Motion Field and Optical Flow

Carlo Tomasi

March 6, 2019

There are many reasons why analyzing the motion of objects in video is useful. Traffic monitoring, gesture recognition, American Sign Language interpretation, tracking of cardiac contractions, analysis of sports video, weather forecasting from satellite imagery, interactive video games are a tiny sample of possible applications. Each frame of a video sequence is just the measurement of a distribution of light on a sensor, and motion is not directly encoded anywhere in a video: Our brains *compute* motion by making assumptions about the world and comparing consecutive vide frames, and computer algorithms for video analysis must do that, too. This inference turns out to be surprisingly difficult, given that our visual systems make it routinely with no apparent effort. This note attempts to give some reasons why by describing how video images are formed, and pointing out some of the assumptions the inference is typically based on. Algorithms for making the inference are described in later notes.

Before it is converted to digital information, the image formed by the lens of a camera is projected onto the camera's sensor, which sees a continuous distribution of light. It is useful to consider this image because it is a real-valued function of two real variables, in contrast with the image we process on the computer, which is a discrete array of integer pixel values. We also view images as changing over time ("video"), as we are interested in image motion. The first two sections below introduce a model of the relationship between the moving light patterns on the sensor and the image arrays we process on a computer. More specifically, section 1 considers a single frame and section 2 describes the motion of these patterns over time. Section 3 then introduces a fundamental assumption that is typically made when analyzing video, namely, that the appearance of a point in the scene remains constant over time, and Section 4 describes a key observability issue that makes the analysis of video motion is fundamentally difficult. Finally, Section 5 categorizes, at a high level, different approaches to estimating visual motion.

1 A Video Model

Every point $\mathbf{x} \in \mathbb{R}^2$ on the sensor sees some pattern of colors $\mathbf{e}(\mathbf{x}, t)$ at time $t \in \mathbb{R}$. This function is called the *irradiance*. While irradiance is generally a spectral distribution of wavelengths, the camera uses three sets of filters to record the amount of light in each of three bands of wavelengths roughly centered around red, green, and blue light. Accordingly, we encode the irradiance by a triple of nonnegative real numbers, $\mathbf{e} \in \mathbb{R}^3$ with $\mathbf{e} \geq 0$.

The output of the camera is an *image sequence* (or *video*)

$$\mathbf{f} \ : \ \Omega \times \mathbb{Z} \to \mathbb{N}^3 \ ,$$

a discrete three-dimensional array $\mathbf{f}(\cdot, n)$ of nonnegative integers at each of a discrete sequence of times n. The *image domain* $\Omega \subset \mathbb{N}^2$ for the discrete space variable $\mathbf{i}$ is a finite rectangular grid of nonnegative integers. While every video has a beginning, and many have an end, it is often mathematically simpler to view every video as being extended indefinitely into the past and future. This is why n is modeled as an integer, $n \in \mathbb{Z}$, rather than a natural number ($n \in \mathbb{N}$).

Each *pixel value* $\mathbf{f}(\mathbf{i}, n)$ is the integral of irradiance over a small, rectangular volume of space and time:

$$\mathbf{f}(\mathbf{i}, n) \;=\; Q\left(\int_{nT-T_e/2}^{nT+T_e/2}\left[\iint_{\mathbf{i}P-P_s/2}^{\mathbf{i}P+P_s/2} \mathbf{e}(\mathbf{x}, t)\; d\mathbf{x}\right] dt \;+\; \boldsymbol{\nu}(\mathbf{i}, n)\right)\;. \tag{1}$$

In this expression, the positive real number P is the *pixel pitch*, that is, the distance between adjacent pixels, which we assume to be the same in the vertical and horizontal direction. The positive real number P_s is the *pixel size*, that is, the part of the pixel pitch over which each pixel senses light. The positive real number T is the *frame interval*, namely, the time elapsed between a frame and the next, and the positive real number $T_e \leq T$ is the *exposure time*, that is, the part of T over which incoming light is recorded. Finally, $\boldsymbol{\nu}$ is measurement noise, and the function $Q(\cdot)$ *quantizes* the measured values, that is, it rounds them to integers. This function also clips values to be between 0 and 255.

2 The Motion Field

As the world and/or the camera move, a point in the scene typically projects to different points on the sensor plane at different times, and the sensor position of the point's image is therefore described by a function of time. Consider the irradiance $\mathbf{e}(\mathbf{x}, s)$ at a particular point $\mathbf{x}$ on the sensor plane and time s. Point $\mathbf{x}$ is the projection of some point in the world (possibly infinitely far away), and we denote by $\mathbf{y}(\mathbf{x}, s, t)$ the trajectory that the projection of that world point traces on the sensor plane as a function of time t, so that in particular

$$\mathbf{y}(\mathbf{x}, s, s) = \mathbf{x}\;. \tag{2}$$

The image velocity of $\mathbf{y}(\mathbf{x}, s, t)$ at time t is

$$\mathbf{w}(\mathbf{x}, s, t) \;\stackrel{\mathsf{def}}{=}\; \frac{\partial \mathbf{y}(\mathbf{x}, s, t)}{\partial t} \tag{3}$$

and the velocity at time s,

$$\mathbf{v}(\mathbf{x}, s) \;\stackrel{\mathsf{def}}{=}\; \mathbf{w}(\mathbf{x}, s, s) \tag{4}$$

is called the *motion field* at $(\mathbf{x}, s)$.

At time $t > s$, the point is at position

$$\mathbf{y}(\mathbf{x}, s, t) = \int_s^t \mathbf{w}(\mathbf{x}, s, \tau)\, d\tau$$

(from the fundamental theorem of calculus and the definition of $\mathbf{w}$) and the vector difference

$$\mathbf{d}(\mathbf{x}, s, t) \;\stackrel{\mathsf{def}}{=}\; \mathbf{y}(\mathbf{x}, s, t) - \mathbf{y}(\mathbf{x}, s, s) \tag{5}$$

is called the *displacement* at $\mathbf{x}$ between times s and t.

The world points that are visible at a particular pixel position at time s may become *occluded* at time t, that is, either become hidden behind some world object or leave the field of view of the camera altogether. Conversely, points that are visible at time t may be occluded at time s. For the points that are occluded at either time the displacement is undefined. All other points at one time have a *corresponding* point at the other time. If transparent objects are ignored, which is common practice, each point at one time has *at most one* corresponding point at the other time.

The function $\mathbf{v} : \mathbb{R}^3 \to \mathbb{R}^2$ is usually modeled as piecewise smooth, and its discontinuities are smooth surfaces in $\mathbb{R}^3$ called *motion boundaries*. They are typically caused by objects in the world that move in front of parts of other objects that move differently. Sometimes the term "motion boundaries" refers to the curves obtained by restricting a motion-boundary surface to a specific time s.

Eulerian and Lagrangian Viewpoint The two functions $\mathbf{v}$ and $\mathbf{w}$ are very closely related to each other (through equation 4), but they describe the motion of points in the image from two very different perspectives. The motion field $\mathbf{v}$ reflects an *Eulerian* view of things: Imagine Euler sitting at a fixed point $\mathbf{x}$ on the sensor plane. He sees the image of a world point go by at time s and measures its velocity $\mathbf{v}(\mathbf{x}, s)$ on the image plane. At a different point t in time, a different world point will generally pass by at $\mathbf{x}$. Euler will still be there, measuring *that* point's motion.

In contrast, the trajectory velocity $\mathbf{w}(\mathbf{x}, s, t)$ reflects a *Lagrangian* perspective: Lagrange is actually riding the image point $\mathbf{y}(\mathbf{x}, s, t)$, following it as it moves around the image, and keeps measuring its image velocity.

Only at time s is Lagrange in the same place as Euler, a very fleeting encounter of two great mathematicians.[1]

The Optical Flow The motion field is not always observable in a video, and not every change in video corresponds to a motion of visible point in the scene. For instance, a perfectly smooth, evenly colored sphere rotating around its axis produces no changes in the image: The motion field is nonzero, but it leaves no trace in the video. Conversely, a static scene lit by a moving light source produces changes in the image, but none of the points that are visible in the video move: The motion field is zero, but the image changes over time.

Thus, the estimation of the motion field from video suffers from a fundamental observability issue, and the term *optical flow* has been used in the literature to denote whatever aspects of the motion field are observable. For instance, with the rotating sphere, we would say that the motion field is nonzero but the optical flow is zero. In the case of lighting change described above, the reverse is typically the case: The optical flow is nonzero but the motion field is zero. Thus, in some sense, the motion field is the true motion (*i.e.*, the projection of world motion onto the image plane), while the optical flow is the measured motion.

However, there is no precise definition for "optical flow," and as a result this term is often used interchangeably with "motion field." We will do so as well in these notes, but it is important to remain cognizant of this fundamental conceptual discrepancy between what we would like to measure (the noumenon, in Kant's terminology) and what we can actually measure from video (the phenomenon).

[1] Although Euler (1707-1783) and Lagrange (1736-1813) never met in person, they had an active correspondence by mail between 1754 and 1775 around the topic of the calculus of variations, which they each invented independently. Their lifetimes overlapped by about 48 years.

3 The Optical Flow Constraint Equation

To measure the motion field from video it is necessary to make some assumption on how frames in a video sequence relate to each other. Without such an assumption, the frames could be images of entirely unrelated scenes, and the notion of "image motion" would then be meaningless. After all, images do not move, but are merely arrays of irradiance measurements, and one needs to *interpret* the changes in the observed brightness patterns as motion by assuming that different frames are related views of the same scene.[2]

The most widely made assumption in the literature on visual motion analysis is that the appearance of any given point in the world does not change over time, at least over small temporal intervals, as the point moves across the field of view: A point in the world that looks pink and of a certain brightness now will be of the same pink and brightness a second from now if we keep looking at it. This assumption of *constant appearance* is violated as a result of changes in shading or illumination, or because the same point in the world may appear different when viewed from different viewpoints because it reflects light differently in different directions. Nonetheless, the assumption holds for a wide variety of circumstances, and is arguably one of the weakest (*i.e.*, most generic) assumptions one can make on how images in video relate to each other.

Mathematically, the assumption of constant appearance can be expressed by taking a Lagrangian point of view: Pick some point in the world, and ride its projection $\mathbf{x}$ as it moves through the image. Then, the irradiance $\mathbf{e}(\mathbf{x}(t), t)$ is constant:

$$\frac{d\mathbf{e}(\mathbf{x}(t), t)}{dt} \stackrel{\text{def}}{=} \lim_{\Delta t \to 0} \frac{\mathbf{e}(\mathbf{x}(t + \Delta t), t + \Delta t) - \mathbf{e}(\mathbf{x}(t), t)}{\Delta t} = 0 \,. \tag{6}$$

In this differential form, the constant-appearance assumption is called the *optical flow constraint equation*. Expanding equation 6 through the chain rule of partial differentiation yields

$$\frac{\partial \mathbf{e}}{\partial \mathbf{x}^T} \frac{d\mathbf{x}}{dt} + \frac{\partial \mathbf{e}}{\partial t} = 0$$

or, from equations 2, 3, and 4,

$$\frac{\partial \mathbf{e}}{\partial \mathbf{x}^T} \mathbf{v} + \frac{\partial \mathbf{e}}{\partial t} = 0 \,. \tag{7}$$

This is a system of up to three equations in the two components of $\mathbf{v}$, the unknown motion field. The derivatives of the irradiance $\mathbf{e}$ can be measured or at least estimated from video, so that *equation 7 is the mathematical basis for the computation of the motion field under the assumption of constant appearance.*

Notation: For color images, the vector $\mathbf{e}$ is a column vector with three components, and the vector $\mathbf{x}^T$ is a row vector with two components. The expression $\frac{\partial \mathbf{e}}{\partial \mathbf{x}^T}$ is the matrix of the partial derivatives of every component of $\mathbf{e}$ (one row per component) with respect to every component of $\mathbf{x}^T$ (one column per component):

$$\frac{\partial \mathbf{e}}{\partial \mathbf{x}^T} \stackrel{\text{def}}{=} \begin{bmatrix} \frac{\partial e_1}{\partial x_1} & \frac{\partial e_1}{\partial x_2} \\ \frac{\partial e_2}{\partial x_1} & \frac{\partial e_2}{\partial x_2} \\ \frac{\partial e_3}{\partial x_1} & \frac{\partial e_3}{\partial x_2} \end{bmatrix} .$$

[2]Let this remark sink in for a moment: Image motion is not "directly visible," but is rather the result of a computation (what we can call an "interpretation" of what we see). Eyes are not enough to "see," we also need a brain.

This matrix is called the *Jacobian* of $\mathbf{e}$ with respect to $\mathbf{x}$. Row i of the Jacobian is the transpose of the gradient of e_i with respect to $\mathbf{x}$, also called the *spatial gradient* of e_i:

$$\left[\begin{array}{cc} \frac{\partial e_i}{\partial x_1} & \frac{\partial e_i}{\partial x_2} \end{array} \right] = (\nabla e_i)^T \quad \text{for} \quad i = 1, 2, 3 \; .$$

This Jacobian is *partial* because $\mathbf{e}$ also depends on time t. The full Jacobian would be

$$\left[\begin{array}{ccc} \frac{\partial e_1}{\partial x_1} & \frac{\partial e_1}{\partial x_2} & \frac{\partial e_1}{\partial t} \\ \\ \frac{\partial e_2}{\partial x_1} & \frac{\partial e_2}{\partial x_2} & \frac{\partial e_2}{\partial t} \\ \\ \frac{\partial e_3}{\partial x_1} & \frac{\partial e_3}{\partial x_2} & \frac{\partial e_3}{\partial t} \end{array} \right] .$$

4 The Aperture Problem

With color images, $\mathbf{e} \in \mathbb{R}^3$, and equation 7 is a system of three equations in the two components of $\mathbf{v}$, the unknown motion field. However, the three color components of an image are often very strongly correlated to each other, and the partial Jacobian $\frac{\partial \mathbf{e}}{\partial \mathbf{x}^T}$ is likely to be poorly conditioned, and perhaps even have rank 1: While the red, green, and blue components are different from each other, their spatial changes are often the result of shading or changes of lighting, which affect the three components multiplicatively in the same way: If red gets twice as dark when looking at two different pixels, so do green and blue if the change is caused by brightness and not hue variations. In that case, the system 7 degenerates to a single independent equation in two variables:

$$\frac{\partial e}{\partial \mathbf{x}^T} \mathbf{v} + \frac{\partial e}{\partial t} = 0 \quad \text{with} \quad e \in \mathbb{R} \; . \tag{8}$$

This degeneracy is called the *aperture problem*, and makes the motion field unobservable without further assumptions. The aperture problem is unavoidable with black-and-white video ($\mathbf{e} \in \mathbb{R}$), and is pervasive for color video as well. Let us look at this problem in the black-and-white case.

Since $\mathbf{v}$ appears in equation 8 through an inner product, only its component along the spatial gradient of e can be observed, assuming that this gradient is nonzero. If we let

$$\nabla e(\mathbf{x}) \stackrel{\text{def}}{=} \frac{\partial e(\mathbf{x})}{\partial \mathbf{x}}$$

be the spatial gradient (a column vector) of the image irradiance and we assume that $\nabla e(\mathbf{x}) \neq \mathbf{0}$ at $\mathbf{x}$, then the scalar

$$v(\mathbf{x}) \stackrel{\text{def}}{=} \|\nabla e(\mathbf{x})\|^{-1} \, [\nabla e(\mathbf{x})]^T \, \mathbf{v}(\mathbf{x})$$

is the component of the motion field along the spatial gradient of e and is called the *normal component* of the motion field.

Thus, when $\nabla e(\mathbf{x}) \neq \mathbf{0}$ and the optical flow constraint equation 8 holds, only the normal component of the motion field is observable, while the component orthogonal to $\nabla e(\mathbf{x})$ is not. In addition, the earlier discussion of the difference between motion field and optical flow in the case of the rotating sphere is an example where the motion field is completely unobservable even if $\nabla e(\mathbf{x}) \neq \mathbf{0}$: In that case, the assumption of constant appearance, and therefore the optical flow constraint equation, is violated, because a fixed point on the rotating sphere changes in appearance (brightness) as it enters areas with different amounts of lighting.

The aperture problem arises even if one considers a small but non-infinitesimal patch of the image. For instance, if the patch covers a part of the image where the brightness or color pattern is

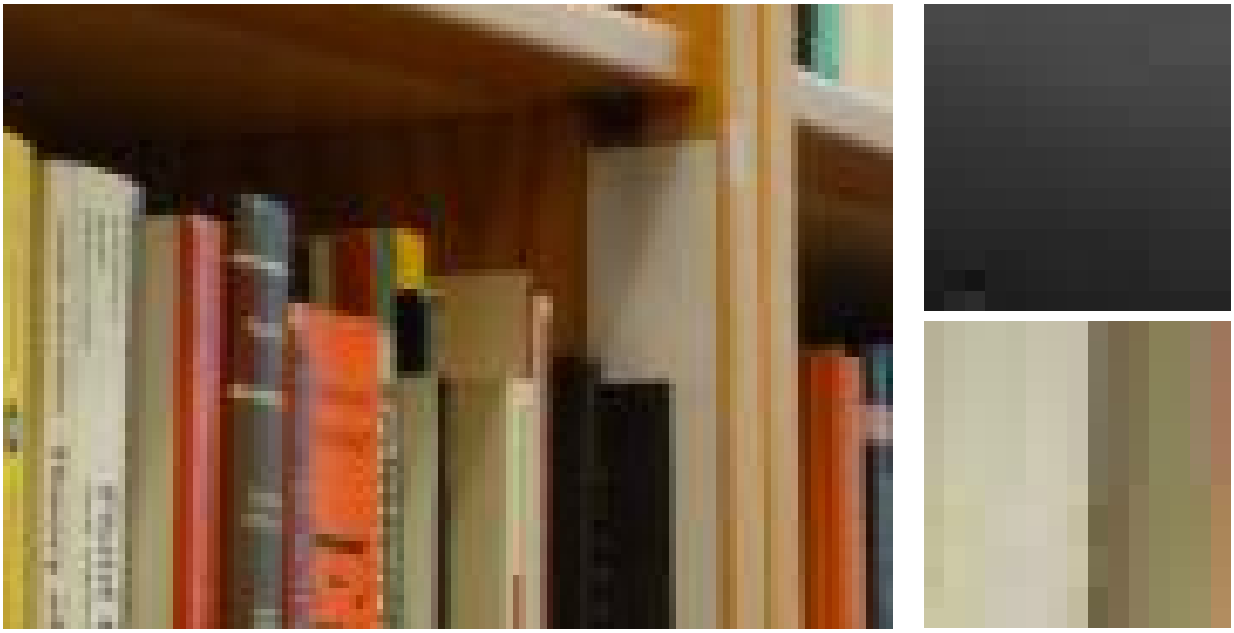

Figure 1: The 15×15 patch at the top right comes from the relatively blank area at the top left of the low-resolution image detail shown on the left. Shifting that patch around by a few pixels does not change its contents appreciably, and any changes that do occur may be small compared to image noise. The 15×15 patch at the bottom right comes from around one of the many vertical edges in the image. Shifting the patch by even a small amount horizontally, say to the right, will cause the edge in it to move to the left, so the horizontal component of motion can be reliably determined along a vertical edge. However, if the patch is shifted vertically by a small amount, changes in it are minor, and may again be overwhelmed by image noise: The vertical component of motion cannot be measured reliably.

more or less uniform, such as in the middle of a blank wall, a small motion inside the window may not lead to noticeable differences in the window contents. In that case, the window's brightness distribution does not allow determining the displacement of the point feature reliably. A less severe version of the aperture problem arises when the feature window straddles a straight intensity edge. In that case, motion across the edge is clearly visible, but motion along it is not. Figure 1 illustrates.

5 Estimating the Motion Field

Video frames are recorded at regular time intervals, as shown in equation 1. Estimating the motion field amounts to using these frames to determine a displacement (equation 5) for each pixel from one frame to the next.

Displacements are not necessarily integer-valued, even in the absence of occlusions: A point visible at the center of a particular pixel in the first frame is not necessarily visible at the center of some pixel in the second frame, but may rather be visible at some position between pixel centers. Because of this, the computation of displacements cannot be framed as a partial bipartite matching (match pixels to pixels), but is rather cast as the computation of a vector field defined on either frame: For each pixel $\mathbf{i} \in \Omega$ in frame $n \in \mathbb{Z}$ one assigns a real-valued displacement $\mathbf{d}(\mathbf{i}P, nT, (n+1)T)$ or, if the point becomes occluded in frame $n+1$, no displacement. If the frame interval T is very

small, this displacement may yield an acceptable approximation for a multiple of the optical flow or motion field $\mathbf{v}$:

$$\mathbf{u}(\mathbf{i}, n) \stackrel{\text{def}}{=} T\mathbf{v}(\mathbf{i}P, nT) \approx \mathbf{d}(\mathbf{i}P, nT, (n+1)T) \tag{9}$$

at that pixel and time.

Because of the aperture problem, the motion field at pixel position $\mathbf{i}$ must be estimated from the pixel values (in two or more video frames) in an area of the image that extends beyond the point $\mathbf{i}$ itself. This area is called the *support* of the computation. Two types of support are common:

- The support is a square centered at $\mathbf{i}$ and with a side of $s = 2h + 1$ pixels, where h is a positive integer, so that s is odd. The square is called the *window* centered at $\mathbf{i}$. The motion field at two different points $\mathbf{i}$ and $\mathbf{i}'$ is estimated through separate computations, even if when $\|\mathbf{i} - \mathbf{i}'\|_\infty \leq s$, so that the two windows overlap.

  When a large window is used, the brightness pattern within the window is likely to be more varied than in a smaller window, and the aperture problem is less likely to arise. However, a single displacement is reported for the entire window, even if multiple displacements occur within it. As a consequence, larger windows lead to lower variance but higher bias in the results, and displacements computed from overlapping windows are potentially more similar to each other than the true displacements are. In other words, low variance (that is, low sensitivity to noise) leads to low spatial resolution in the resulting displacements.

- The support is the entire image. In this case, the motion is estimated jointly at all pixel positions, and the assumption of a single displacement within the support is of course unwarranted. Instead, each pixel is assigned a (possibly different) motion field vector, and some form of regularization is used to address the uncertainty implied by the aperture problem. Typically, the estimated motion field is made to be spatially smooth by levying a penalty on differences between field values at neighboring pixel positions. Again, a larger penalty leads to lower variance but tends to smooth the motion field estimates, with the effect being particularly felt (high bias) along motion discontinuities.

Sample methods for visual motion estimation with each type of support are discussed in subsequent notes.