Vision Research 150 (2018) 44–53

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Motion estimation: A biologically inspired model

L. Bowns

Cambridge Computational Biology Institute, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Rd, Cambridge CB3 0WA, United Kingdom

ARTICLE INFO

Keywords:
Motion estimation
Motion model
Optic flow
Optical flow
Component level feature model
CLFM
Human motion processing
Intersection of constraints
IOC
Contrast invariant
Plaids
Random pixels
Biologically inspired motion processing
Gabor filters
Spatio-temporal energy models
Primary visual cortex
V1
MT
V5
Middle temporal

ABSTRACT

When humans (or robots) move through a scene, the scene can be represented as an optic flow (optical flow) field that contains vectors representing all of the movement within the scene projected onto a two-dimensional sensor. A simultaneous sample of these resulting vectors contain a good deal of information. A general model of motion estimation of local vectors would therefore be valuable. This paper addresses the estimation of motion vectors with uniform motion over a fixed time period. Previous reported attempts at computing motion estimation have been dominated by the machine vision community, however, these attempts are not specifically concerned with biological plausibility. Here, the author presents a model of motion estimation that computes motion based on filtering the moving image into sinusoidal responses varying in spatial frequency and orientation similar to the early visual responses found in human vision. Unlike similar spatio-temporal energy models "motion energy" is not computed. The model is mathematically explicit and simulated in MATLAB. It has been tested using over 7000 synthetic moving images with known veridical velocity (ground truth). These images range from sparse translating patterns containing 1 to 10,000 random pixels, to dense narrow band sinusoidal patterns. Simulation results show that the model correctly estimates motion trajectories between 84% and 100% angular direction error (within ± 2°), and displacement error (within ± 1 pixel). The results remain robust at different contrasts. In addition, a number of psychophysical and physiological results are examined in the context of the model.

1. Introduction

The term 'Optic Flow' was first introduced by J.J. Gibson in the 1940s (Gibson, 1950). Optic flow refers to a field of vectors representing the instantaneous flow of motion on the retina (or retinae) as an observer moves through an environment. 'Optical Flow' is now frequently used by computer scientists to refer to similar fields resulting from both human observers and non-human agents. The motion may be caused by the observer's movement through the scene, or by objects moving within the scene. Each vector in the field represents the speed and direction of motion at a specific position and at a specific time projected from the 3D scene onto the 2D retina. A simultaneous sample of these vectors contain important information about the objects in the scene, e.g. relative depth; shape of objects; or signatures of specific biological motion. Although there are a number of excellent approaches to computing optical flow published in the machine vision literature (Fortun, Bouthemy, & Kervrann, 2015), these have been largely unconcerned with biological plausibility, and often favour "gradient" based models for extracting individual velocity vectors (see Section 4.1 for a discussion of gradient models). Gradient models generally operate on the two dimensional image and therefore do not reflect typical characteristics of human motion processing. Human motion appears to be extracted following filtering operations that decompose the two dimensional image into sets of sinusoids. Therefore this paper describes a model of motion estimation of individual vectors that incorporates this well established result, and knowledge gained from vision science (physiological and psychophysical) that better represent the characteristics of human motion processing.

The most ubiquitous models that are more characteristic of human local motion estimation, are "spatio-temporal energy models". These models are based on "motion energy" that is computed directly from the amplitude or power of the object's underlying Fourier components (or wavelet equivalent (Adelson & Bergen, 1985)). This generally leads to the requirement of some form of normalising mechanism, because an important property of motion estimation is that it is has to be independent of the object's colour, brightness, or contrast. For example, it is not important to know that an oncoming vehicle is black or white, just what direction it is moving in and at what speed. See Section 4.2 for

E-mail address: lb16@cam.ac.uk.

https://doi.org/10.1016/j.visres.2018.07.003
Received 9 October 2017; Received in revised form 17 July 2018; Accepted 20 July 2018

a discussion of a more recent motion energy model. One of the strengths of the Component Level Feature Model (CLFM) (Bowns, 2011) is that, although it shares some of the established biological characteristics of spatio-temporal energy models, it is invariant to contrast, and therefore does not require a normalising mechanism. The implementation, however, was restricted to extracting direction up to a reflection for two component sinusoidal translating patterns (i.e. plaids), and therefore falls short of a general model of motion estimation. This paper uses some of the basic ideas underlying the CLFM but the model is enhanced so that (1) the model is mathematically explicit, and programmed in MATLAB. (2) absolute direction (and not just relative direction) is computed, (3) displacement (and therefore speed) is computed; (4) results are reported for a wider range of synthetic patterns, i.e. plaids that vary across a greater range of properties, and random pixel patterns with varying densities ranging from 1 pixel to 10,000 pixels. The next section provides equations describing each step of CLFM, together with a verbal description. Each step is introduced with a brief rationale of why the step is necessary and its relevance to the human visual system. The simulation results are then presented. CLFM is discussed in the context of other types of models, and a physiological interpretation of CLFM is suggested by comparing the specific steps of the model with those of a spatio-temporal energy model. Finally, frame by frame analysis of atypical plaids illustrates how CLFM can produce behaviour resembling that of human observers, and in doing so eliminate the need for supplementary mechanisms that spatio-temporal energy models have required.

2. Computing the component level feature model

Step 1: Filter each frame of the moving image with a bank of oriented Gabor filters.

Images can be broken down into sinusoidal components that vary in orientation, spatial frequency, contrast, and phase. Humans and other mammals have evolved neurones to extract sinusoidal patterns at different spatial frequencies and orientations at an early stage of visual processing to efficiently encode images (Campbell & Robson, 1968). Neurones in visual area V1 in mammals have been shown to be selective for spatial frequency and orientation (Hubel & Wiesel, 1962), and to respond to motion perpendicular to the orientation in both visual areas V1 and MT (Albright, 1984; Foster, Gaska, Nagler, & Pollen, 1985; Movshon & Blakemore, 1973). Oriented Gabor filters are often used to simulate receptive fields that respond to both spatial frequency and orientation, Adelson and Bergen (1985), Daugman (1984), Watson and Ahumada (1985).

$$r(x, y, \omega, \theta, t) = I(x, y, t) * G(\omega, \theta) \tag{1}$$

Eq. (1) shows a specific response to filtering a single frame t, from an image sequence t = [1,n], with a Gabor filter with parameters spatial frequency ω and orientation θ; with choice of $\omega = [\omega_1, \omega_m]$ and then for each ω there is a corresponding set of orientations $\theta = [\theta_1, theta_p]$. R will be used to describe the set of responses at time t.

Step 2: Select the two largest filter responses at the same spatial frequency but at different orientations for each frame.

It is common practice in vision research to assume efficiency through the use of lossy selection, i.e. abandoning unnecessary information. Subsequent processes only require two responses at the same spatial frequency and at two different orientations. However, as will become apparent, using more than two orientations should not change the outcome of the results.

$$(\omega 1, \theta 1) = \underset{\omega,\theta}{\operatorname{argmax}}(R(x, y, \omega, \theta, t)) \tag{2}$$

$$r_1(x, y, t) = R(x, y, \omega 1, \theta 1, t) \tag{3}$$

$$r_2(x, y, t) = \max_{\theta, \theta \neq \theta 1} R(x, y, omega1, theta, t) \tag{4}$$

For each position (x,y) in a given frame t find the spatial frequency and orientation, ($\omega 1$, $\theta 1$) at which the response is maximum (Eq. (2)), and denote this maximum response by $r_1(x, y, t)$ (Eq. (3)). Then find the orientation $\theta 2$ for which the response at spatial frequency $\omega 1$ has the second highest value, and denote this $r_2(x, y, t)$ (Eq. (4)).

Step 3: Extract the mean values from the maximal filter responses r_1 and r_2 for each frame.

The mean values from r_1 and r_2 are going to be used to compute velocity in accordance with the "Intersection of Constraints Rule" (IOC) (Fennema & Thompson, 1979). The IOC and how the mean values are used to compute the IOC will be described in the next step. The mean values of the filter responses are extracted by convolving r_1 and r_2 with a Laplacian of a Gaussian operator. This has the effect of aligning the mean values with the zero-crossings of the image function. These zero-crossings are then thresholded about the zero-crossing. Gabor filter responses are both sinusoidal and oriented, extracting the means therefore produces oriented line segments. These lines have double the frequency of the original spatial frequency of the filter responses r_1 and r_2. Most importantly, these lines are invariant to contrast.

$$Z_1(x, y, t_1^n) = (r1 * \nabla^2 G_\sigma)[-0.00001, +0.00001] \tag{5}$$

$$Z_2(x, y, t_1^n) = (r2 * \nabla^2 G_\sigma)[-0.00001, +0.00001] \tag{6}$$

Eqs. (5) and (6) produce two images Z_1 and Z_2 of thresholded zero-crossings by convolving the maximal responses r_1 and r_2 with the Laplacian of a Gaussian $\nabla^2 G_\sigma$ ($\sigma = 2$) and thresholding about the zero-crossing for each time frame t = [1,n], see Fig. 1.

Step 4: Compute the Intersection of Constraints (IOC) from Z_1 and Z_2.

The idea underlying the IOC was introduced in the context of a gradient approach to image motion processing (Fennema & Thompson, 1979), and later shown to be relevant to human perception (Adelson & Movshon, 1982; Bowns, 1996; Bowns, 2001b; Bowns & Alais, 2006; Quaia, Optican, & Cumming, 2016). The IOC is a solution to the "aperture" problem, whereby one-dimensional moving images do not have a unique velocity when viewed or processed through an aperture. The IOC provides a solution to this problem by using two one-dimensional varying images to produce a unique solution. In addition to solving the aperture problem for one-dimensional images, the IOC predicts the correct velocity for any moving image. Fig. 2 illustrates the IOC using a velocity space diagram with two velocity vectors. Each vector represents the velocity of a one-dimensionally varying periodic pattern translating perpendicular to its own orientation. The angle of each vector corresponds to the direction, the length corresponds to the speed. The IOC is computed from the intersection of two lines drawn perpendicular to each vector – the "constraint lines". Fig. 1 shows this for two displacements at $t = 1$, $t = 2$. This solution has been applied to velocity vectors derived from Gabor filters using "spatio-temporal energy" (Adelson & Bergen, 1985; Rust, Mante, Simoncelli, & Movshon, 2006). However, the vectors used in these papers are not invariant to contrast, and the mechanism is not mathematically explicit. The vectors proposed here are derived from the zero-crossings obtained in step 3, Z_1, Z_2, that are invariant to contrast. The lines in Z_1 and Z_2 have the same orientation of the filters responses r_1, and r_2, and act as the constraint lines for computing the IOC. The lines are displaced as a function of the phase of the input from which they are extracted, and therefore this displacement corresponds to the velocity vectors (or displacement vectors). Intersections between Z_1 and Z_2 displaced over time correspond to the velocity as determined by the IOC.