Multiscale Pixel Spatiotemporal Information Flows

Felix Y. Zhou
Lyda Hill Department of Bioinformatics
UT Southwestern Medical Center
Dallas, TX 75390
felix.zhou@UTSouthwestern.edu

Roshan Ravishankar
Lyda Hill Department of Bioinformatics
UT Southwestern Medical Center
Dallas, TX 75390
roshan.ravishankar@UTSouthwestern.edu

Abstract

We develop a formal algorithmic framework to compute *multiscale pixel spatiotemporal information flows* which capture, in an unbiased manner, salient causal relationships between pixels across space and time. Real spatiotemporal dynamical systems such as cellular morphodynamics are inherently complex, nonlinear and evolve over time in response to feedbacks. This makes it highly challenging to directly model, simulate, or fit observed phenomena from first principle physics. Oftentimes neither the salient variables nor the key relationships are known *a priori* to include in a mathematical model. Even if a model was possible, we may be limited in our ability to sample the necessary information for exact system identification and verification. Alternatively, causal measures have been developed to identify potential causal relationships statistically from only observational timeseries. However such measures have largely only been studied for unstructured 1D timeseries where objects-of-interest have been pre-segmented and tracked over time. This restricts their application either to analyse general video dynamics, where individual objects are impossible to define or difficult to segment, or to understand potential causal relationships between subparts of objects. Here we propose a formal definition of a *pixel spatiotemporal information flow* as a spatiotemporal derivative of a pixel intensity timeseries to extract the dense pixel-to-pixel information transfer in 2D + time videos using any desired 1D causal measure, in a general and multiscale manner. Applying our framework, we discover salient pixel-to-pixel information highways in videos of diverse phenomena spanning traffic and crowd flow, collision physics, fish swarming, moving camouflaged animals, human action, embryo development, cell division and cell migration.

1 Introduction

This work considers the problem of extracting salient causal dynamic patterns from video. Specifically, we are interested in capturing whether an individual pixel's intensity changes over a given time interval causally influences another pixel's intensity changes at a later timepoint.

Capturing semantically meaningful video dynamics over arbitrary time intervals is challenging. In general, videos are capture the temporal evolution of real spatiotemporal dynamical systems which are inherently complex, nonlinear, and nonstationary. Moreover videos may be affected by acquisition artifacts such as camera shake. Consequently videos cannot generally be described by first principles physical differential equations. Over short-time intervals, $[t, t + \delta t]$, e.g. frame-to-frame changes, optical flow is the most popular approach in computer vision to capture the potential influence of individual pixels on its neighbors through intensity variation. Many methods of optical flow estimation have been proposed including block-based matching (Liu and Delbruck [2018]), gradient based methods (Farnebäck [2003]) and learning based methods (Dosovitskiy et al. [2015]). For longer-

A causal view on dynamical systems workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

time intervals over many frames, existing works have explored using the short time dynamics captured by optical flow as a supplementary input to deep learning architectures to improve performance in applications such as action recognition (Sun et al. [2018]) and object segmentation (Lamdouar et al. [2020]). However deep learning training is expensive and limited by the availability of appropriate training data. In addition, for novel scientific discovery in fields such as biology, labelled data are often too expensive to acquire, ambiguous or outright impossible to annotate. Thus optical flow and statistical modelling remains an important tool for applications such as cellular tracking and inference of cell-cell interaction networks (Zhou et al. [2019]).

Classically, optical flow computes for every pixel the physical displacement $(\Delta x, \Delta y)$ to a pixel position (x, y) in the current video frame at time t such that the pixel intensity $I(x+\Delta x, y+\Delta y, t) \approx I(x, y, t+1)$, is preserved in the next video frame at time $t+1$ and pixel position (x, y). The $(\Delta x, \Delta y)$ optical flow displacement thus represent a pixel-wise 'causal information flow' where the magnitude measures the 'strength' of interaction between the pixel at (x, y) and neighboring pixels in the $(\Delta x, \Delta y)$ direction. However the physical displacement of individual pixels between two timepoints is only one process contributing to all possible 'causal information flows' within a video. In particular, displacement captures only the egocentric motion of single actors and misses communal interactions between several actors within a video such as: changes in the walking of pedestrians to avoid other pedestrians and traffic; the changing traffic on highways; the physical collisions between snooker balls; the influence of cellular crowding in confluent tissue; and the coordination between individual body parts to execute a stereotypic action. Importantly these interactions are complex; a temporal delay between interacting pixels requires observation over multiple frames. In addition, interactions may be transient or sporadic. In all such cases their detection over a given time interval can easily be under- or over-estimated even when the frame-to-frame optical flow is temporally averaged.

Alternatively, statistical causal measures for measuring the information transfer between 1-dimensional (1D) timeseries have been developed and extensively studied in scientific disciplines such as econometrics (Granger [1969]), ecology (Sugihara et al. [2012], Detto et al. [2012]) and neuroscience (Seth et al. [2015]). Unlike optical flow, these measures do not require a physical notion of information. Instead a score of causal information is provided by assaying the extent of co-fluctuation between individual timeseries when observed over a given window of time. A plethora of causal measures have been developed and they can generally be classified into two categories: functional and effective connectivity. Functional connectivity defines causal information by measuring observable statistical dependencies among timeseries and includes well-known families of methods such as correlation, coherence, Granger causality, and transfer entropy (Bastos and Schoffelen [2016]). Effective connectivity defines causal information based on the influence on activity, and depends on an explicit parametric model of the interaction. Examples include structural causal models and dynamic differential covariance (Valdes-Sosa et al. [2011], Chen et al. [2022]).

In this paper, we investigate whether by considering each pixel as a 1D timeseries and using existing 1D causal measures can better discover the complex salient dynamic patterns over arbitrary time intervals of any video with minimal prior assumptions. To do so, we developed a generally applicable and multiscale algorithmic framework, *pixel spatiotemporal information flows* which formally defines how to adapt 1D functional and effective connectivity measures to compute meaningful information flows at individual pixels in real 2D videos. We note this is not as trivial as one might think. Images are structured data such that individual pixels are intrinsically correlated and causally affect any dynamics in their immediate surrounding pixels! As it was unclear to us which 1D causal measures, if any, would be optimal, we operationalised three 1D causal measures, two examples of functional connectivity (maximum cross-correlation, conditional Granger causality) and one example of effective connectivity (dynamic differential covariance). We then applied these pixel spatiotemporal information flows together with the only literature example of a pixel-based information flow we could find, probabilistic canonical correlation analysis (PCCA) (Yamashita et al. [2012]) to a broad variety of real world datasets to discover salient pixel-wise information highways and information sources/sinks. First, we uncover dynamic patterns in a crowd flow segmentation dataset (Ali and Shah [2007]) to qualitatively and quantitatively compare and highlight differences to mean optical flow. Second, we demonstrate the generality of information flows with diverse examples of video dynamical systems from action recognition (Soomro et al. [2012]), collision physics (Yi et al. [2018]), developmental biology (Tomer et al. [2012]), and cell migration (Ulman et al. [2017]). Finally, we compared the ability of optical and information flows on the challenging task of segmenting moving camouflaged animals in the wild (Lamdouar et al. [2020]). We find pixel spatiotemporal information

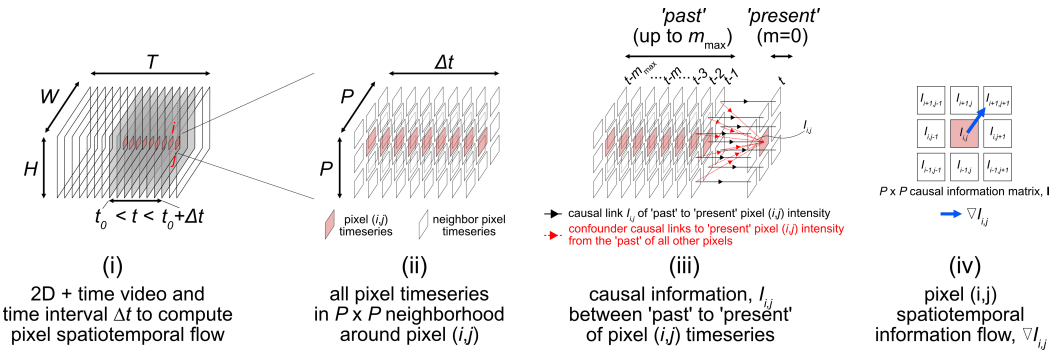

Figure 1: Schematic illustrating the four key steps and mathematical notation used in the definition and computation of a pixel spatiotemporal information flow. For more details see **section 3**.

flows are superior than optical flow in every application where the physical pixel displacement is not explicitly necessary by better leveraging any temporal cues within the given time interval.

2 Related Works

We briefly review existing literature that attempt to discover causal relationships in 2D videos.

Pixel-wise Causal information flow as an alternative to dense optical flow. Yamashita et al. [2012] first introduced the notion of a pixel-based causal flow analogous to dense optical flow for 2D videos and is still the only literature example we could find. They exclusively considered the extension of Granger causality, a functional connectivity measure to the case of a RGB pixel by showing an equivalence to probabilistic canonical correlation analysis (PCCA). Only qualitative comparison to Horn-Schunk optical flow was shown for a few selected examples of the same crowd flow segmentation dataset (Ali and Shah [2007]) we use and only at a single spatial scale. It is thus unclear whether PCCA flow is simply an improved optical flow, whether it can discover dynamic patterns dense optical flow could not, and if other causal measures behave the same or could offer additional insights. Addressing these outstanding important questions was the main technical motivation for this paper. Importantly, PCCA specifically only operationalises Granger causality. A formal algorithmic framework was therefore needed to operationalise other 1D causal measures into its equivalent pixel-based causal flow.

Causal inference of object relationships in videos. Much literature exists related to the inference of causal relationships between 'objects' in videos. These 'objects' are any predefined discrete entity including keypoints, segmentated objects or superpixels - any representation that is not individual pixels. 1D timeseries of any property of these objects are then extracted after temporal tracking. These timeseries can then simply be treated with existing methods for the desired 1D causal measure - typically this is just Granger causality (Narayan and Ramakrishnan [2014], Prabhakar et al. [2010], Swears et al. [2014]). These methods can identify potential causal relationships in video, however their dependence on knowing *a priori* the exact objects of interest means they are unsuited to general application to discover potentially novel causal information flows. Any causal relationships inferred are also clearly limited at the level of object specification. For example if only entire humans are segmented, we cannot infer anything about the coordinated movement of joints in an action recognition video.

Inference of functional connectivity networks Notions of functional connectivity are popular in biology, including in the fields of neuroscience and cell biology where the exact causality is unknown and the system cannot be experimentally perturbed. As in the machine learning literature, works in this area focus on finding potential functional connectivity networks between discrete, predefined entities such as physical partitions of the brain (Bastos and Schoffelen [2016], Noble et al. [2019]) or between segmented cells (Zamir et al. [2022]) and superpixels (Noh et al. [2022]). Here again any 1D causal measures can be directly applied to any extracted average 1D timeseries of object properties.

3 Methods

3.1 Pixel spatiotemporal information flow

Given a grayscale input video, $X \in \mathbb{R}^{H\times W\times T}$ with each video frame of size $H \times W$ pixels, and T total frames, let $X_{i,j,t}$ denote the intensity timeseries of the pixel at position $(i, j) \in [0, H-1] \times [0, W-1]$ over a time interval Δt with start time t_0 i.e. $t \in [t_0, t_0 + \Delta t]$. The pixel spatiotemporal information flow at the pixel position (i, j) is a 2-vector and the spatial gradient $\nabla I_{i,j}$ of a scalar-valued 'causal information' measure, $I_{i,j}$. Thus an input video $X \in \mathbb{R}^{H\times W\times T}$ yields a pixel spatiotemporal information flow, $\nabla I \in \mathbb{R}^{H\times W\times 2}$. We illustrate the concept schematically with the same mathematical notation in Figure 1.

The spatial gradient, $\nabla I_{i,j}$ is evaluated considering the set of all pixel timeseries within a $P \times P$ pixel neighborhood centered around (i, j), i.e. $\mathbf{X}_{i,j,t} = \{X_{(i+\delta_i,j+\delta_j,t)} | \delta_i = -P/2, ..., P/2 \text{ and } \delta_j = -P/2, ..., P/2\}$. Note that each pixel timeseries is a 1D timeseries with respect to time. For each pixel timeseries in the spatial neighborhood we compute the causal information $I_{i,j}$ as the causal influence of its **own** past intensity values to its **present** intensity values, conditioning out any influence from **all other** pixel timeseries according to the chosen 1D causal measure of interest (see **section 3.2** for mathematical details of those measures implemented in this study). This produces a $P \times P$ matrix of causal information values $\mathbf{I} = \{I_{(i+\delta_i,j+\delta_j)} | \delta_i = -P/2, ..., P/2 \text{ and } \delta_j = -P/2, ..., P/2\}$. The spatiotemporal information flow, $\nabla I_{i,j}$ is then computed as the total sum of causal information in $\mathbf{I}$ multiplied by an orientation vector, $O_{i,j}$ with the following $x-$, $y-$ components:

$$O_{(i,j),x} = (-1) \sum_{\delta_j=1}^{P/2} I_{i,j-\delta_j} + \sum_{\delta_j=1}^{P/2} I_{i,j+\delta_j} \tag{1}$$

$$O_{(i,j),y} = (-1) \sum_{\delta_i=1}^{P/2} I_{i-\delta_i,j} + \sum_{\delta_i=1}^{P/2} I_{i+\delta_i,j} \tag{2}$$

$$\nabla I_{i,j} = \sum_{\delta_i,\delta_j} I_{i+\delta_i,j+\delta_j} \sqrt{O^2_{(i,j),x} + O^2_{(i,j),y}} \tag{3}$$

This definition following (Yamashita et al. [2012]) allows us to quantify the mean spatial direction of causal information across the central (i, j) pixel weighted by the total causal influence in the pixel neighborhood. Note in Eqn.1 and 2 in contrast to (Yamashita et al. [2012]), the pixel offset indices are all positive in the summation symbol as we incorporate the sign in the summand.

In practice, to reduce computational time, we only compute $\nabla I_{i,j}$ for all non-overlapping $P \times P$ neighborhoods to generate ∇I for the input video, and use linear interpolation to resize the resulting vector field to the size of the video frame i.e. $\nabla I \in \mathbb{R}^{H\times W\times 2}$. Color RGB videos are first converted to grayscale videos for computation.

3.2 Implemented causal measures as information flows

We implement three literature causal information measures as pixel spatiotemporal information flows to supplement the PCCA flow of (Yamashita et al. [2012]) using the same mathematical notation introduced in **section 3.1** as described below:

Maximum Cross-Correlation (max. CC). Cross-correlation measures the similarity between two timeseries over different temporal lags. The peak position offset of the maximum spatial cross-correlation between a pixel neighborhood at time t and the same neighborhood in the successive frame at time $t + \delta t$ is commonly used to extract dense optical flow in fluid mechanics and biology where it is more commonly referred to as particle image velocimetry (Adrian [2005]). Here we extend correlation into a causal measure for pixel timeseries. Consider $\mathbf{X}_{i,j,t}$, the set of all pixel timeseries in a $P \times P$ neighborhood around (i, j) over a time interval $[t_0, t_0 + \Delta t]$. Construct from $\mathbf{X}_{i,j,t}$, the m-frame lagged signal denoted $\mathbf{X}_{i,j,t} = \{X_{(i+\delta_i,j+\delta_j,t)} | t \in [t_0, t_0 + \Delta t - m]\}$ as the 'past' timeseries of each pixel and the corresponding 'present' timeseries of each pixel and of equal temporal duration, $\mathbf{X}_{i,j,t} = \{X_{(i+\delta_i,j+\delta_j,t)} | t \in [t_0 + m, t_0 + \Delta t]\}$. We then define the $P \times P$ max.

CC causal information matrix, $\mathbf{I}_{CC}$ as the maximum 3D (space + time) correlation value over time,

$$\mathbf{I}_{CC} = \max_t CC(t) \in \mathbb{R}^{P \times P} \tag{4}$$

and

$$CC(t) = \sum_{\delta'_i, \delta'_j, \delta'_t} \mathbf{X}_{i+\delta i', j+\delta j', t-m+\delta t'} \mathbf{X}_{i,j,t}, \in \mathbb{R}^{P \times P \times (\Delta t - m)} \tag{5}$$

computed over spatial and temporal lags using 'same' padding. Note $\mathbf{I}_{CC}$ is only valid when $m > 0$. When $m = 0$, $\mathbf{I}_{CC}$ will be the identity matrix as every pixel is maximally correlated with itself and the pixel spatiotemporal information flow will be 0 everywhere.

Conditional Granger causality (cGC). Granger causality (GC) establishes that a variable, x is 'causal' to another variable y if the additional inclusion of the history of x reduces the uncertainty in the prediction of future values of y. First conceptualized by (Wiener [1956]) GC was operationalized using vector autoregressive (VAR) models by (Granger [1969]). Later, (Geweke [1984]) extended the VAR model to condition out the confounding effect of a third variable z which x and y may both depend upon. The modified method is called conditional Granger causality (cGC). Due to the conceptual simplicity, GC variants are the most popular class of causal measures for testing temporal precedence amongst 1D timeseries (Bressler and Seth [2011]). To adapt the cGC setup to pixel timeseries, consider $\mathbf{X}_{i,j}$, the set of all pixel timeseries in a $P \times P$ neighborhood around (i, j) over a time interval $[t_0, t_0 + \Delta t]$. We will test for each pixel timeseries the reduction in fitting error when predicting the 'present' timeseries when including (full model) or excluding (reduced model) the 'present' timeseries in addition to all 'past' timeseries up to a maximum lag of $m_{\max}$ frames as independent regression variables. Mathematically, construct from $\mathbf{X}_{i,j,t}$, the m-frame lagged 'past' timeseries denoted $\mathbf{X}_{i,j,t-m} = \{X_{(i+\delta_i, j+\delta_j, t)} | t \in [t_0 + m_{\max} - m, t_0 - m + \Delta t]\}$, and the 'present' timeseries which we wish to predict denoted $\mathbf{X}_{i,j,t} = \{X_{(i+\delta_i, j+\delta_j, t)} | t \in [t_0 + m_{max}, t_0 + \Delta t]\}$. The pair of full and reduced VAR regression models in matrix notation for all pixel timeseries is then

$$\mathbf{X}_{i,j,t} = \mathbf{A}_{m=0} \mathbf{X}_{i,j,t} + \sum_{m=1,\ldots,m_{\max}} \mathbf{B}_m \mathbf{X}_{i,j,t-m} + \mathbf{E}^{full}_{i,j,t} \tag{6}$$

$$\mathbf{X}_{i,j,t} = \sum_{m=1,\ldots,m_{\max}} \mathbf{C}_m \mathbf{X}_{i,j,t-m} + \mathbf{E}^{reduced}_{i,j,t} \tag{7}$$

where **A**, **B**, **C** denote respective submatrices of the constant coefficients fitted by least squares ridge regression and **E** the residual error timeseries. The $P \times P$ cGC causal information matrix, $\mathbf{I}_{cGC}$ is computed following (Geweke [1984]) as the log variance ratio of reduced and full residual errors.

$$\mathbf{I}_{cGC} = \ln \frac{\text{var}(\mathbf{E}^{reduced}_{i,j,t})}{\text{var}(\mathbf{E}^{full}_{i,j,t})} \in \mathbb{R}^{P \times P} \text{ (after reshaping)} \tag{8}$$

where the variance operation is over time.

Dynamic differential covariance (DDC). The VAR models that operationalise Granger causality can be seen as modelling a set of timeseries as a state-space dynamical system (Barnett and Seth [2015]). Instead of VAR which requires specifying a maximum lag, (Chen et al. [2022]) proposed a linear ordinary differential equation (ODE) dynamical system could compute causality. They showed that the least-squares fitted ODE model coefficients are a measure of causal effect between every pair of 1D timeseries conditional on all other timeseries. They termed this measure dynamic differential covariance (DDC). We adapt DDC to consider $\mathbf{X}_{i,j,t}$, all pixel timeseries in a $P \times P$ neighborhood around (i, j) over a time interval $[t_0, t_0 + \Delta t]$. The $P \times P$ DDC causal information matrix, $\mathbf{I}_{DDC}$ is defined as the matrix formed by retrieving just the causality measure between desired timeseries pairs in the full pairwise DDC computed according to (Chen et al. [2022]):

$$\frac{d\mathbf{X}_{i,j,t}}{dt} = \mathbf{W} \mathbf{X}_{i,j,t} \quad \text{and} \quad DDC = \mathbf{W} \in \mathbb{R}^{P^2 \times P^2} \tag{9}$$

$$\mathbf{I}_{DDC} = \{W_{k,k} | k = 1, ..., P^2\} \in \mathbb{R}^{P \times P} \text{ (after reshaping)} \tag{10}$$

where $d\mathbf{X}_{i,j,t}/dt = \{dX_{(i+\delta_i, j+\delta_j, t)}/dt\}$ is the time derivative of the timeseries, $\mathbf{W}$ is the constant coefficient matrix estimated using ridge regression and k is the matrix indices of $\mathbf{W}$.

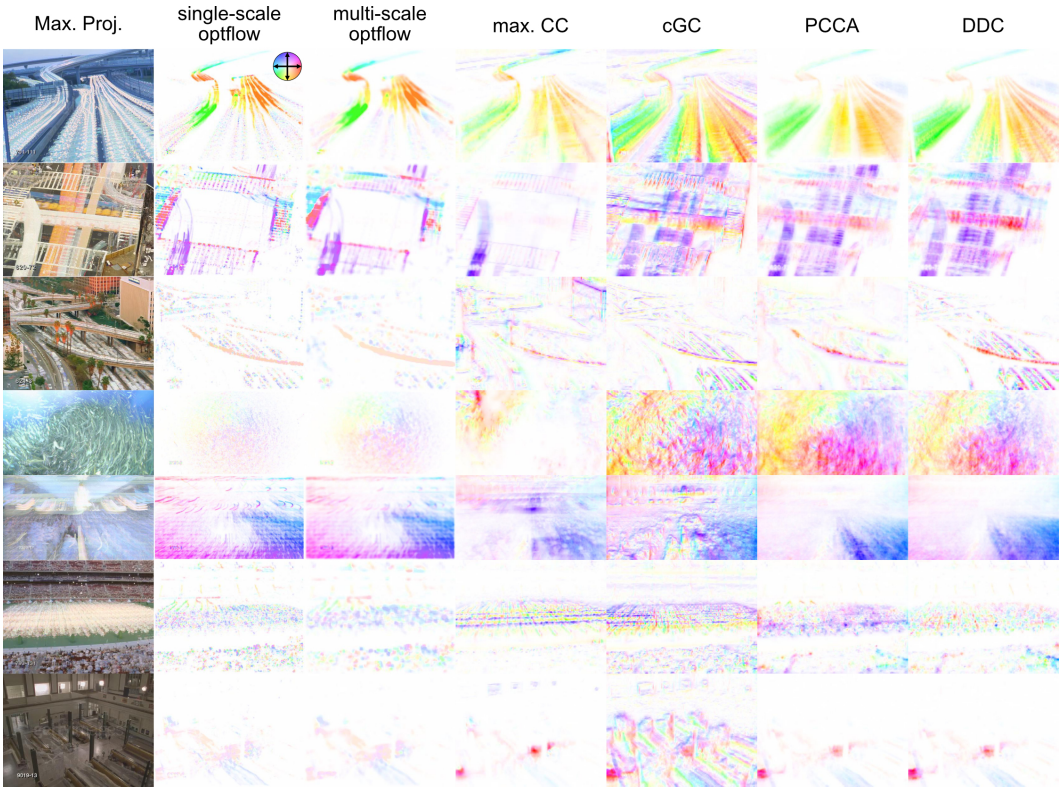

Figure 2: Comparison of the mean single- and multi-scale optical flow with four spatiotemporal information flows. **Row 1:** highway traffic. **Row 2:** pedestrian and motor traffic at a 4-way junction. **Row 3:** Multistory, multi-directional highway traffic. **Row 4:** Anti-clockwise rotating fish swarm. **Row 5:** two way crowd movement indoors. **Row 6:** cheerleading performance in a sports arena. **Row 7:** visitors in an art gallery. Direction of flow is colored per the color wheel

Probabilistic Canonical Correlation Analysis (PCCA). We briefly describe PCCA and defer the reader to (Yamashita et al. [2012]) for mathematical details. (Yamashita et al. [2012]) considered how to apply Granger causality, defined originally for the univariate case, to compute a causal measure between two pixels when each pixel is described by a multivariate timeseries (e.g. in a color image, a pixel is described by three timeseries; red, green and blue). They showed that correcting for the redundant intercorrelations between the multivariate features was the same as using PCCA (Fujita et al. [2009]). This modification allowed them to consider further block causal flow where the (i, j) pixel is replaced by a $P \times P$ pixel block centered at (i, j). Neighboring pixels are analogously replaced by neighboring $P \times P$ pixel blocks. The information flow computation is the same as Eqn.3. We note PCCA is undefined for univariate grayscale images. For comparison here for grayscale video we use the $P \times P$ pixel block definition for PCCA pixel spatiotemporal information flow.

3.3 Multiscale pixel spatiotemporal information flow

The causal information flow between pixels is dependent in part on the spatial separation of individual interacting objects in the video. Consequently a prespecified $P \times P$ neighborhood may be too small or too large a window to capture all potential causal relationships. Therefore we propose to compute the pixel spatiotemporal information flow in a multiscale manner. We note it is not fully clear how exactly to do this (Valdes-Sosa et al. [2011]). Unlike optical flow which yields a physical displacement vector, information flow vectors do not have physical meaning. In favor of keeping things simple we implemented multiscale flow using the same Gaussian image pyramids as used in multiscale optical flow schemes and an average pooling scheme to combine the flows from different scales. Given a grayscale input video sequence, $X \in \mathbb{R}^{H \times W \times T}$ with a frame size of $H \times W$ pixels, T total frames, and an array of N desired downsampling factors, $s = \{s_1, s_2, ..., s_N\}$ of increasing magnitude, compute the dowsampled video sequences with linear

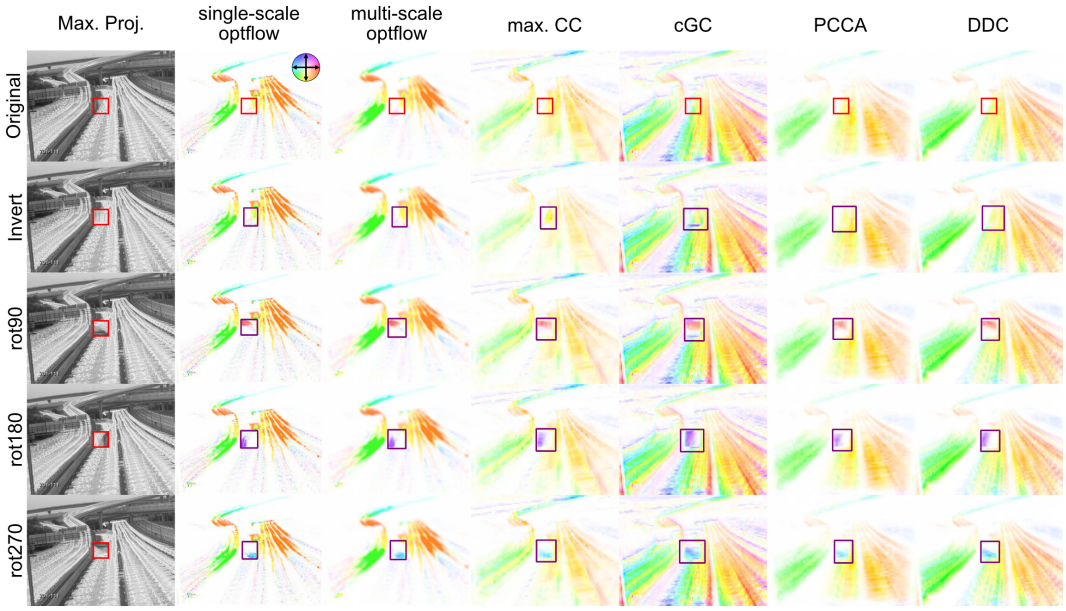

Figure 3: Perturbations of a traffic flow by inverting or rotating the pixel intensity inside a fixed region of interest (red box), 50x50 pixels in size and the computed flows and their predicted perturbed region (purple box). Direction of flow is colored per the color wheel

interpolation, $\mathbf{X}_s = \{\sigma_1(X_{s_1}) \in \mathbb{R}^{H/s_1 \times W/s_1 \times T}, \sigma_2(X_{s_2}) \in \mathbb{R}^{H/s_2 \times W/s_2 \times T}, ..., \sigma_N(X_{s_N}) \in \mathbb{R}^{H/s_N \times W/s_N \times T}\}$ where $\sigma(\cdot)$ denotes an optional isotropic Gaussian smoothing with the specified smoothing σ. For each video sequence compute the respective information flow fields, $\nabla I = \{\nabla I_{s_1} \in \mathbb{R}^{H/s_1 \times W/s_1 \times 2}, \nabla I_{s_2} \in \mathbb{R}^{H/s_2 \times W/s_2 \times 2}, ..., \nabla I_{s_N} \in \mathbb{R}^{H/s_N \times W/s_N \times 2}\}$. The multiscale pixel spatiotemporal information flow, $\nabla I_{ms} = \text{mean}\{\nabla I\}$ is the mean flow after first linearly upsampling individual flows to the image dimensions of the largest scale, s_1. We note unlike optical flow, no scaling factors are applied to the upsampled flows before taking the mean. We reason that the 1D causal measures are already a 'normalised' non-physical unit and independent of the magnitude of the image pixel intensity. Consequently the magnitudes of information flows at individual scales are all comparable and should not be weighted. Using the mean as the agglomerative function is thus in the spirit of obtaining the net causality across scales at each pixel position as well as a quantitative measure of a spatially-persistent causality. In Appendix A we illustrate how this multiscale implementation allows us to fully capture the highway traffic flow filmed from a perspective view such that cars further back are much smaller than those in the front. In contrast the single scale information flows selectively captured only those cars in the back, middle or front. The remainder of the paper always refers to the multiscale pixel spatiotemporal flow implementation.

4 Experiments

We conduct qualitative and quantitative experiments on various video datasets to verify our implementation and demonstrate the utility of the four pixel spatiotemporal information flows in comparison to classic optical flow. Optical flow was computed with the algorithm of (Farnebäck [2003]) which is readily available in the OpenCV library at both single and multiple scales.

4.1 Crowd flow segmentation dataset

4.1.1 Crowd flow information flows

Crowd flow segmentation aims to extract salient motion patterns of high density moving objects where individual objects cannot be independently identified and segmented. Following (Yamashita et al. [2012]) we computed optical and information flows for the dataset of (Ali and Shah [2007]) comprising 30 diverse crowd flow videos, including highway traffic, pedestrian crossing, fish swarms

Table 1: Comparison of optical and information flows to detect perturbed video regions on 4 different perturbations of the crowd flow segmentation dataset. Bold numbers indicate the best value in column.

	Invert		**rot90**		**rot180**		**rot270**	
Flow	mAP_{25}	IoU_{25}	mAP_{25}	IoU_{25}	mAP_{25}	IoU_{25}	mAP_{25}	IoU_{25}
Single-scale optflow	36.9	35.3	93.3	**70.4**	93.3	**68.6**	93.3	**70.4**
Multi-scale optflow	37.6	29.2	**100.0**	64.2	96.7	62.2	**100.0**	63.8
max. CC	**74.8**	**46.4**	84.7	51.3	90.6	51.8	84.7	52.1
cGC	16.3	12.8	30.1	22.6	22.6	19.8	27.2	21.1
PCCA	14.4	12.1	95.3	44.8	**100.0**	46.4	**100.0**	46.4
DDC	47.8	22.6	93.3	55.3	93.3	54.5	93.3	55.0

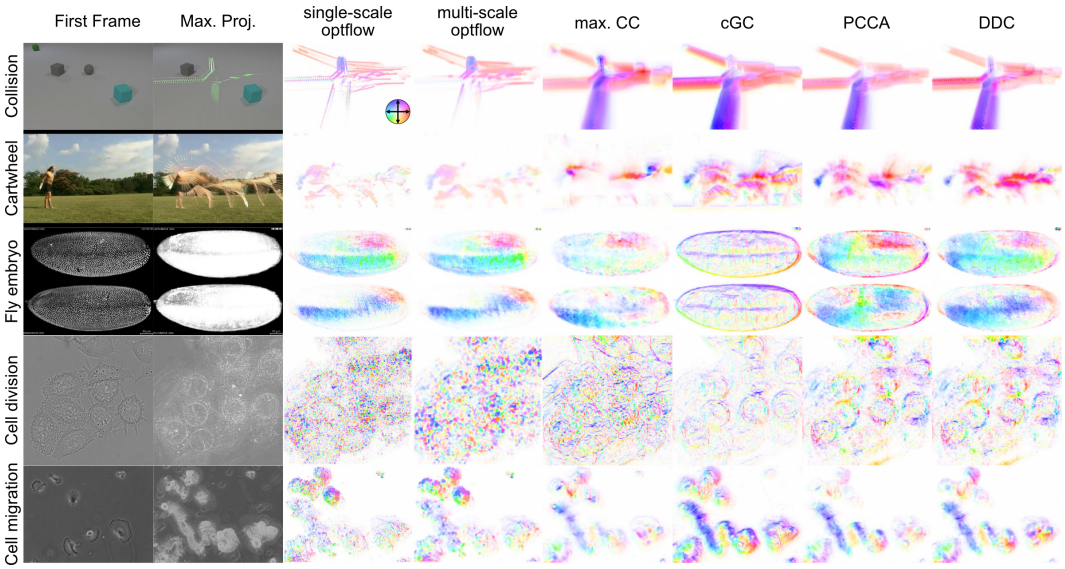

Figure 4: Comparison of the mean single- and multi-scale optical flow with four spatiotemporal information flows. **Row 1:** collision scence. **Row 2:** cartwheel scene. **Row 3:** *Drosophila* embryo. **Row 4:** cell divsion. **Row 5:** cell migration. Direction of flow is colored per the color wheel

and cheerleading. For all videos, we use the full temporal duration, a 3x3 pixel neighborhood, and downsampling scale factors $s = \{1, 2, 4, 8\}$ for multiscale flows. We use a maximum lag $m_{\max}$ of 1 frame for max. CC, cGC and PCCA. Figure 2 shows that mean optical flow (1st two columns) oversmooths motion patterns over time phasing out any transiently occurring patterns or patterns involving only small pixel displacements. The smoothing will be worse, the longer the video duration. In contast, spatiotemporal information flows better attends to the distinct motion patterns present as evidenced by their ability to recover more homogeneous and spatially complete 'information highways'. Notably, due to differences in the modeling assumptions of individual causal measures, the different information flows do exhibit differences in their behavior. Maximum cross-correlation (max. CC) appears to be midway between mean optical flow and DDC flow. It can produce more homogeneous representation of dynamical patterns but can also be sensitive to motion artifact. In the multi-story highway example (Row 3), max.CC over enhances the top-right static buildings. For fish swarming (Row 4) max.CC fails to find the dominant anti-clockwise swirling in the second half of the video. Meanwhile, conditional Granger causality (cGC) is highly sensitive and finds all potential information sources as seen by the sharp imprintings of the zebra crossing (Row 2) and individual people (Row 5) in the final flow. However it also appears cGC is less able to 'weight' the relative intensity of motion patterns within a single video. This results in blurring or overemphasis of information highways when the information is a composition of transient dynamical patterns. Notably, under cGC the full stadium (Row 6) and art gallery (Row 7) is highlighted. Lastly PCCA and DDC appear similar and strikes a good balance between max.CC and cGC. They find the dominant flow patterns whilst better suppressing less interesting or background sources of motion variation.

Table 2: Comparison of performance of optical and information flows to segment moving camouflaged animals over a range of IoU cutoffs. Bold highlights the best value per row.

Cutoffs	**Single-scale optflow**	**Multi-scale optflow**	**max. CC**	**cGC**	**PCCA**	**DDC**
mAP_5	19.9	19.3	10.9	17.7	21.3	**22.9**
IoU_5	4.2	2.2	2.2	4.8	**5.0**	3.3
mAP_{15}	6.7	8.0	8.5	13.6	**15.5**	13.7
IoU_{15}	4.0	2.1	2.2	4.7	**4.9**	3.2
mAP_{25}	3.9	3.9	5.5	9.8	**10.6**	7.7
IoU_{25}	3.6	1.9	2.0	4.3	**4.6**	2.8

This is nicely seen in the case of the stadium where both highlight the walkway between the seated spectators instead of highlighting everything (optical flow and cGC) or nothing (max.CC).

4.1.2 Perturbed crowd flow information flows

To understand quantitatively the behavior of the different information flows, we generated perturbed versions of the crowd flow segmentation dataset by inverting the pixel intensities in the central 50x50 region of each video every frame or rotating this region by 90, 180, 270 degrees. Flows were computed with the same parameters as previously. Figure 3 shows an example of the perturbation and the computed flows. Visually all flows detect a difference in the flow direction after region rotation. However flow differences appear less apparent for intensity inversion. We posed the detection of the perturbed region as a bounding box detection problem and measured the detection performance of each flow. Predicted bounding boxes were obtained by thresholding the magnitude difference in flow between the unperturbed and perturbed video and keeping the largest connected component. Table 1 summarises the detection results. When rotated, optical flow and PCCA optimized for motion flow performed best. However both max. CC and DDC perform competitively. Notably when pixel intensity was inverted such that the motion remained the same, cGC and PCCA performed the worst but both max. CC and DDC outperform optical flow variants. Averaging the mAP across all four perturbations, max. CC was best (83.7), then multi-scale optical flow (83.6), DDC (81.9), single-scale optical flow (79.2), PCCA (77.4) and cGC worst (24.1).

4.2 Information flows on diverse motion videos

The crowd flow videos primarily exhibit dynamic patterns driven by the persistent, unidirectional motion of objects. we therefore extracted spatiotemporal information flows next for videos exhibiting diverse dynamics and acquired from different sources to test generality using the same parameters as for crowd flow, (Figure 4, see Appendix B for links to the datasets used). On simulated collisions of objects (Yi et al. [2018]) information flows accurately extract the collision information highways irrespective of whether the objects were initially present or subsequently have left the frame. Optical flow could only track the object corners, (Row 1). On complex human actions like a cartwheel, information flows capture coordinated body part motion to recreate a 'flow' of key snapshot postures (Row 2). *Drosophila* (fruitfly) embryos undergo characteristic, motion patterns during development for correct gene patterning (Alberts [2017]). These patterns occur in sequential order, are transient, and comprise both small and dramatic motion patterns, separated by periods of stationarity (Zhou [2017]). Optical flow and max. CC capture just the largest motion patterns, cGC capture the subtle vibration in the stationary periods, only PCCA and DDC capture a superposition of all occurring patterns (Row 3). In cell division, where motion is mostly stochastic morphological fluctuations with division occurring over only a brief moment, only the dynamical systems based information flows (cGC, PCCA and DDC) recovered the circular flow of cell division. Lastly, we tested cell migration where individual cells also actively change shape whilst moving (Row 5). Optical flow obtains a patchy flow of the migration. In contrast information flow better extracts spatiotemporally consistent histories of both the subcellular morphodynamics and the cell migration trajectory of the cell centroid. In conclusion, information flows are generally applicable, and demonstrate an enhanced capacity to retain temporal history and uncover salient dynamic patterns even over long video durations and for complex nonlinear dynamics.

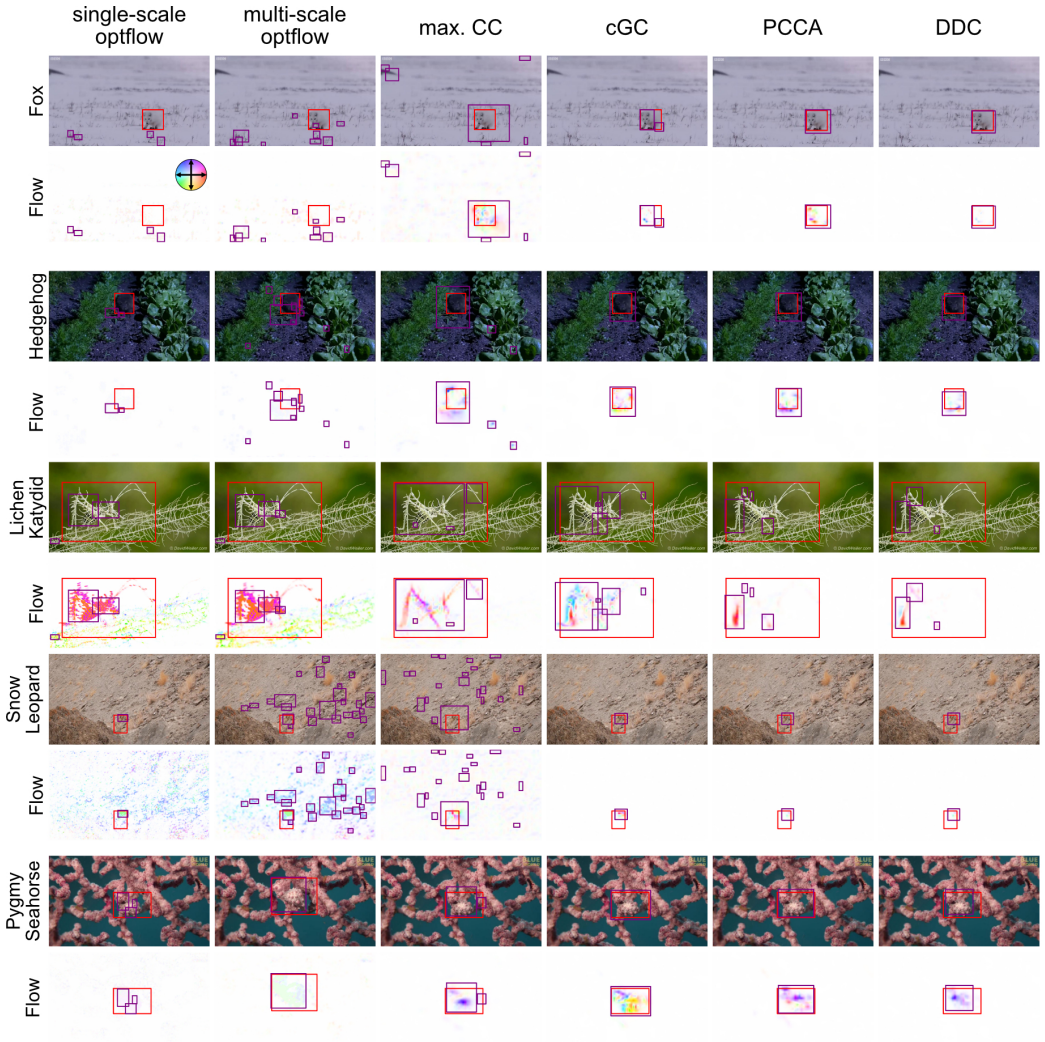

Figure 5: Comparison of the mean single- and multi-scale optical flow with spatiotemporal information flow detection of different camouflaged moving animals in video frames. Red box = ground truth. Purple box = detection by binary thresholding of flow and running connected components. Direction of flow is colored per the color wheel.

4.3 Moving camouflaged animals (MoCA) dataset

The previous experiments all used the full available temporal duration of a given video (ranging from 10s to 1000s of frames) to compute flows. From these experiments we can conclude that information flows better attend to salient motion patterns in a video over long times. From the extensive literature on applying causal measures to 1D timeseries we know that these measures are expected to perform optimally when dynamic processes are stationary and when timeseries are sampled over long durations e.g. DDC (Chen et al. [2022]). Our last experiment seeks therefore to assess whether information flows could also even offer advantages over very short timeseries of a few frames. This is important practically as events may be sporadic or chaotic. We conduct the test with the task of detecting camouflaged animals through their subtle movements. The moving camouflaged animals (MoCA) is a comprehensive dataset with 141 videos of animals. Each video is of a single animal instance which lasts 10s to 100s of frames. The animal has been annotated with bounding boxes every 5 frames. We test the ability of individual flows to detect the animal in every annotated frame given only a video subsequence of temporal duration 11 frames. 11 frames was chosen as 5 frames either side of the annotated frame. Bounding boxes were generated from the magnitudes of the computed information flows by binary thresholding (median + std magnitude cutoff). Removal

of overly small connected component regions (<250 pixels) followed by morphological dilation (disk kernel with 3 pixel radius) and morphological holefilling was used to postprocess the initial binarisation before bounding box extraction for each remaining individual connected components. No non-max bounding box suppression as commonly used in detection applications was used. The majority of MoCA videos exhibit significant camera motion which required sophisticated registration. As the correction of these camera-related acquistion artifacts is not in the scope of this paper, we computed flows only on a subset of 40 videos for which we manually identified to have little to no visible camera artifact (see Appendix B). Notably, these videos exhibit animals of different sizes and shapes. Frames do not necessarily exhibit any animal motion - the majority in fact does not. We did not however constrain prediction and evaluation to only frames of animals moving as it does not affect fair comparison between the different flows, and we find it also illuminating to understand the extent of the dataset that cannot be detected using any form of flow. To make the total computation time tractable for the total number of frames in this reduced 40 video dataset (2,321 total annotated frames), we resized all videos whilst preserving the aspect ratio so that the height of all video frames was 512 pixels. We then computed flows on the resized videos with the same parameters as for crowd flow. Table 2 shows the quantitative detection results for different IoU cutoffs. Notably cGC, PCCA, DDC information flows significantly outperform and are more robust than optical flow. DDC achieves the highest mean average precision of 22.8 with IoU cutoff of 5. Across all cutoffs, PCCA is best overall and is always better than cGC but DDC performs competitively. To investigate the globally low IoU cutoffs we visualised the detections (purple boxes) relative to the ground truth bounding box (red box), Figure 5. We found that the low cutoffs were justified. Contrary to detecting the contour of the whole animal, information flows by definition targetly aims to isolate the precise motion sources and sinks as evidenced by their flow magnitude being localised and near to the moving arms, legs and heads of individual animals. We note from the fox and hedgehog examples that cGC, PCCA and DDC compute less noisy flows even when motion was minimal and to generate significantly reduced extraneous bounding boxes due to background motion. In the lichen katydid example, all four information flows precisely focus and isolate the organism's moving appendages. A similar pattern is found in the snow leopard case. Lastly, all methods perform well when there is clear relative motion of the animal to its background such as in the pygmy seahorse case. Overall the qualitative visualisation show that PCCA and DDC are the best and nearly inseparable. We hypothesise from the lichen katydid and pygmy seahorse examples where PCCA and DDC differ the most, that the use of a pixel block-based flow in PCCA yields slightly larger bounding boxes that help it to overlap better with the ground-truth annotations which are based on the full animal.

5 Conclusion

In conclusion, we have introduced a formal notion of a multiscale pixel spatiotemporal information flow in this paper which enables the operationalisation and application of 1D causal measures to discover causal information relationships amongst individual pixels in 2D + time video over short and long times. We demonstrated through application to diverse video datasets that these information flows possess the ability to capture community structure in complex dynamical systems, and provides a more robust and informative alternative to standard computer vision optical flow approaches. For simplicity, we implemented our multiscale framework based on Gaussian image pyramids used in optical flow estimation and combined the flows from different scales using average pooling. Future work will investigate more optimal methods to combine information flows across individual scales. We would like to investigate if the spatial Gaussian pyramids could also be extended to time to allow the combination of spatiotemporal flows computated over different sized time intervals in addition to different spatial scales. We hypothesise the resultant flow may better adapt to the spatiotemporal complexity of the video e.g. to better handle intervals of minimal movement. Encouraged by the performance of the DDC flow based on a linear ODE system with constant coefficients, we would like to also experiment with using more complex ODE models to better handle nonlinear dynamics.

6 Code Availability

An extensible Python library for implementing the four multiscale pixel spatiotemporal information flows in this paper is available in the following GitHub, `https://github.com/DanuserLab/spatiotemporal_information_flows`.

7 Acknowledgements and Funding

We gratefully acknowledge mentorship and support from Prof. Gaudenz Danuser. This project was funded by an NIH grant, grant ID:R35 GM136428 to Prof. Gaudenz Danuser. We also thank Prof. Jungsik Noh for discussions on the computational implementation of conditional Granger causality.

References

Ronald J Adrian. Twenty years of particle image velocimetry. *Experiments in fluids*, 39(2):159–169, 2005.

Bruce Alberts. *Molecular biology of the cell*. WW Norton & Company, 2017.

Saad Ali and Mubarak Shah. A lagrangian particle dynamics approach for crowd flow segmentation and stability analysis. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6. IEEE, 2007.

Lionel Barnett and Anil K Seth. Granger causality for state-space models. *Physical Review E*, 91(4): 040101, 2015.

André M Bastos and Jan-Mathijs Schoffelen. A tutorial review of functional connectivity analysis methods and their interpretational pitfalls. *Frontiers in systems neuroscience*, 9:175, 2016.

Steven L Bressler and Anil K Seth. Wiener–granger causality: a well established methodology. *Neuroimage*, 58(2):323–329, 2011.

Yusi Chen, Burke Q Rosen, and Terrence J Sejnowski. Dynamical differential covariance recovers directional network structure in multiscale neural systems. *Proceedings of the National Academy of Sciences*, 119(24):e2117234119, 2022.

Matteo Detto, Annalisa Molini, Gabriel Katul, Paul Stoy, Sari Palmroth, and Dennis Baldocchi. Causality and persistence in ecological systems: a nonparametric spectral granger causality approach. *The American Naturalist*, 179(4):524–535, 2012.

Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Scandinavian conference on Image analysis*, pages 363–370. Springer, 2003.

André Fujita, Joao Ricardo Sato, Kaname Kojima, Luciana Rodrigues Gomes, Masao Nagasaki, Mari Cleide Sogayar, and Satoru Miyano. Identification and quantification of granger causality be-tween gene sets. *arXiv preprint arXiv:0911.1159*, 2009.

John F Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388):907–915, 1984.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

Min Liu and Tobi Delbruck. Adaptive time-slice block-matching optical flow algorithm for dynamic vision sensors. 2018.

Sanath Narayan and Kalpathi R Ramakrishnan. A cause and effect analysis of motion trajectories for modeling actions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2633–2640, 2014.

Stephanie Noble, Dustin Scheinost, and R. Todd Constable. A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage*, 203:116157, 2019. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2019.116157.

Jungsik Noh, Tadamoto Isogai, Joseph Chi, Kushal Bhatt, and Gaudenz Danuser. Granger-causal inference of the lamellipodial actin regulator hierarchy by live cell imaging without perturbation. *Cell Systems*, 2022.

Karthir Prabhakar, Sangmin Oh, Ping Wang, Gregory D Abowd, and James M Rehg. Temporal causality for the analysis of visual events. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1967–1974. IEEE, 2010.

Anil K Seth, Adam B Barrett, and Lionel Barnett. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297, 2015.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.

Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. Optical flow guided feature: A fast and robust motion representation for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1390–1399, 2018.

Eran Swears, Anthony Hoogs, Qiang Ji, and Kim Boyer. Complex activity recognition using granger constrained dbn (gcdbn) in sports and surveillance video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Raju Tomer, Khaled Khairy, Fernando Amat, and Philipp J Keller. Quantitative high-speed imaging of entire developing embryos with simultaneous multiview light-sheet microscopy. *Nature methods*, 9(7):755–763, 2012.

Vladimír Ulman, Martin Maška, Klas EG Magnusson, Olaf Ronneberger, Carsten Haubold, Nathalie Harder, Pavel Matula, Petr Matula, David Svoboda, Miroslav Radojevic, et al. An objective comparison of cell-tracking algorithms. *Nature methods*, 14(12):1141–1152, 2017.

Pedro A Valdes-Sosa, Alard Roebroeck, Jean Daunizeau, and Karl Friston. Effective connectivity: influence, causality and biophysical modeling. *Neuroimage*, 58(2):339–361, 2011.

Norbert Wiener. The theory of prediction. *Modern mathematics for engineers*, 1956.

Yuya Yamashita, Tatsuya Harada, and Yasuo Kuniyoshi. Causal flow. *IEEE Transactions on multimedia*, 14(3):619–629, 2012.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. *Advances in neural information processing systems*, 31, 2018.

Amos Zamir, Guanyu Li, Katelyn Chase, Robert Moskovitch, Bo Sun, and Assaf Zaritsky. Emergence of synchronized multicellular mechanosensing from spatiotemporal integration of heterogeneous single-cell information transfer. *Cell Systems*, 13(9):711–723.e7, 2022. ISSN 2405-4712. doi: https://doi.org/10.1016/j.cels.2022.07.002.

Felix Zhou. *Phenotyping cellular motion*. PhD thesis, University of Oxford, 2017.

Felix Y Zhou, Carlos Ruiz-Puig, Richard P Owen, Michael J White, Jens Rittscher, and Xin Lu. Motion sensing superpixels (moses) is a systematic computational framework to quantify and discover cellular motion phenotypes. *elife*, 8:e40162, 2019.

A Information flow at different scales

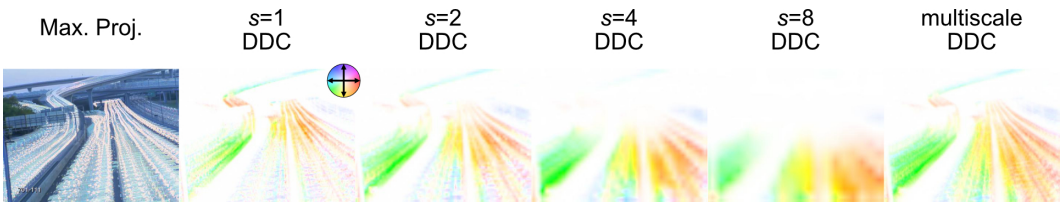

Figure 6: Visualisation of the individual information flow at each of the 4 dowsampling factors compared to the multiscale flow, the average of the 4 for dynamic differential covariance flow. Direction and magntiude of flow is colored per the color wheel.

B Datasets used

Below are links to the datasets used in this paper.

- Crowd flow segmentation dataset (Ali and Shah [2007]), https://www.crcv.ucf.edu/research/data-sets/crowd-segmentation.
- CLEVRER: CoLlision Events for Video REpresentation and Reasoning dataset for videos of collisions, (Yi et al. [2018]). http://clevrer.csail.mit.edu/.
- HMDB: a large human motion database for cartwheel example, (Kuehne et al. [2011]). https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/#Downloads.
- Tomer et al. [2012] supplementary movie 3 for *Drosophila* fruitfly embryo development example.
- HeLa cells on a flat glass (DIC-C2DH-HeLa.zip) for cell divison and Glioblastoma-astrocytoma U373 cells on a polyacrylamide substrate (PhC-C2DH-U373.zip) from the cell tracking challenge, (Ulman et al. [2017]). http://celltrackingchallenge.net/2d-datasets/
- MoCA : Moving camouflaged animals dataset, (Lamdouar et al. [2020]). https://www.robots.ox.ac.uk/ vgg/data/MoCA/. We specifically used the following 40/141 videos with least camera artifacts for evaluating animal detection: arctic_fox_2, arctic_wolf_0, crab_2, dead_leaf_butterfly_1, desert_fox, egyptian_nightjar, flatfish_3, flower_crab_spider_0, flower_crab_spider_1, flower_crab_spider_2, fossa, grasshopper_1, grasshopper_2, hedgehog_1, hedgehog_3, hyena, jerboa, jerboa_1, lichen_katydid, meerkat, mongoose, moth, pallas_cat, polar_bear_3, potoo, pygmy_seahorse_4, rabbit, rodent_x, rusty_spotted_cat_0, rusty_spotted_cat_1, smallfish, snow_leopard_4, snow_leopard_5, snow_leopard_10, snowy_owl_0, snowy_owl_2, spider_tailed_horned_viper_1, spider_tailed_horned_viper_3, white_tailed_ptarmigan, wolf.