2D projections of moving points. The ability for humans to perceive depth from moving points and objects, known as the kinetic depth effect (Wallach & O’Connell, 1953), has been extensively studied in both psychology and computer vision. The kinetic depth effect has been investigated through a wide array of psychophysical experiments suggesting that humans can generate stable precepts of 3D structures under a wide variety of conditions (Petersik, 1979; Braunstein et al., 1987; Todd et al., 1988; Dosher et al., 1989; Sperling et al., 1989; Braunstein, 2014).

Computational models have been developed to estimate 3D point-cloud structure from multiple views of an object or scene. Multiview geometry (Longuet-Higgins, 1981; Tsai & Huang, 1984; Hartley, 1997; Hartley & Sturm, 1997; Hartley & Zisserman, 2004) and factorization methods (Tomasi & Kanade, 1992; Kanade & Morris, 1998) have been used to estimate the 3D structure from rigid body motion. Factorization methods have been extended to estimate nonrigid structure from motion by introducing

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

additional constraints encouraging low-rank (Bregler et al., 2000; Dai et al., 2014), union-of-subspaces (Agudo et al., 2018; Zhu et al., 2014), and block sparsity (Kong & Lucey, 2016, 2019). These techniques assume prior knowledge of the types of transformations present in temporal visual inputs (i.e., rotation and translation used to represent camera motion) and rely on a mathematical specification of how to apply rotational and translational motion through matrix multiplication.

Additionally, many neural network–based models have been introduced for estimating depth from the motion two-dimensional points. These models use 3D depth labels (Eigen et al., 2014; Ladicky et al., 2014; Liu et al., 2015), knowledge of camera viewpoints (Godard et al., 2017; Garg et al., 2016; Xie et al., 2016), and temporal consistency (Zhou et al., 2017) to supervise learning. Several methods use an image reconstruction objective by estimating depth, in one image, transforming it, and projecting it to compare against a second image (Godard et al., 2017; Garg et al., 2016; Xie et al., 2016). All of these methods for estimating structure from motion, while very successful at estimating camera motion and depth, have requirements that make them poor representations of the neural mechanisms for learning types of motion, inferring motion in scenes, and estimating point depths. We cannot assume that vision systems know ground-truth point depths or how to apply natural transformations during the development of mechanisms for estimating motion and depth of scenes. In this work, we learn a representation of 3D transformations from observed point motion itself, without a prior assumption of how these transformations affect points in a scene. Importantly, this model is learned using only 2D moving points without requiring ground-truth knowledge of point depth. This work does not aim to compete with current structure-from-motion techniques used in computer vision but instead aims to add to our understanding about how the models of motion that are essential for mental structure-from-motion tasks are developed internally.

3 Model Description

We aim to develop a model for learning 3D rotational transformations from 2D projected inputs and to use that model to describe how humans may employ motion cues to recover the 3D structure of objects in their environment. This perceptual setting is visualized in Figure 1 where an object is transforming in 3D but the visual inputs are in the form of 2D projections on the retina. In particular, each object is represented as a combination of 3D key points $\mathbf{x}^{(i)} \in \mathbb{R}^3$, $i = \{1, \ldots, N_P\}$ that are projected to 2D point locations $\mathbf{y}^{(i)} \in \mathbb{R}^2$. We assume rigid body motion and incorporate a transformation model that can constrain the possible 3D motion between different transformed viewpoints. This provides the structure necessary to infer depth from points on a moving object. We develop a method that uses a generative manifold model as a representation of transformations, and we show

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

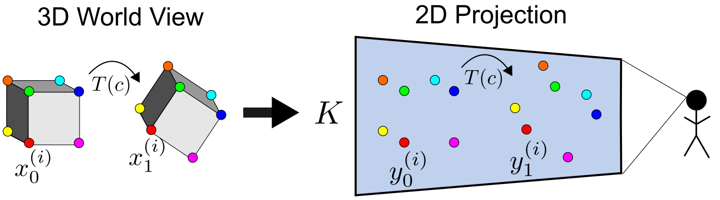


Figure 1: Visualization of the 3D depth inference problem. Three-dimensional points on an object are jointly transformed in the 3D worldview, and the visual inputs are in the form of 2D projected points.

that we can learn rotational transformation operators and use them to accurately infer point depth from rotating points and scenes. We build up to the learnable model of natural transformations in two steps. In the first step, we assume the 3D rotational transformation model is known, and we develop a method for inferring the depth of 2D projections of rotating points. In the second step, we use the depth inference approach from the first step to develop a learning model that can adapt the transformation representation to ensure it corresponds to the real-world transformations. We will preface the descriptions of each of these tasks with an overview of the transport operator model, a learnable generative manifold model.

3.1 Transport Operator Model. The transport operator technique is a specific manifold learning model that learns to transform points through nonlinear Lie group operators, known as transport operators, that transverse a manifold (Rao & Ruderman, 1999; Miao & Rao, 2007; Culpepper & Olshausen, 2009; Sohl-Dickstein et al., 2010; Cohen & Welling, 2014; Hauberg et al., 2016; Connor & Rozell, 2020; Connor, Canal, & Rozell, 2021). Lie group operators represent infinitesimal transformations that can be applied to data through an exponential mapping to transform points along a manifold. In particular, this model learns a dictionary of M transport operators $\mathbf{\Psi_m}$ that each represent a different transformation. These operators are effective for representing an internal transformation model for a few reasons. First, once learned, the transport operators are stored as a representation of possible transformations that may be experienced or observed. This means they can be reused in the future when the same type of transformation is visualized. Second, the transport operator model is a generative manifold model, meaning that it can interpolate and extrapolate new views of points undergoing a learned transformation. This provides a way of creating an internal visualization of how an object transforms similar to what humans describe when performing mental transformation tasks (Zacks & Michelon, 2005). Finally, transport operators can be used to infer the relationship between points in two separate viewpoints and define the 3D transformations between them.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

With the transport operator model, the relationship between two individual 3D points $\mathbf{x}_0^{(i)}$ and $\mathbf{x}_1^{(i)}$ is defined as follows:

$$\mathbf{x}_0^{(i)} = \mathrm{expm}\left(\sum_{m=1}^{M} \boldsymbol{\Psi}_m c_m\right) \mathbf{x}_1^{(i)} + \mathbf{n},$$

$$\mathbf{n} \sim \mathcal{N}(0, I) \quad c_m \sim \mathrm{Laplace}\left(0, \frac{1}{\zeta}\right), \tag{3.1}$$

where $\mathbf{c} \in \mathbb{R}^M$ is the set of coefficients that specifies the local structure of transformations between $\mathbf{x}_0^{(i)}$ and $\mathbf{x}_1^{(i)}$. Given this relationship between points, the original work from Culpepper and Olshausen (2009) defines the negative log posterior of the model as

$$E_\Psi = \frac{1}{2}\left\| \mathbf{x}_0^{(i)} - \mathrm{expm}\left(\sum_{m=1}^{M} \boldsymbol{\Psi}_m c_m\right) \mathbf{x}_1^{(i)} \right\|_2^2 + \frac{\gamma}{2} \sum_m \|\boldsymbol{\Psi}_m\|_F^2 + \zeta \|\mathbf{c}\|_1, \tag{3.2}$$

where $\|\cdot\|_F$ is the Frobenius norm and $\gamma, \zeta \geq 0$. The first term is a data fidelity term that specifies how well $\mathbf{x}_0^{(i)}$ can be represented as a transformed version of $\mathbf{x}_1^{(i)}$ when the transformations are constrained by the current dictionary of operators $\boldsymbol{\Psi}$. The data fidelity objective term is an indication of how well the transport operators fit the data manifold. The second term is a Frobenius norm regularizer on the dictionary elements that constrains the growth of the dictionary magnitudes and helps identify how many operators are necessary for representing transformations on the data manifold. The third term is the sparsity regularizer, which encourages each transformation between point pairs to be represented with a sparse set of coefficients.

Given a dictionary of operators $\boldsymbol{\Psi}$, the 3D transformation between $\mathbf{x}_0^{(i)}$ and $\mathbf{x}_1^{(i)}$ can be estimated by inferring a set of transport operator coefficients $\mathbf{c}$. This inference is performed by minimizing E_Ψ when $\gamma = 0$. If the operators need to be learned or adapted, E_Ψ is used as an objective for transport operator training as well. Training proceeds by alternating between performing coefficient inference between point pairs while fixing the transport operators and taking gradient steps on the transport operators while fixing the coefficients. This process of alternating between coefficient inference and dictionary updates is standard for sparse dictionary learning (Olshausen & Field, 1997).

We adapt the transport operator model to use time-varying views of transforming points in a 2D projection plane to learn a generative motion model. We begin by developing an inference procedure that enables joint depth estimation and coefficient inference from pairs of points of 2D inputs in different viewing frames.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

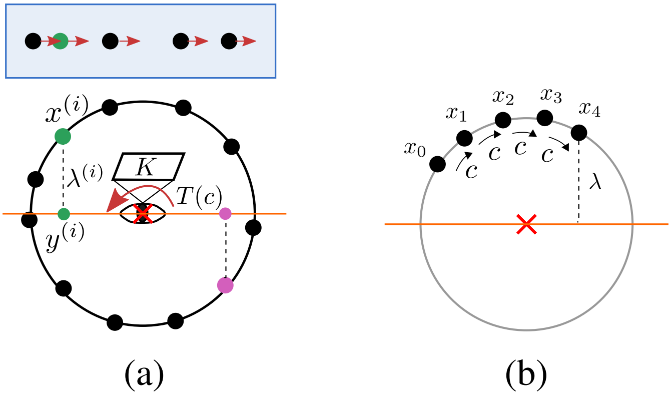


Figure 2: (a) Top-down views of the depth inference problem setup for points rotating on a cylinder. The 3D points $\mathbf{x}^{(i)}$ have an associated depth $\lambda^{(i)}$. Each point is projected onto the orange viewing plane using the projection matrix $\mathbf{K}$. This results in the 2D projected points $\mathbf{y}^{(i)}$. The 3D points are rotating counterclockwise around the axis and the points in the blue shaded box on top indicate the direction of motion of the projected points. (b) Visualization of the inference window sequence for a single point. The inference window is made up of several frames of transformed points. We assume that the transformation speed is constant between the frames, resulting a constant coefficient value representing the transformation from one frame to the next. The depth λ is inferred for the final frame in the sequence.

3.2 Depth Inference with Projected Inputs. In this section, we assume that the rotational transport operators are either known a priori or already learned. We describe the training procedure in section 4.2. Figure 2a shows a top-down view of the setup of this problem. The eye located at the red × in the center represents the viewer at the origin. The placement of the viewer at the origin is natural for learning a representation of self-motion in an egocentric viewing framework where the human is the origin. However, the model is flexible, and the same setup can be used to infer object motion in the allocentric viewing framework (see appendix A for more details). Each 3D point $\mathbf{x}^{(i)}$ is projected onto the viewing plane to a corresponding 2D point $\mathbf{y}^{(i)}$ with an associated depth $\lambda^{(i)}$: $\mathbf{y}^{(i)} = \mathbf{K}\mathbf{x}^{(i)}$. The matrix $\mathbf{K}$ is the orthographic projection matrix defined as $\mathbf{K} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$ in all of our experiments. This projection matrix corresponds to setting the viewing plane to the xy-plane and defining the unknown depth as the z-coordinate of the 3D input points. It is assumed that the $\mathbf{K}$ is known during processing. We observe N_P points transforming jointly on a rigid object and concatenate all N_P points into a matrix: $\mathbf{X}_0 = \left[\mathbf{x}_\mathbf{0}^{(\mathbf{1})} .... \mathbf{x}_\mathbf{0}^{(\mathbf{N_P})}\right]$.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

The relationship between points in two consecutive frames at $t = 0$ and $t = 1$ is defined as

$$\mathbf{Y}_0 = \mathbf{K}\mathbf{T}(\mathbf{c})\widehat{\mathbf{X}}_1\left(\boldsymbol{\lambda}\right) + \mathbf{W}, \tag{3.3}$$

where $\mathbf{W}$ is a gaussian noise matrix, $\widehat{\mathbf{X}}_1$ is a matrix of estimated 3D point locations associated with $\mathbf{Y}_1$, and $\mathbf{T}(\mathbf{c})$ is the matrix exponential of a weighted combination of transport operators that can each represent a different type of motion:

$$\mathbf{T}(\mathbf{c}) = \operatorname{expm}\left(\sum_{m=1}^{M} \boldsymbol{\Psi}_m c_m\right). \tag{3.4}$$

In equation 3.3, we define $\mathbf{Y}_0$ at $t = 0$ as a transformation of points at $t = 1$ in order to estimate the 3D point locations in $\widehat{\mathbf{X}}_1$ at $t = 1$ in a causal manner, as we describe below. To compute $\widehat{\mathbf{X}}_1$, we reverse the process of the projection matrix in two steps. First, we concatenate the $\mathbf{Y}_1$ with a vector of zeros in the z-coordinate position that is lost during projection:

$$\widetilde{\mathbf{X}}_1 = \begin{bmatrix} \mathbf{y}_{\mathbf{1}}^{(\mathbf{1})} & \cdots & \mathbf{y}_1^{(N_P)} \\ 0 & \cdots & 0 \end{bmatrix}. \tag{3.5}$$

Second, we add the depths to the newly introduced dimension. To do this, we compute the outer product between the standard basis vector associated with the axis lost during projection $\mathbf{e}_z$ and the depth vector $\boldsymbol{\lambda} \in \mathbb{R}^{N_P}$, resulting in a matrix with two rows of zeros and one row containing the estimated depths, and add that to $\widetilde{\mathbf{X}}_1$:

$$\widehat{\mathbf{X}}_1\left(\boldsymbol{\lambda}\right) = \widetilde{\mathbf{X}}_1 + \mathbf{e}_z\boldsymbol{\lambda}^\top. \tag{3.6}$$

This model for incorporating the estimated depths can be integrated into the data fidelity term of the objective function in equation 3.2 and used for jointly inferring the depth $\boldsymbol{\lambda}$ and coefficients $\mathbf{c}$ between point pairs:

$$L_{\text{df}} = \frac{1}{2}\sum_{i=1}^{N_P} \|\mathbf{y}_0^{(i)} - \mathbf{K}\mathbf{T}(\mathbf{c})\widehat{\mathbf{x}}_1^{(i)}\left(\boldsymbol{\lambda}^{(i)}\right)\|_2^2 \tag{3.7}$$

$$= \frac{1}{2}\operatorname{trace}\left(\left(\mathbf{Y}_0 - \mathbf{K}\mathbf{T}(\mathbf{c})\widehat{\mathbf{X}}_1(\boldsymbol{\lambda})\right)^\top\left(\mathbf{Y}_0 - \mathbf{K}\mathbf{T}(\mathbf{c})\widehat{\mathbf{X}}_1(\boldsymbol{\lambda})\right)\right). \tag{3.8}$$

We add two more constraints to this model to improve the consistency of accurate depth estimation. First, we incorporate a gaussian prior on the

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

depths, which constrains them to magnitudes consistent with the ground-truth depths of the rotating objects. Second, we group several consecutive views of the transforming points to reverse the projection procedure on points in the final frame in the sequence. We refer to this sequence of frames as the inference window. During inference and learning, we use ground-truth knowledge of point correspondences between frames. From this inference window, we can obtain a causal estimate of the depth in the final frame and infer a fixed set of coefficients that represents the transformations between each consecutive view. This assumes a fixed transformation speed over multiple frames, which can be seen as an extension of the slowness principle to natural transformations that persist over time (Wiskott & Sejnowski, 2002). Figure 2b shows this setting where the same coefficients **c** are inferred between points in each neighboring frame and the depth is inferred for the final point in the sequence. Using more than two motion frames for depth inference provides additional information that can be used to resolve depth ambiguities. To model this setting, we generalize equation 3.3 for N_T viewing frames,

$$\mathbf{Y}_{N_T-n} = \mathbf{K}\mathbf{T}^{n}(\mathbf{c})\widehat{\mathbf{X}}_{N_T}(\boldsymbol{\lambda}) = \mathbf{K}\mathbf{T}(n\mathbf{c})\widehat{\mathbf{X}}_{N_T}(\boldsymbol{\lambda}), \quad n = \{1, \ldots, N_T\}, \tag{3.9}$$

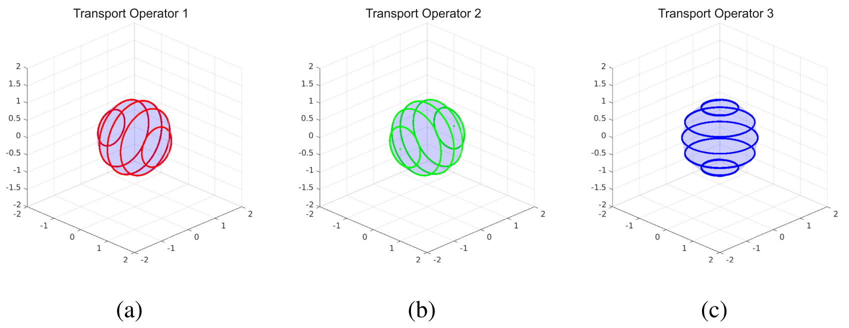
where the change from $\mathbf{T}^n(\mathbf{c})$ to $\mathbf{T}(n\mathbf{c})$ is possible because raising an exponent to the power of n is the same as applying the same transformation $\mathbf{T}(\mathbf{c})$ n times and thus multiplying its transformation coefficients by n.

We define an objective that leverages multiple views and a depth regularizer:

$$\begin{aligned} L = {} & \frac{1}{2N_T}\sum_{n=1}^{N_T}\sum_{i=1}^{N_P}\left[\|\mathbf{y}^{(i)}_{N_T-n} - \mathbf{K}\mathbf{T}(n\mathbf{c})\widehat{\mathbf{x}}^{(i)}_{N_T}(\boldsymbol{\lambda})\|_2^2\right] + \zeta\,\|\mathbf{c}\|_1 \\ & + \frac{\beta}{2}\|\boldsymbol{\lambda}\|_2^2 + \frac{\gamma}{2}\sum_{m}\|\boldsymbol{\Psi}_m\|_F^2. \end{aligned} \tag{3.10}$$

Notably the objective, equation 3.10, is nonconvex, presenting the possibility that inference may result in local minima. We take several steps to avoid the local minima. First, the gaussian prior on the depth is a regularizer that can incorporate information about expected depths of observed points, constraining the inference. Second, we perform inference for the same inference window several times using several random restarts. That is, we randomly sample a new initialization and infer coefficients and depths using that starting point. In particular, we use random initializations with several different variances to account for depths and transformations of different magnitudes. When using several random initializations, we choose the inferred output associated with the lowest final objective from inference.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024



(a) (b) (c)

Figure 3: Trajectories generated by ground-truth rotational transport operators. Each line represents the trajectory of an individual transport operator dictionary element applied to one of several example starting points selected on the sphere. These three operators generate rotation around each of the three principal axes.

With this objective, the depth $\boldsymbol{\lambda}$ and the coefficients $\mathbf{c}$ can be jointly inferred for sequences of transforming points. See section D for more details on the inference process. To highlight the effectiveness of this inference model, we will examine how accurately it can infer depths and transformations using ground-truth rotational operators and explore the requirements for the inputs that lead to robust depth estimation.

4 Results

4.1 Depth Inference Experiments. Three-dimensional rotational matrices can be defined as elements of the 3D rotational group SO(3), and ground-truth rotational transport operators can be derived from elements of the $\mathfrak{so}(3)$ Lie algebra (Hall, 2015). Figure 3 shows the trajectories of these ground-truth 3D rotational operators, each rotating around one of the principal axes. These plots are generated by selecting a few example starting points on a sphere and applying individual dictionary elements $\boldsymbol{\Psi}_\mathbf{m}$ to each point as they evolve over time: $\mathbf{x}_t^{(i)} = \text{expm}(\boldsymbol{\Psi}_m \frac{t}{T})\mathbf{x}_0^{(i)}, t = 0, \ldots, T$.

Using equation 3.10 and the fixed $\boldsymbol{\Psi}$ representing ground-truth rotational operators, we jointly infer the coefficients and depths from a sequence of transforming points. Figure 2b shows a visualization of this inference setting for a single point. Given N_T views of rotating points, we infer the depth λ for the projected points in the last viewpoint $\mathbf{Y}_{N_T}$, as well as the coefficients $\mathbf{c}$ that correspond to the shared transformation between every pair of consecutive views in the sequence. This ensures that the depths at time N_T are inferred by samples preceding it in the motion sequence, making this a causal estimate.

Figure 4 shows examples of depth inference for points on the surfaces of three different shapes. The plots in the first column show the visual input

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

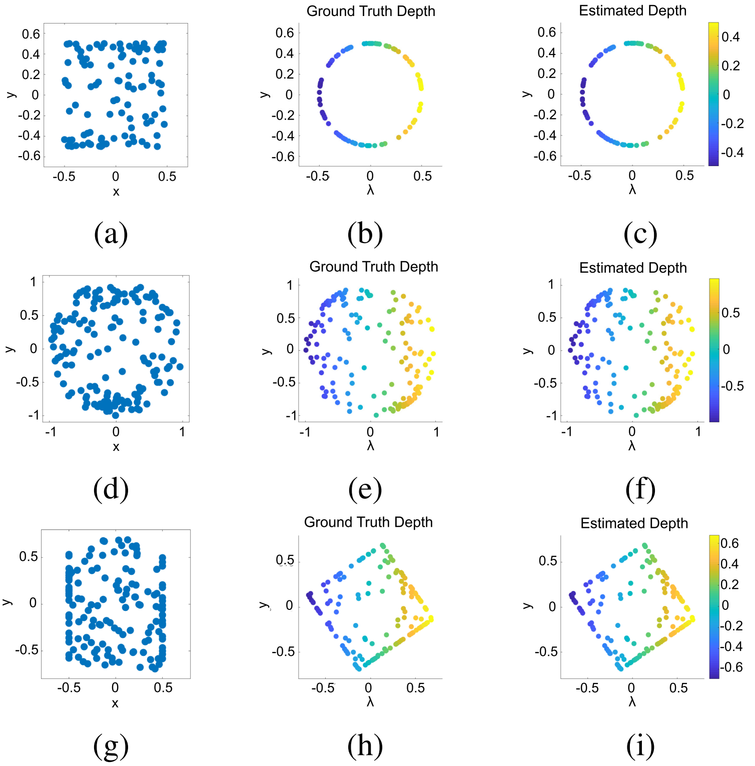


Figure 4: Inferred depths for points on the surface of different shapes. The first column shows the 2D point projections in the final viewing plane. The second column shows a side view of the ground-truth 3D point stimuli where the x-axis in the plots is the depth axis. The third column shows a side view of the estimated depths for the projected points. The points in the second and third columns are colored by the ground-truth depth. (a–c) Cylinder, (d–f) Sphere, and (g–i) Cube.

of projected points in the final viewing plane of the sequence. The plots in the second column show a side view of the point stimuli where the ground-truth depth locations are shown on the x-axis of the plot. The plots in the third column show a side view with the estimated depth locations for each of the points. In both the second and third columns, the points are colored by the ground-truth depths. This shows that the estimated depths correspond with the ground-truth depths for a variety of shapes.

We quantitatively evaluate the accuracy of the inferred depths for many trials to analyze the impact of various model parameters. There are three parameters of particular interest during inference. First, we are interested

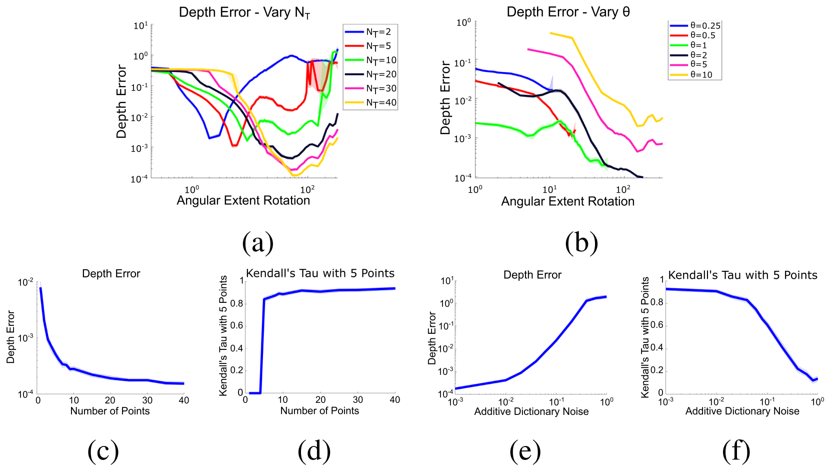
Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

in the impact of the perceptual extent of rotation viewed in a sequence of frames. The perceptual extent of rotation is a combination of two parameters: the number of frames in a rotation sequence N_T and the ground-truth rotation angle between each frame in the rotation sequence θ. The full angular extent of rotation viewed is $\theta_{\text{path}} = N_T\theta$. Experiments have shown that larger angular extents of rotation can lead to more accurate depth estimates for human subjects (Hildreth et al., 1990). Next we are interested in the impact of the number of jointly transforming points N_P. This indicates the amount of coherent rotational motion viewed in the input stimulus. Psychophysical experiments have been run that indicate that a greater number of coherently moving points leads to a more robust depth percept (Todd et al., 1988; Dosher et al., 1989; Sperling et al., 1989; Braunstein et al., 1987).

In order to quantitatively evaluate the success of inferred depth, we use two metrics. The first is the mean squared error between the estimated depths and the ground-truth depths for the N_P rotating points. Ideally depth inference would result in low MSE between the estimated depth and the ground-truth depth. However, with a rotational transformation, as we are working with here, there exists a depth-angle ambiguity. Namely, when viewing projected points $\mathbf{y}_0^{(i)}$ and $\mathbf{y}_1^{(i)}$ from two separate views, they could be either projections of points with large depths that undergo rotation with a smaller angle or points with small depths that undergo rotation with a larger angle (see appendix B for a visualization). While it is not ideal for the depth to be off by a scaling factor, the inferred structure can still be accurate. Additionally, this depth ambiguity is observed in experiments with human subjects (Todd et al., 1988). In psychophysical experiments, one metric for determining accuracy of a percept is comparing the estimated ordering of point depths to the ground-truth ordering of point depths (Hildreth et al., 1990). To analyze the accuracy of the inferred structure in the presence of potential scaling in depth, we compare the ordering of the inferred depths to that of the ground-truth depths using Kendall's tau rank correlation coefficient (Kendall, 1938). We compare Kendall's tau between all points N_P as well as between five randomly selected points. We choose to compare the ordering of five randomly selected points in order to define a metric that can be used to fairly compare the performance as the number of points increases. With greater N_P, even if depths are accurate within some error range, there is a greater chance of incorrectly ordering a few points because there is a greater point density. Therefore, comparing five randomly sampled points should provide a consistent metric as we vary N_P.

Figures 5a and 5b show the median depth error as we vary the angular extent of rotation θ_{path}. In Figure 5a, each line represents a different number of frames N_T. The error bars in all plots represent the bootstrap confidence interval. For each line in Figure 5a, because the values of N_T are fixed, moving along the x-axis corresponds to increasing the angle between frames θ. Each of these lines has a clear minimum, and this minimum occurs at an angular extent in the range of N_T to $2N_T$. This corresponds to an angle θ

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024



(a) (b)

(c) (d) (e) (f)

Figure 5: Quantitative metrics for depth inference when varying the angular extent of rotation, the number of coherently transforming points N_P, and the standard deviation of gaussian noise added to the ground-truth operators. (a) Median depth error as angular extent increases. Each line is generated with different numbers of frames in the inference window N_T. The optimal performance occurs for angular extents in the range of N_T to $2N_T$. (b) Median depth error as angular extent increases. Each line is generated with different angles of rotation between sequence frames θ. A rotation angle of $\theta = 2$ results in the lowest depth error at 180° of rotation. (c) Median depth error as N_P increases. (d) Mean Kendall's tau for five randomly selected points as N_P increases. Values of this metric for $N_P < 5$ are set to zero because there are not enough points to compare five randomly selected points. (e) Median depth error as the standard deviation of dictionary noise increases. (f) Mean Kendall's tau for five randomly selected points.

between frames of 1° to 2°. Up to this minimum value, the depth error decreases as the rotational extent increases.

Figure 5b breaks down the performance for individual values of θ. As θ increases up to $\theta = 2°$, the depth error decreases with increasing θ. For θ greater than 5°, the performance starts to degrade. This is consistent with the patterns in Figure 5a, and it indicates that rotation angles that are too large between frames (corresponding to fast rotational motion) result in less accurate depth inference.[1] In the remaining tests of inference performance

[1] We should note that inference optimization experiences inaccuracy even with a large number of random restarts as greater numbers of frames are used in the inference window, leading to large increases in depth error for $N_T > 50$ in many settings. Therefore, in Figure 5b, we only display lines until an angular extent of rotation for which the

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

with ground-truth operators, unless otherwise stated, we set $N_T = 30$ and $\theta = 2$. See appendix D for more details on model parameters.

In these results, we see the impacts of the nonconvex objective function on the performance of this model. At smaller rotational extents, the inference objective does not receive enough information to allow for successful disambiguation between many possible solutions. This can result in the inference procedure finishing in local minima with low objective function values but with depths and coefficient values that do not reflect the true point geometry and rotation sequence. With larger rotational extents, the inference objective receives a more complete view of the transformation that better constrains the optimization, resulting in low objective function values corresponding more directly with accurate depth inference. With high-speed transformations that have large rotation angles between views, the magnitudes of the coefficients corresponding to the true rotation increase. This increases the space of possible coefficient values. Therefore, randomly selecting initializations of the coefficients in the neighborhood of the true minimum is less likely, which results in more solutions that correspond to inaccurate local minima. The optimal values between N_T and $2N_T$ have a large enough rotational extent such that low objective function values correspond to accurate depths but small enough rotational extent that we can randomly initialize inference with coefficients that result in minima close to the ground-truth depth values.

While the characteristics above present challenges for accurate optimization, they are informative about the situations in which we could expect successful depth inference. Namely, we see the best depth inference for angular extents between N_T and $2N_T$ with the angle between frames θ of around 2°. These constraints that are necessary for effective depth inference in our model correspond to real-world constraints. For instance, with this model of depth estimation, we would anticipate a human's depth percept to improve with greater rotational extent, and we would anticipate difficulty with inferring depth when the speed of rotation is too great. Interestingly, there is psychophysical research that shows that humans build up an accurate estimate of depth when viewing longer rotational sequences (Hildreth et al., 1990) as well as research that shows human subjects identify the depth of moving points less often with larger angular rotation between frames (Mather, 1989).

Figures 5c and 5d show the median depth error and the mean Kendall's tau as we vary N_P. This shows that the depth estimation improves as more points are added with a large performance improvement from $N_P = 1$ to $N_P = 10$. We reason that this improvement is due to the reduction in transformation ambiguity that results from seeing more points rotating jointly.

nonconvex optimization inaccuracy affects the solutions. The angular extent where this occurs is smaller for the smaller values of θ because they require larger values of N_T to achieve the same angular extent of rotation.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

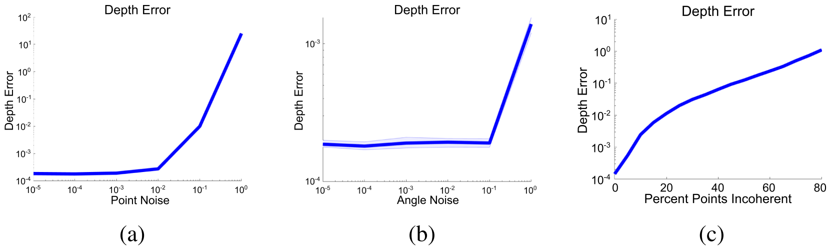


Figure 6: Depth error for depth inference in the presence of mismatch between the model assumptions and stimulus characteristics. (a) Depth error as the standard deviation of additive gaussian noise in the 2D point locations increases. The model is robust to noise with a standard deviation up to around 10^{-2}. (b) Depth error as we increase the standard deviation of the gaussian noise added to the magnitude of the rotation angles between frames in the inference window θ. The model is robust to noise with a standard deviation up to around 10^{-1}. (c) Depth error as the percent of incoherently moving points is increased. The depth error quickly increases after 5% to 10% of points are incoherent.

The greater number of points on a rigid object undergoing the same transformation, the more information our model has about the accurate transformation and depth. Going forward, we use $N_P = 20$. Research in structure from motion has shown that increasing the number of transforming points improves the general depth percept (Todd et al., 1988; Dosher et al., 1989; Sperling et al., 1989), but it may not increase the accuracy of the inferred depths (Braunstein et al., 1987).

To better understand the performance of this model in the real world, we analyze the robustness of our model when there is a mismatch between the theoretical assumptions and the real-world data characteristics. Figure 6 shows the impact of adding noise to the point locations, adding variation to the rotational speed, and varying the percentage of points moving in a coherent direction.

Figure 6a shows the depth error as we add gaussian noise with progressively larger standard deviations to the 2D point locations. This is similar to a human having a slightly inaccurate estimate of the points' locations in space. Results indicate that the model is robust to noise with a standard deviation of up to around 10^{-2}. For reference, points in these experiments have values in the range of $[-1, 1]$.

Figure 6b shows the impact of variations to the speed of rotation between frames. Our current model assumes that each frame in the inference window results from rotations with the same speed and direction as preceding frames. In natural systems, objects are unlikely to undergo rotation at the exact same speed over many frames. To test the effect of variations in the rotation speed, we create rotation sequences in which each frame uses the

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

same axis of rotation, but the magnitude of the rotation angle is varied by adding gaussian noise with a standard deviation scaled by the angle of rotation.[2] Our model is robust to variations in rotation speed up to a standard deviation of around 10^{-1}.

Finally, we analyze the performance of our model in the presence of points that are moving incoherently with the rotating points. Figure 6c shows the depth error as we increase the percentage of incoherently moving points when $N_P = 100$. This shows that incoherent points significantly affect the accuracy of the 3D structure estimate. This analysis begins to quantify impacts of experimenting in settings closer to the real world.

The final quantity we analyze in this controlled setting with ground-truth rotational operators is the effect of adding gaussian noise to the operators. Noisy operators depart from the ground-truth rotational transformations, and analyzing the performance with noisy operators can indicate the impact of accurate rotational transformation models on effective depth inference. Figures 5e and 5f show the median depth error and mean Kendall's tau metrics as noise is added to the ground-truth operators. Both metrics indicate that the depth inference is robust to noise with a standard deviation of around 10^{-3} - 10^{-2} but performance decreases sharply with noise larger than that. This shows that the model can perform effectively with some transformation inaccuracy, but performance decreases with increasingly inaccurate transport operators. This highlights the necessity of accurate rotational operators and inspires the learning and adaptation procedure introduced and analyzed in the next section.

4.2 Learning 3D Transport Operators from 2D Projected Inputs. A stated goal of this work is to develop a model that can learn 3D transformational representations from rotating 2D projected input points. The learning procedure is a straightforward extension of the coefficient and depth inference model from the previous sections. Training of the transport operator dictionary elements is performed using gradient descent. For each training step, a sequence of projected rotated points $\mathbf{Y}_n$, $n = \{1, \ldots, N_T\}$ is generated. First, the dictionary weights are fixed, and the depth and coefficients are inferred. Then, fixing the depth and coefficients, the gradient on the dictionary elements is computed using the objective in equation 3.10 with $\zeta = \beta = 0$. If this gradient step improves the objective, then it is accepted. Otherwise, the step is rejected and the learning rate is decreased. (See appendix E for more details on the training procedure.)

With this training procedure, we are able to learn rotational transport operators from randomly initialized operators. Figure 7 shows the trajectories of the operators during one training run in which the number of dictionary elements M is set to 3. At the beginning of training, the trajectories

[2] The angles used in this experiment have a mean of 2°.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

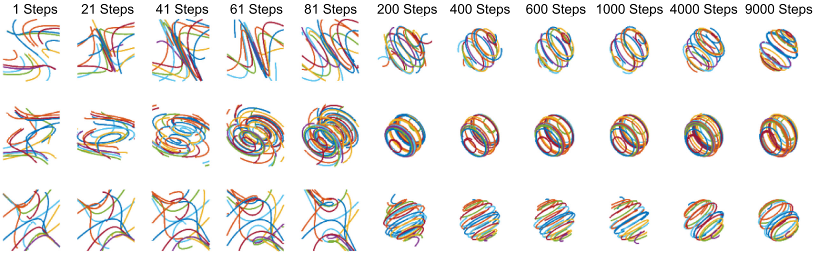


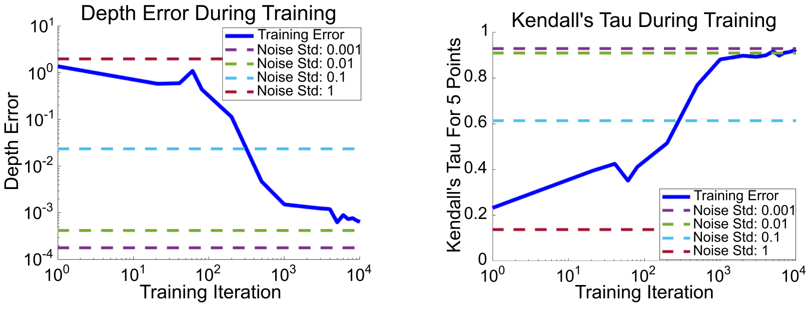
Figure 7: Transport operator trajectories during training. Each row represents one of the three learned operators. Each column shows the trajectories at a different training step. The operators begin with random initializations at step 1 and quickly reach a rotation structure around 200 steps. From 200 steps to 9000 steps, the operators vary relatively slowly, resulting in operators with clear rotational structure at the end of training.

do not correspond to common geometric transformations, but they quickly adapt to represent near-rotational operators with trajectories similar to the ground-truth operators shown in Figure 3. In appendix E, we show an example of learning rotational operators from a dictionary with six operators.

We quantitatively compare these operators to the ground-truth operators using the same depth MSE and Kendall's tau metrics employed to analyze inference success. We can compute the depth error and Kendall's tau metrics for inferred depths using operators at various points during training and compare them to the metric values resulting from depth inference using the ground-truth operators with noise added. This gives us a proxy for estimating the deviation between the learned operators and ground-truth rotational operators. In Figure 8, we show the depth error and Kendall's tau for depths inferred using transport operators at different points in the training procedure. For reference, we also plot straight lines that correspond to the values for these metrics in Figures 5e and 5f, which are computed using ground-truth rotational operators with added noise with standard deviations of 10^{-3}, 10^{-2}, 10^{-1}, and 1. This shows that our method learns operators that are close in structure to the ground-truth operators, and the performance they achieve is similar to ground-truth operators with additive gaussian noise with a standard deviation of 10^{-2}. Additionally we see that the depth inference performance with learned operators improves significantly over the first 100 to 200 training steps but requires fine-tuning for many steps after that to achieve optimal performance.

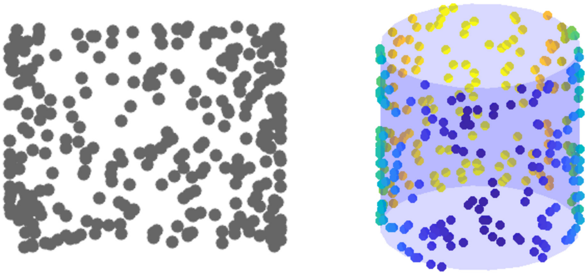
4.3 Kinematogram Experiments. Random dot kinematograms are displays of dots on the surface of or within invisible rotating shapes. Still frames of kinematogram inputs appear as random dots with no perceptible structure (see Figure 9a). However, the motion of the dots elicits the

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024



(a) (b)

Figure 8: Inferred depth metrics using operators at different steps in training. Dashed lines represent values of error metrics for depths inferred using ground-truth operators with additive gaussian noise with the standard deviation specified in the legend. These values are obtained from the plots in Figures 5e and 5f. (a) Median depth error for depth inference performed with operators at different steps in training. The depth error decreases significantly after 200 training steps and continues to decrease until the end of training. The depth error achieves a value consistent with the estimates using ground-truth operators with an additive gaussian noise with a standard deviation of 10^{-2}. (b) Mean Kendall's tau for five randomly selected points. Kendall's tau increases significantly around 200 training steps and starts to plateau around 1000 training steps. It reaches a value consistent with ground-truth operators with a noise standard deviation of 10^{-3}.



(a) (b)

Figure 9: Visualization of kinematogram visual stimuli. (a) Example of a 2D kinematogram stimulus, which is the projection of random dots on the surface of a cylinder. (b) 3D ground-truth structure of the points in the kinematogram stimulus. The points are randomly sampled from the cylinder surface and colored by their depth.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

perception of a 3D structure. Figure 9 shows the 2D projection of random points along with the 3D structure of the points on the surface of a cylinder. This perception of depth through motion is termed the "kinetic depth effect" (Wallach & O'Connell, 1953). The random dot kinematogram visual stimulus has been used for many structure-from-motion experiments because it isolates the use of motion cues from the use of other possible depth cues. We use our depth inference model with transport operators learned from 2D projections of rotational motion in order to estimate depths for random points that are located within the volume of invisible rotating shapes. We compare characteristics of our experimental results to the performance of humans on structure-from-motion tasks with random dot kinematograms.

For these experiments, we create kinematogram stimuli by randomly selecting N_P 3D points within the volume of a cylinder. Sequences are generated by rotating points around the x-axis at a rotational speed specified by the angle between frames θ. The points in each frame are ortographically projected to the xy-plane. Point correspondences are estimated by pairing nearest neighbors in the projected inputs from one frame to the next using the Hungarian algorithm (Kuhn, 1955). We use the inference procedure described in section 3.2 to infer the depth and coefficients. During inference, we use an inference window of $N_T - 1$ preceding frames to infer depths for the points in the current frame. We can vary the parameters of the stimuli and the inference procedure and analyze their impact.

Figure 10 shows depths that are inferred for a random dot kinematogram sequence on a cylindrical structure by minimizing the objective in equation 3.10. In this experiment, $N_P = 20$, $\theta = 2^\circ$, and $N_T = 30$. Each line in the top and middle plots of Figure 10a is the depth for one of five stimulus points. In the early stages of the kinematogram sequence, the number of frames in the inference window is only as large as the number of frames that have appeared (which is less than N_T). Once more than N_T frames have appeared, the depth and coefficient inference will make use of only the current frame and the $N_T - 1$ preceding frames. This buildup in the angular extent of rotation explains the larger depth errors early in the sequence, and we will analyze this further below.

The estimated depth in Figure 10 has discontinuities that result in the sign of the depth switching. This natural phenomenon is due to the fact that the orthographically projected random dot kinematogram stimulus is a bistable perceptual representation (Andersen & Bradley, 1998). That is, it is an ambiguous representation in which there are two correct perceptual structures. All the points could be rotating in a clockwise direction with a specific combination of positive and negative depths, or they could be rotating in a counterclockwise direction with the opposite combination of positive and negative depths. Each of these perceptual estimates is equally correct for the stimulus. Therefore, when computing the error metrics, we correct for the direction of the inferred rotation so it corresponds to the

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

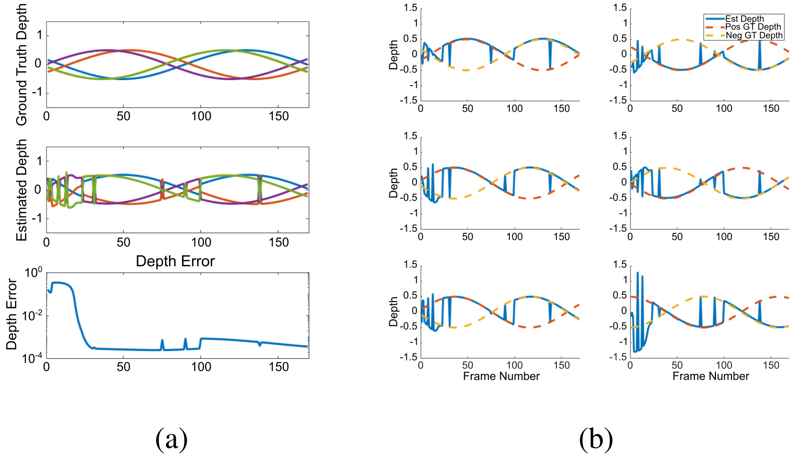
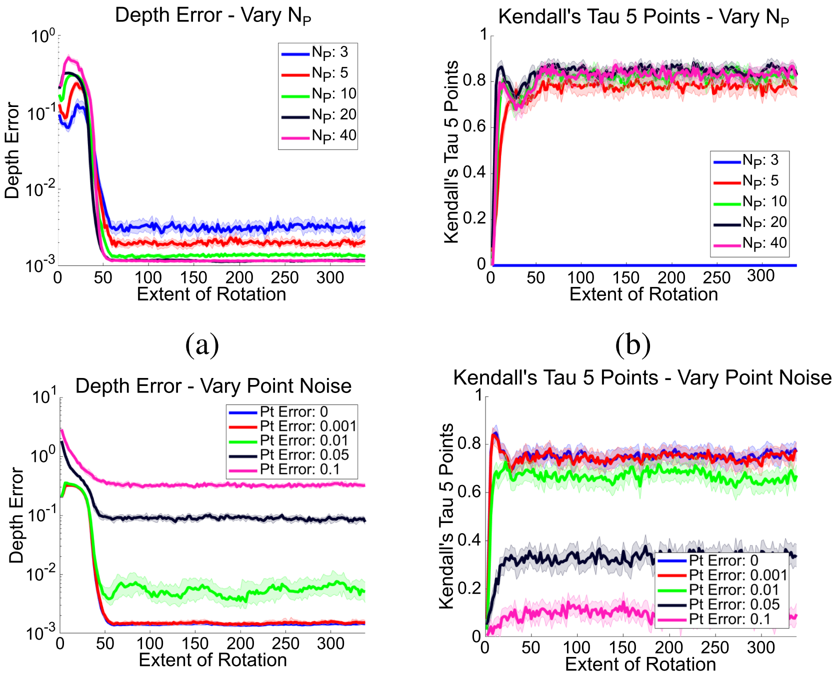


Figure 10: Example of depths inferred for random dots in a kinematogram sequence. (a) In the top plot, each line represents the ground-truth depth of a single random point over the rotational sequence of the kinematogram. In the middle plot, each line represents the estimated depth for the same points as in the top plot. The bottom plot shows the depth error between the estimated and ground-truth depths over the sequence. (b) Each plot shows the estimated depth for a single point with the sequences of positive and negative ground-truth depths overlaid.

ground-truth direction (which is clockwise in all of our experiments). This is done by generating a path with the inferred transformation coefficients and identifying the rotation direction of the points on that path. If the inferred rotation is moving in a counterclockwise direction, we reverse the signs of the depths prior to computing the error metrics. The bottom plot of Figure 10a shows the depth error for the kinematogram sequence. The depth error is high at the beginning of the sequence due to the limited angular extent of rotation. As the angular extent of rotation increases, the depth error decreases and remains low even while the signs of the depths switch. In Figure 10b, we overlay the estimated depth for individual points on top of sequences with both positive and negative ground-truth depth values. This shows, whichever direction of rotation is inferred, that the depths are aligned with either the positive or negative ground-truth depth values.

This bistable phenomenon is observed in pyschophysical experiments as well. Specifically, subjects incorrectly identify the rotation direction of orthographically projected stimuli 50.3% of the time (Petersik, 1979). In the experiments shown in Figure 11a, the clockwise rotation is estimated 50.04% of the time.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024



(a) (b)

(c) (d)

Figure 11: Quantitative metrics for random dot kinematogram depth estimates. Depth error and Kendall's tau: (a, b) as N_P increases. (c, d) as the standard deviation of noise added to the point locations increases.

Figure 11 contains plots demonstrating the performance of our model on random dot kinematogram stimuli as we vary parameters of both the inference algorithm and the kinematogram inputs. These plots compute the depth MSE and Kendall's tau for five randomly selected dots in the stimulus. Figures 11a and 11b examine the influence of the number of stimulus points.[3] Increasing the number of points improves the accuracy of the depth estimates but does not significantly affect the accuracy of the depth ordering. The Kendall's tau values for $N_P \geq 10$ have a spike at the beginning of the kinematogram sequence. This is due to the trade-off between the data fidelity term and the depth regularizer term in the objective. With fewer rotational frames in the inference window early in the kinematogram sequence,

[3] Note that in this experiment, we use ground-truth correspondences between sample points in each kinematogram frame in order to focus on the impact of N_P on the inferred depth sequence independent of our point correspondence technique.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

the optimization results in large magnitudes for the inferred depths that lead to small errors in the data fidelity term but large values for the depth regularizer. Therefore, the ordering of the depths is accurate, but the exact depth values are inaccurate because they are off by a scale in magnitude. As the kinematogram sequence continues, with more frames in the inference window, the magnitudes of the depths decrease and reduce the depth regularizer term, but this leads to an increase in the data fidelity term associated with less accurate depth ordering. We see this as the depth error decreases (because depth magnitudes are reducing) in conjunction with a decrease in Kendall's tau values. Decreasing the number of points eliminates this spike.[4] We examine the impact of adding gaussian noise to the point locations in Figures 11c and 11d. The depth is consistently accurate with point location noise up to a standard deviation of 10^{-2} and depth error increases after that. The introduction of point noise also eliminates the spike in Kendall's tau at the beginning of the sequence.

This perceptual buildup of an accurate estimate of point depths is observed in structure-from-motion experiments (Hildreth et al., 1990). Hildreth et al. performed experiments where they displayed orthographic projections of three points rotating about a central axis and asked subjects to order the depths of the three points. They computed the percentage of correct ordering responses—a metric similar in nature to our Kendall's tau metric. They found that the percent of correct depth ordering increased as the angular extent of rotation increased up to about 40° of rotation, after which it plateaued. We observe the same buildup and plateau of point ordering accuracy (as judged by Kendall's tau). They also observed degradation in performance as gaussian noise was added to the point locations, as we see in Figure 11d.

5 Discussion

The main contribution of this work is a generative model framework for learning and inference of 3D manifold-based transformations from 2D projections. A key innovation of the model is an inference procedure that jointly estimates scene geometry (point depth) and transformation parameters from a sequence of 2D views via gradient descent through a transport operator. Using this procedure, we show that it is possible to learn, without any prior knowledge of transformations or point depth, the dictionary elements that generate rotational motion from 2D projections of rotating points. This model lays the groundwork for explaining the development and adaptation of internal representations of natural variations that are observed in the world. Additionally, our depth inference model enables the

[4] The Kendall's tau we report in Figure 11b is for five randomly selected stimulus points, so the value for $N_P = 3$ is set to zero because there are fewer than five points to use for this metric computation.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

investigation of data characteristics that may influence the capacity for accurate depth estimation. This allows us to connect model performance with various data characteristics and algorithmic parameters to human performance on perceptual studies.

An important factor in accurate depth estimation and ordering is that a large angular extent is spanned by the set of input frames used for inference (see Figures 5 and 11). This supports the notion that humans build up their perception of 3D structure during random dot kinematogram rotation sequences (Hildreth et al., 1990). We also show that increases in the number of random dot stimuli result in improvement in depth inference performance (see Figures 5c, 5d, 11a, and 11b). This connects to kinematogram experiments that indicate a greater number of coherently transforming points results in a stronger depth percept from moving points (Todd et al., 1988; Dosher et al., 1989; Sperling et al., 1989; Braunstein et al., 1987). Our model also demonstrates the same direction switching phenomenon with the bistable kinematogram stimulus that humans perceive (Petersik, 1979).

5.1 Psychophysical Implications. Our model has the flexibility to adapt to many different test scenarios that are inspired by human performance on mental rotation and structure-from-motion tasks. In our experimentation, we tuned parameters like the inference window length N_T, the angle between frames θ, and the number of stimulus points N_P to achieve the most accurate depth estimates. However, experiments show that even when humans perceive the correct shape, they often have inaccurate estimates of depth magnitude (Todd et al., 1988), especially when viewing limited numbers of transforming points (Dosher et al., 1989). From our model performance, this may suggest that humans rely on a smaller angular extent of rotation for inferring depth or that they do not utilize a prior on expected depths. In the future, we can vary our model parameters to explore comparisons with potentially inaccurate human depth estimation in various test settings.

In this work, we do not directly relate the internal rotation model developed here to the rich area of mental rotation experiments. In particular, the seminal work in that area suggests a monotonically increasing relationship between rotation angle between views of an object and the human reaction time (Shepard & Metzler, 1971). The manifold-based model presented here may have a similar connection between processing time and rotation angle because the transport operators can generate transformations similar to the internal representation of 3D rotations described by humans in these studies. It may be fruitful to examine the performance of our transformation model on mental rotation tasks and compare to human performance on similar tasks.

5.2 Future Improvements. Ultimately, addressing the underlying neural mechanisms of 3D perception will require formulating a more

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

biologically plausible model. The inference and learning in the current model are performed using quasi-Newton optimization and gradient descent, respectively. The optimization objective is nonconvex and does not naturally lend itself to a parallel representation similar to neural architectures. Moving forward, we suggest developing an optimization procedure that is more biologically plausible.

As our focus in this work is on the development of a transformation learning framework, we assume ground-truth point correspondences for the learning and inference experiments in section 4.1 through section 4.2. However, identifying point correspondences from different views of the same scene is a challenging task and one that has been a focus of many computer vision algorithms (Ullman, 1979; Fischler & Bolles, 1981; Zbontar et al., 2016; Luo et al., 2016). Going forward, incorporating point correspondence estimates into this framework will lead to a more versatile, biologically plausible model.

Finally, a step toward biological plausibility is extending this model to be robust to incoherent point motion and additional moving objects. An initial approach to improving the robustness to incoherent motion is to employ random sample consensus (RANSAC; Fischler & Bolles, 1981). With this method, transport operator coefficients could be estimated from random subsets of points in the scenes, and the final transformation parameters could be chosen as those that explain the transformation between the largest number of random subsets of points.

Appendix A: Egocentric versus Allocentric

The model presented can be applied to two viewing frameworks: the allocentric framework in which points rotate around the observer and the egocentric framework in which the observer rotates with respect to the surrounding world. When the motion is centered around the origin (i.e., when the origin of the observer coordinate system and the world coordinate system is the same), the allocentric and egocentric learning frameworks utilize the exact same model. Comparing Figures 2a and 12a, the only difference between the allocentric and egocentric frameworks when motion is centered at the origin is the direction in which the projected points move with respect to the rotational motion. When the viewer rotates in a counterclockwise direction in the egocentric framework, the projected points with positive depth (i.e., points in front of the viewer) move to the right in the viewing plane. When the points rotate in a counterclockwise direction in the allocentric framework, the projected points with positive depth move to the left in the viewing plane. Therefore, the model and experiments can correspond interchangeably to the egocentric or allocentric frameworks.

For the work presented here, we are assuming that the motion generated by the transformations we wish to learn is centered at the viewer location.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

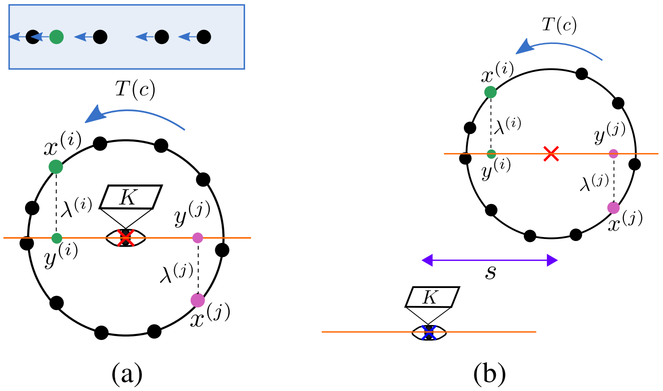


Figure 12: Top-down views of the depth inference problem setup in an allocentric framework where the object rotates around the observer. The 3D points $\mathbf{x}^{(i)}$ have an associated depth $\lambda^{(i)}$. Each point is projected onto the orange viewing plane using the projection matrix $\mathbf{K}$. This results in the 2D projected points $\mathbf{y}^{(i)}$. (a) Example when the origin of the motion is located at the viewer location. The 3D points are rotating counterclockwise around the axis, and the points in the blue shaded box on top indicate the direction of motion of the projected points. (b) Example when the origin of motion is offset from the viewer location.

However, if the origin of transformational motion is offset from the viewer location (the example shown in Figure 12b), this model can be easily extended to incorporate an origin offset. This offset may be known a priori or estimated by viewing the point motion over several frames.

Appendix B: Depth-Angle Ambiguity

Figure 13 shows a visualization of the depth-angle ambiguity that exists with orthographic projection. When viewing projected points $\mathbf{y}_0^{(i)}$ and $\mathbf{y}_1^{(i)}$ from two separate views, these points could be either projections of points with large depths that undergo rotation with a smaller angle θ_a or points with small depths that undergo rotation with a larger angle θ_b.

Appendix C: Visualization of Operator Noise

Figures 5e and 5f show the quantitative impact that adding noise to ground-truth rotation operators has on the accuracy of inferred depth. To provide an intuitive understanding of the effect of noise on the operators, Figure 14 shows the example trajectories for operators with increased noise standard deviation.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

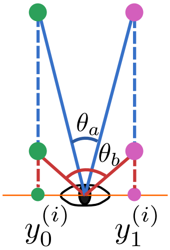


Figure 13: Example of depth-angle ambiguity. Points $\mathbf{y}_0^{(i)}$ and $\mathbf{y}_1^{(i)}$ are projections of a 3D point that is transformed from view 0 (green points) to view 1 (pink points). These points could result from a small rotation angle θ_a on points with large depth magnitudes or a larger rotation angle θ_b with smaller depth magnitudes.

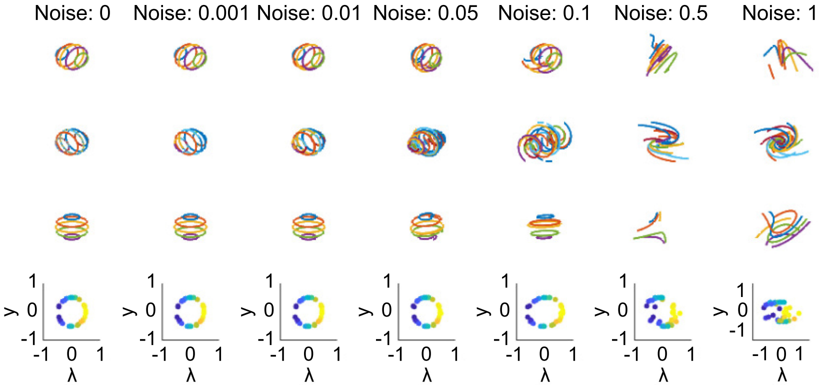


Figure 14: Examples of noisy operator trajectories. Each column shows examples of the ground-truth rotational operators with additive gaussian noise with increasing standard deviation. Rows 1–3 show the trajectories for the three rotational operators. Row 4 shows depth inferred for points projected from a rotating cylinder using the three operators in each column. The operators do not vary much in appearance from the ground-truth operators (first column) with noise standard deviations of 0.01 or less. For noise standard deviations larger than that, the operators diverge from rotational operators, and the point depths no longer look like a cylinder.

Appendix D: Inference Details

We perform depth inference in sections 4.1 and 4.2 using the objective in equation 3.10. In the inference experiments in section 4.1, we set $\gamma = 0.1$, $\zeta = 0.01$, and $\beta = 10^{-3}$. These parameters are selected because they yield accurate inferred depth estimates. We add gaussian noise with a standard

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

deviation of 10^{-3} to the ground-truth operators in most experiments in section 4.1. Unless otherwise stated, we set $N_P = 20$, $\theta = 2^\circ$, and $N_T = 30$. For the quantitative analysis of inference on many trials, we use input points that are 2D projections of random 3D points that are undergoing rotation about a randomly selected 3D axis. The rotation angle used for a given trial is sampled from a distribution with a mean of θ and a standard deviation of 0.516°.

Because the training objective is nonconvex, optimization may result in local minima. To avoid resulting in local minima, we perform inference for the same inference window several times using several random restarts. That is, we randomly sample a new intialization and infer coefficients and depths using that starting point. This often results in different final inferred outputs. We choose the inferred output associated with the lowest final objective from inference. For the experiments in section 4.1, we use five random restarts.

For the inference experiments, we compute the mean squared error between the ground-truth depths and estimated depths and Kendall's tau between the ordering of the truth depths and estimated depths. As mentioned in section 4.3, the kinematogram stimulus is bistable, which means it can result in two separate percepts (i.e., clockwiserotation or counterclockwise rotation). We observe switching in inferred direction and signs of the depths with our model and account for that when computing the depth inference metrics. Specifically we generate a path with the transformation defined by the inferred coefficients and observe the direction of motion of that path to determine which direction of rotation we inferred. If the inferred rotation is counterclockwise, then we multiply the depths by -1 because the rotation of the points in ground-truth sequence is clockwise.

Appendix E: Training Details

We train the transport operators using gradient descent. The operators are intialized with gaussian noise with a standard deviation of 0.3. The training run that resulted in the operators used in the kinematogram experiments used the parameters in Table 1.

We begin training at a specific learning rate and increase it if there is a successful learning step (i.e., one that decreases the learning objective) or decrease it if there is a failed learning step. While large learning rates aid in efficient gradient steps in the beginning of training, we find that decreasing the learning rate consistently toward the end of training leads to more stable final transport operator representations. Specifically, we start decreasing the learning rate at 3000 training steps and decay it by a multiplication factor of 0.9997 at each step.

Figure 15 shows the trajectories of operators learned when $M = 6$. This highlights the usefulness of the Frobenius norm regularizer on the

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

Table 1: Training Parameters for Learning Rotational Operators from 2D Projected Inputs.

Training Parameters
$M : 3$
$lr_{\text{begin}} : 0.5$
$\zeta : 0.1$
$\gamma : 0.15$
β: 10^{-4}
$\theta : 10^{\circ}$
$N_T : 20$
$N_P : 20$
Training steps: 10,000
Number of restarts for coefficient inference: 25

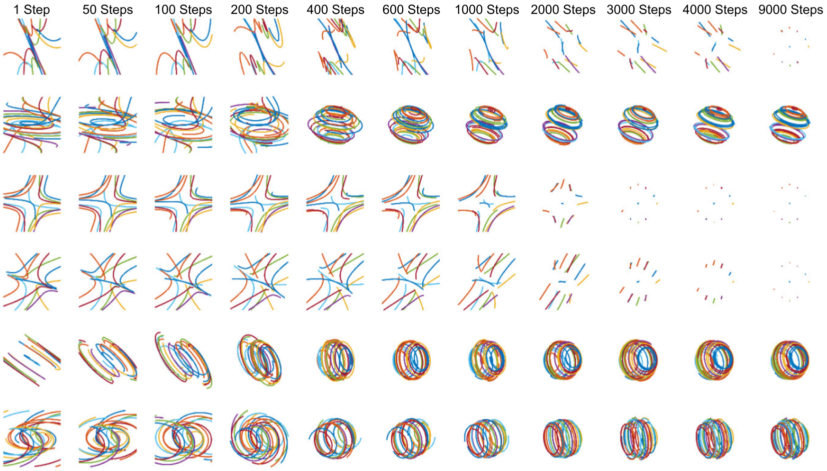


Figure 15: Transport operator trajectories during training. Each row represents one of the six learned operators. Each column shows the trajectories at a different training step. The operators begin with random initializations at step 1 and reach a rotation structure around 400 steps. After that, the three operators that do not represent rotation have their magnitudes reduced because they are not being used. At the end of training, there are three operators with clear rotational structure.

dictionary elements. If a transport operator is not being used for representing manifold paths, then its magnitude is reduced to nearly zero. Training with six operators utilizes the following parameters: $\zeta = 0.1$, $\gamma = 0.08$, $\beta = 10^{-4}$, $\theta = 10^{\circ}$, $N_T = 20$, $N_P = 20$, 10,000 training steps, and 25 random restarts.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

Table 2: Parameters for Random Dot Kinematogram Experiments.

Kinematogram Inference Parameters
ζ : 0.01
γ : 0.1
β: 0
θ : 2°
ξ : 0
N_T : 30
N_P : 20
Number of restarts for coefficient inference: 5

Appendix F: Kinematogram Experimental Details

For the kinematogram experiments, we use the base set of parameters shown in Table 2. For individual experiments, we vary subsets of the parameters from these baseline values. As with the other experiments, we correct for depth sign switching before computing the error metrics for the kinematogram tasks.

Appendix G: Kinematogram Dynamic Regularization

Our model has the capability of reducing the amount of depth sign switching and transformation direction switching by adding a dynamic regularizer during inference. We assume that there is a constant speed of rotation and encourage the transport operator coefficients to be similar from one frame to the next. Note that we already utilize this coefficient consistency in the inference procedure by inferring the same coefficients for all frames in the inference window. The additional regularizer encourages the set of coefficients to be similar from one inference window to the next. This amounts to adding a term to the inference objective that incorporates the previous coefficient estimate:

$$\begin{aligned} L = {} & \frac{1}{2N_T}\sum_{n=1}^{N_T}\sum_{i=1}^{N_P}\left[\|\mathbf{y}^{(i)}_{N_T-n} - \mathbf{KT}(n\mathbf{c})\widehat{\mathbf{x}}^{(i)}_{N_T}(\boldsymbol{\lambda})\|_2^2\right] + \zeta\|\mathbf{c}\|_1 \\ & + \frac{\beta}{2}\|\boldsymbol{\lambda}\|_2^2 + \frac{\xi}{2}\|\mathbf{c}-\mathbf{c}_{\text{prev}}\|_2^2. \end{aligned} \tag{G.1}$$

With the addition of the dynamic regularizer on the coefficients, we infer smooth depth estimates for kinematogram sequences. Figure 16 shows the estimated depths and depth error with dynamic regularization. This regularization eliminates sign-flipping, but it also affects the depth error in the beginning frames of the kinematogram sequence. The initial depth

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

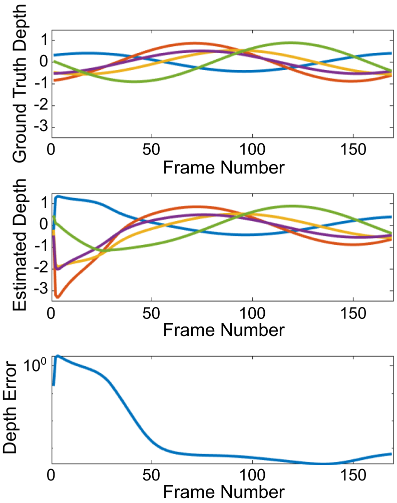
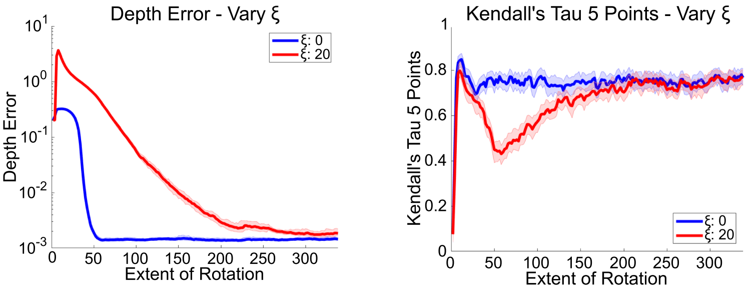


Figure 16: Example of depths inferred for random dots in a kinematogram sequence with a dynamic regularizer. The dynamic regularizer removes the sign switching of the depth values observed in Figure 10, but it also leads to larger error early in the kinematogram sequence.



(a) (b)

Figure 17: Quantitative metrics for random dot kinematogram depth estimates with and without dynamic regularization.

estimates are less accurate because they use only a small window of frames for depth inference, and the coefficient regularizer encourages coefficient estimates in later frames to be similar to the initial inaccurate estimates. The depth estimates eventually achieve low error as the inference window gets larger.

The plots in Figure 17 compute the depth MSE for all N_P points and Kendall's tau for five randomly selected dots in the stimulus. Figures 17a

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

and 17b each show the impact of including the dynamic regularizer from equation G.1. When $\xi = 0$, the median depth error looks the same as in Figure 10, where it begins high due to the limited angular extent of rotation and decreases as the kinematogram sequence continues. The Kendall's tau values with both $\xi = 0$ and $\xi = 20$ have a spike at the beginning of the kinematogram sequence. Without the dynamic regularizer, the Kendall's tau value plateaus and remains around the same value until the end of the sequence. As we saw in Figure 16, the integration of the dynamic regularizer on the coefficients leads to a delay in achieving optimal accuracy for depth estimates measured by both the depth MSE and Kendall's tau.

References

Agudo, A., Pijoan, M., & Moreno-Noguer, F. (2018). Image collection pop-up: 3D reconstruction and clustering of rigid and non-rigid categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2607–2615).

Andersen, R. A., & Bradley, D. C. (1998). Perception of three-dimensional structure from motion. *Trends in Cognitive Sciences*, *2*(6), 222–228. 10.1016/S1364-6613(98)01181-4

Braunstein, M. L. (2014). *Depth perception through motion*. Academic Press.

Braunstein, M. L., Hoffman, D. D., Shapiro, L. R., Andersen, G. J., & Bennett, B. M. (1987). Minimum points and views for the recovery of three-dimensional structure. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(3), 335. 10.1037/0096-1523.13.3.335

Bregler, C., Hertzmann, A., & Biermann, H. (2000). Recovering non-rigid 3D shape from image streams. In *Proceedings. of the IEEE Conference on Computer Vision and Pattern Recognition.* (vol. 2, pp. 690–696).

Cohen, T., & Welling, M. (2014). Learning the irreducible representations of commutative Lie groups. In *Proceedings of the International Conference on Machine Learning* (pp. 1755–1763).

Connor, M., Canal, G., & Rozell, C. (2021). Variational autoencoder with learned latent structure. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* (pp. 2359–2367).

Connor, M., Fallah, K., & Rozell, C. (2021). Learning identity-preserving transformations on data manifolds. *Transactions on Machine Learning Research*, *2023*.

Connor, M., & Rozell, C. (2020). Representing closed transformation paths in encoded network latent space. In *Proceedings of the AAAI Conference on Artificial Intelligence* (vol. 34, pp. 3666–3675). 10.1609/aaai.v34i04.5775

Cooper, L. A., & Shepard, R. N. (1973). Chronometric studies of the rotation of mental images. In W. G. Chase (Ed.), *Visual information processing* (pp. 75–176). Elsevier.

Culpepper, B. J., & Olshausen, B. A. (2009). Learning transport operators for image manifolds. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems, 23* (pp. 423–431). Curran.

Dai, Y., Li, H., & He, M. (2014). A simple prior-free method for non-rigid structure-from-motion factorization. *International Journal of Computer Vision*, *107*, 101–122. 10.1007/s11263-013-0684-2

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. 10.1016/j.tics.2007.06.010

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. 10.1016/j.neuron.2012.01.010

Dosher, B. A., Landy, M. S., & Sperling, G. (1989). Ratings of kinetic depth in multidot displays. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 816. 10.1037/0096-1523.15.4.816

Eigen, D., Puhrsch, C., & Fergus, R. (2014). *Depth map prediction from a single image using a multi-scale deep network*. arXiv:1406.2283.

Fefferman, C., Mitter, S., & Narayanan, H. (2016). Testing the manifold hypothesis. *Journal of the American Mathematical Society*, *29*(4), 983–1049. 10.1090/jams/852

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395. 10.1145/358669.358692

Fukumi, M., Omatu, S., & Nishikawa, Y. (1997). Rotation-invariant neural pattern recognition system estimating a rotation angle. *IEEE Transactions on Neural Networks*, *8*(3), 568–581. 10.1109/72.572096

Funt, B. V. (1983). A parallel-process model of mental rotation. *Cognitive Science*, *7*(1), 67–93.

Garg, R., Bg, V. K., Carneiro, G., & Reid, I. (2016). Unsupervised CNN for single view depth estimation: Geometry to the rescue. In *Proceedings of the European Conference on Computer Vision* (pp. 740–756).

Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 270–279).

Hall, B. (2015). *Lie groups, Lie algebras, and representations: An elementary introduction*. Springer.

Hartley, R. I. (1997). In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*(6), 580–593. 10.1109/34.601246

Hartley, R. I., & Sturm, P. (1997). Triangulation. *Computer Vision and Image Understanding*, *68*(2), 146–157. 10.1006/cviu.1997.0547

Hartley, R. I., & Zisserman, A. (2004). *Multiple view geometry in computer vision* (2nd ed.). Cambridge University Press.

Hauberg, S., Freifeld, O., Larsen, A. B. L., Fisher, J., & Hansen, L. (2016). Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In *Proceedings of the International Conference onArtificial Intelligence and Statistics* (pp. 342–350).

Hildreth, E. C., Grzywacz, N. M., Adelson, E. H., & Inada, V. K. (1990). The perceptual buildup of three-dimensional structure from motion. *Perception and Psychophysics*, *48*(1), 19–36. 10.3758/BF03205008

Inui, T., & Ashizawa, M. (2011). Temporo-parietal network model for 3D mental rotation. In Y. Ysamaguchi (Ed.), *Advances in cognitive neurodynamics (II)* (pp. 91–95). Springer.

Just, M. A., & Carpenter, P. A. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, *92*(2), 137. 10.1037/0033-295X.92.2.137

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

Kanade, T., & Morris, D. D. (1998). Factorization methods for structure from motion. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *356*(1740), 1153–1173. 10.1098/rsta.1998.0215

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1/2), 81–93. 10.1093/biomet/30.1-2.81

Kong, C., & Lucey, S. (2016). Prior-less compressible structure from motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4123–4131).

Kong, C., & Lucey, S. (2019). Deep non-rigid structure from motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1558–1567).

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, *2*(1–2), 83–97. 10.1002/nav.3800020109

Ladicky, L., Shi, J., & Pollefeys, M. (2014). Pulling things out of perspective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 89–96).

Lamm, C., Windischberger, C., Moser, E., & Bauer, H. (2007). The functional role of dorso-lateral premotor cortex during mental rotation: An event-related fMRI study separating cognitive processing steps using a novel task paradigm. *NeuroImage*, *36*(4), 1374–1386. 10.1016/j.neuroimage.2007.04.012

Liu, F., Shen, C., Lin, G., & Reid, I. (2015). Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(10), 2024–2039. 10.1109/TPAMI.2015.2505283

Longuet-Higgins, H. C. (1981). A computer algorithm for reconstructing a scene from two projections. *Nature*, *293*(5828), 133–135. 10.1038/293133a0

Luo, W., Schwing, A. G., & Urtasun, R. (2016). Efficient deep learning for stereo matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5695–5703).

Mather, G. (1989). Early motion processes and the kinetic depth effect. *Quarterly Journal of Experimental Psychology Section A*, *41*(1), 183–198. 10.1080/14640748908402359

Miao, X., & Rao, R. P. (2007). Learning the Lie groups of visual invariance. *Neural Computation*, *19*(10), 2665–2693. 10.1162/neco.2007.19.10.2665

Nistér, D. (2005). Preemptive RANSAC for live structure and motion estimation. *Machine Vision and Applications*, *16*(5), 321–329.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325. 10.1016/S0042-6989(97)00169-7

Petersik, J. T. (1979). Three-dimensional object constancy: Coherence of a simulated rotating sphere in noise. *Perception and Psychophysics*, *25*(4), 328–335. 10.3758/BF03198812

Pollefeys, M., Nistér, D., Frahm, J.-M., Akbarzadeh, A., Mordohai, P., Clipp, B., . . . Towles, H. (2008). Detailed real-time urban 3D reconstruction from video. *International Journal of Computer Vision*, *78*(2), 143–167. 10.1007/s11263-007-0086-4

Rao, R. P., & Ruderman, D. L. (1999). Learning Lie groups for invariant visual perception. In S. Solla, T. Leen, & K. Müller (Eds.), *Advances in neural information processing systems*, *12* (pp. 810–816). Curran.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

Reichelt, S., Häussler, R., Fütterer, G., & Leister, N. (2010). Depth cues in human visual perception and their realization in 3D displays. In B. Javidi & J.-Y. Son (Eds.), *Three-dimensional imaging, visualization, and display 2010 and display technologies and applications for defense, security, and avionics IV* (Vol. 7690, p. 76900B). SPIE.

Seepanomwan, K., Caligiore, D., Cangelosi, A., & Baldassarre, G. (2015). Generalisation, decision making, and embodiment effects in mental rotation: A neurorobotic architecture tested with a humanoid robot. *Neural Networks*, *72*, 31–47. 10.1016/j.neunet.2015.09.010

Shepard, R. N., & Cooper, L. A. (1986). *Mental images and their transformations.* MIT Press.

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

Snavely, N., Seitz, S. M., & Szeliski, R. (2006). Photo tourism: Exploring photo collections in 3D. *ACM Transactions on Graphics*, *25*(3), 835–846. 10.1145/1141911.1141964

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

Sperling, G., Landy, M. S., Dosher, B. A., & Perkins, M. E. (1989). Kinetic depth effect and identification of shape. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(4), 826. 10.1037/0096-1523.15.4.826

Todd, J. T., Akerstrom, R. A., Reichel, F. D., & Hayes, W. (1988). Apparent rotation in three-dimensional space: Effects of temporal, spatial, and structural factors. *Perception and Psychophysics*, *43*(2), 179–188. 10.3758/BF03214196

Tomasi, C., & Kanade, T. (1992). Shape and motion from image streams under orthography: A factorization method. *International Journal of Computer Vision*, *9*(2), 137–154. 10.1007/BF00129684

Tsai, R. Y., & Huang, T. S. (1984). Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *6*(1), 13–27. 10.1109/TPAMI.1984.4767471

Ullman, S. (1979). The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *203*(1153), 405–426.

Wallach, H., & O'Connell, D. (1953). The kinetic depth effect. *Journal of Experimental Psychology*, *45*(4), 205. 10.1037/h0056880

Wiskott, L., & Sejnowski, T. J. (2002). Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, *14*(4), 715–770. 10.1162/089976602317318938

Xie, J., Girshick, R., & Farhadi, A. (2016). Deep3D: Fully automatic 2D-to-3D video conversion with deep convolutional neural networks. In *Proceedings of the European Conference on Computer Vision* (pp. 842–857).

Zacks, J. M., & Michelon, P. (2005). Transformations of visuospatial images. *Behavioral and Cognitive Neuroscience Reviews*, *4*(2), 96–118. 10.1177/1534582305281085

Zbontar, J., & LeCun, Y. (2016). Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine learning Research*, *17*(1), 2287–2318.

Zhou, T., Brown, M., Snavely, N., & Lowe, D. G. (2017). Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1851–1858).

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024

Zhu, Y., Huang, D., De La Torre, F., & Lucey, S. (2014). Complex non-rigid motion 3D reconstruction by union of subspaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1542–1549).

Received March 14, 2023; accepted March 6, 2024.

Downloaded from http://direct.mit.edu/neco/article-pdf/36/11/2505/2474305/neco_a_01695.pdf by Ramona Marchand on 01 November 2024