PERGAMON

Vision Research 38 (1998) 3633–3653

Vision Research

Motion analysis by feature tracking

M. Michela Del Viva, M. Concetta Morrone *

Istituto di Neurofisiologia del CNR, V. S. Zeno 51, 56100 Pisa, Italy

Received 10 December 1996; received in revised form 27 October 1997

Abstract

We have developed a two-stage model of motion perception that identifies moving spatial features and computes their velocity, achieving both high spatial localisation and reliable estimates of velocity. Features are detected in each frame by locating the peaks of the spatial local energy functions, as for stationary images (Morrone MC and Burr DC. Proc R Soc Lond 1988;B235:221–245.). The energy functions are calculated for different scales and orientations, and integrated within a temporal Gaussian window. The velocity of features is determined by the direction of maximal elongation of the energy in space-time, evaluated by calculating the three characteristic curvatures of the energy at each feature point. To circumvent the aperture problem, the energy maps are blurred in space by various amounts, and velocity is computed separately for each spatial blur. The Weber fraction of the local curvatures (curvature contrast) describes the spatio-temporal energy elongation at each feature point, giving a reliability index for each velocity estimate. For each point, the velocity of the spatial blur that yielded the highest curvature contrast was selected, with no further constraints, such as rigidity of motion. Dynamic recruitment of operators of different size allows maximum flexibility of the analysis, allowing it to simulate human visual performance in the detection of noise images, transparent motion, some motion illusions, and second-order motion.

Keywords: Motion modelling; Feature tracking; Motion transparency; Second-order motion; Spatio-temporal filters.

1. Introduction

One of the main tasks of a biological or artificial visual motion system is to determine the form of moving objects, as well as computing their velocity. Many current algorithms, based either on gradient models [1–6] or spatio-temporal energy models [7–10] may be successful in calculating velocity information (optic flow), but usually perform poorly in the analysis of the spatial structure of moving images. To obtain a stable velocity field, these models integrate heavily over space, losing spatial localisation. Subsequent reconstruction of form from segmentation of the optic flow, using appropriate Gaussian [11] or Laplacian filters [12] usually results in a large error in localisation. The complementary approach used by models [13–15] that first segment the visual scene (by edge detection, [16]) and then evaluate velocity by tracking the edges over time, yield much better localisation and can be used to analyse simultaneously the velocity and the form of objects [17,18]. However, the problem of tracking features in dense visual clutter is very challenging for artificial vision systems and usually requires some prior knowledge of the object shape, of its motion and environment [18–20].

An important problem that all motion models have to solve is the determination of the velocity, often an under-constrained or ill-posed problem [21,22]. An extreme example of the common inherent ambiguity in velocity information is given by the so-called 'barber pole illusion', where the motion of a circular spiral painted on the surface of a rotating cylinder is not perceived veridically, but as vertical motion [23]. This illusion results from the 'windowing' effect of the pole on which the stimulus moves, and reflects a general problem inherent to many computer algorithms, when the 'window' or 'aperture' of the operators used to extract motion is small compared with the moving stimulus. Of the several general solutions proposed to solve the aperture problem, two have been particularly successful [1,14,24]. Both solutions require the determination of the velocity component orthogonal to the prevailing local orientation of the signal, followed by

* Tel.: +39 50 559719; fax: +39 50 559725; e-mail: concetta@neuro.in.pi.cnr.it.

0042-6989/98/$19.00
PII: S0042-6989(98)00022-4

integration of this velocity vector over space to derive the direction and speed of the object motion. The two solutions differ in the spatial region of the summation: Hildreth's solution uses a weighted integration along the zero crossing of the image, usually corresponding to the object contours, while the original suggestion of Horn and Schunck uses an isotropic integration over the full 2D image. Both solutions present advantages and disadvantages (see Ref. [24], for discussion) and the performance of one over the other depends on the type of motion and of the particular contour. More recent computational models resolve the aperture problem using a battery of spatio-temporal filters. These models differ on how the various filter responses are combined to derive velocity: some compare the response for each spatial position [5,9]; others integrate the velocity field over the entire 2D image [8,25], or recursively select the response from large to small scales [26]. Some evaluate separately the three components of the flow field in terms of local translation, shear, rotation and expansion [3,11], while others generalise the integration approach of the orthogonal velocity using a regularisation theory [27]. All these widely different strategies are able to resolve the aperture problem with variable success rates, depending on the particular property of the visual scene. They also support the original idea that the solution of the aperture problem requires integration of the velocity or the spatio-temporal input signals. However, the necessary integration stage needed to solve the aperture problem will also restrict the performance of the same models for many other motion tasks, such as their ability to determine the shape of the moving object or to determine the motion of transparent objects or surfaces. Transparency represents a particularly strong challenge to theories of motion processing, usually requiring strategies of velocity integration that are quite different from those required for the aperture problem. Apart from transparency, there are several other motion functions that require specific and variable integrative processing of the local signals over space and time (see Ref. [28]): image segmentation requires the detection of spatial discontinuity in velocity without integration over the border, as well as the separation of points belonging to transparent surfaces (while integrating similar nearby velocities); optic flow processing requires extensive integration and the recognition of large-scale differentials such as divergent and rotary motion etc.; and as previously discussed, computing object motion requires the combination of locally ambiguous 'aperture-based' measurements of the motion of contour segments and surface markings over the object.

Recent psychophysical evidence suggests that the human visual system may solve the aperture problem by detecting the motion of 2D image cues like line-terminators [29], dots, or any other features with unambiguous motion and extending the resulting motion to all or part of the image. The presence of such features on an image can dramatically change the motion perception, like the illusory deformation of a translating quasi-linear curve [30–32] or the Barber-pole illusion [33]. These and other psychophysical results [34–36] also challenge the theory of intersection of constraints (IOC: see Ref. [37]) as a general strategy used by the visual system to resolve the windows problem, and suggest an alternative one based on the average velocity between ambiguous and unambiguous motion signals [29]. It is worthwhile to note that many of the computation models previously described [8,9,14,24] implement, with different strategies, the idea of the IOC to resolve the windows problem and would face difficulties in predicting or simulating the psychophysical results described here.

The initial requirement for all the behavioural applications of visual motion is a local measurement of motion. This process, which initially occurs in area V1 of the primate visual system, has now been quite thoroughly investigated in psychophysical, neurophysiological, anatomical and computational terms. There is now clear evidence of an early stage of motion processing mediated by elementary filters with a clear directional selectivity in space-time, well described by their elongated spatio-temporal receptive fields ([38–40]; see also Refs. [7,41–43]). These detectors probably correspond to the linear stage of motion energy mechanisms [7,8] or a more general Reichardt mechanism [44–47] subserving the computation of the local optic flow. In addition to these mechanisms, however, recent evidence suggests the existence of another initial stage of motion analysis that is not directional selective, probably subserving the perception of what it is usually referred to as 'non-Fourier' or (more correctly) as 'second-order' motion analysis [48–50]. This analysis requires an early spatial non-linearity, after which standard motion energy or Reichardt detectors would suffice to analyse the motion signal [48,51]. Some studies [52,53] also suggest that the second-order motion system is functionally independent from the first-order system. Other experiments have measured motion sensitivity to sequences of frames with congruent motion of zero-crossings at very different spatial scale [54,55]. These studies suggest that the visual system may also use a feature tracking scheme to analyse the motion of objects. It is worthwhile noting that both the feature tracking mechanisms and the second-order motion mechanisms need a spatial non-linearity, and could in principle be the same. At this stage, it is far from clear how all these hypothetical biological mechanisms for the analysis of the motion signal do really interact and co-operate to subserve motion perception. Indeed it is not certain that they are separate, but could be reduced to a single unified mechanism. In this respect, a computational model that

uses simple and biological plausible mechanisms could be an important tool to clarify the essential properties necessary to accomplish the various tasks of motion perception.

One of our present aims is to investigate if a single adaptive form of integrative processing could accommodate simultaneously the conflicting requirements posed by the aperture problem, transparency motion and global motion. Given the clear psychophysical evidence of the importance of the analysis of 2D feature motion, of the existence of an average velocity strategy and of the accurate detection of the form of the object in motion, we developed a computational model based on the feature tracking strategy.

Here we extend the model of local energy [56–59] to develop an algorithm that extracts the spatial features in a moving image, then computes their velocity (similar in principle to earlier feature-tracking models of Refs. [13,14]). The presence of an early spatial non-linearity and the use of non-directional temporal filters will confer, as demonstrated by Chubb and Sperling [48], the ability to detect and analyse second-order motion. The use of a battery of multiple-sized spatial integrators will confer the ability to achieve both fine spatial localisation and reliable estimation of velocity. Dynamic recruitment of operators of different sizes, based on the evaluation of the spatio-temporal orientation, will allow maximum flexibility to the analysis, with small integrators to recognise transparency and motion discontinuity, and large operators to disambiguate the aperture problem and to calculate the velocity of noisy images.

In the following section we will present and motivate separately these three main aspects of the model: spatio-temporal non-linear properties of the front stage; the parallel analysis, at different spatial scales, of the second stage; the concept of orientation contrast and its associated law for dynamic recruitment. In Section 3 we will show how the different properties will contribute to the ability of the model to fulfilling many of the demanding tasks imposed by our visual system.

2. Algorithm description and implementation

2.1. Front-end filtering properties

Tracking contours in a cluttered environment is a very difficult task. However, it could be greatly simplified if the input image, that usually comprises of complex luminance profiles with positive and negative peaks, could be regularised and transformed into a well behaved positive function. The task would become even easier if the tracking were restricted to only the local maxima of this well behaved function. The local energy model [56–59] performs an input transformation that fulfils both properties: the output is regular and positive, it eliminates redundant information (see Ref. [60]), considering only local maxima, that have been shown to coincide with salient visual features. The technical details of the implementation are given below, and further details and explanations are available in Morrone and Burr [61].

The local energy function $E(x, y, t)$ is computed by multiplying the image Fourier Transform $I_f(w_x, w_y, w_t)$ with pairs of band-pass oriented spatial operators in quadrature phase ($F_e(w_x, w_y)$ and $F_o(w_x, w_y)$: see Fig. 1). The quadrature phase constraint forces the use of oriented front-end filters. We used similar operators to those extensively employed for feature localisation in stationary images in previous research [57,61]. The shapes of the filters were given by the product of two Gaussians in the domain of log spatial frequency and orientation:

$$F_e(X, Y) = e^{-\left(\frac{(\ln X/f_p)^2}{2\sigma_X^2} + \frac{Y}{2\sigma_Y^2}\right)}$$

$$F_o(X, Y) = ie^{-\left(\frac{(\ln X/f_p)^2}{2\sigma_X^2} + \frac{Y}{2\sigma_Y^2}\right)}\sin(X)$$

where X and Y are respectively the frequencies parallel and orthogonal to the preferred orientation a of the filter (ie. $X = x\cos(a) + y\sin(a)$ and $Y = x\sin(a) - y\cos(a)$). The parameter f_p is the peak spatial frequency along the preferred orientation and σ_x and σ_y are the parameters governing the bandwidth of the filters along the two orientations. To simulate known selectivity of human motion psychophysical receptive fields [62,63], σ_x and σ_y were chosen to achieve a spatial frequency bandwidth of ± 0.28 logarithmic units and an angular bandwidth of $\pm 45°$ at half height. This yields receptive fields with a length to width ratio of about 1. Given the broad band of the filters, four preferred orientations (0, 90, -45 and 45°) were sufficient to span the orientation range and give satisfactory localisation of all oriented features, with considerable savings in computation (Fig. 1). To cover the range of spatial frequencies, three different sized filters were used: the smallest scale had a peak frequency at 1/4 of the input sampling frequency, the medium at 1/8 and the largest at 1/16 of the sampling frequency.

For each scale and orientation, a local energy function was computed as the square-root of the sum of the squares of the even and odd filter outputs [56], as schematically illustrated in the inset of Fig. 1. To keep the computational cost within reasonable limits and to deal with only one function at each scale, the spatial local energy functions, calculated separately for the four orientations, were added together. Given the broad band of the original filters, the resulting total energy function was a near isotropic transformation in space, obtaining a regular positive function appropriate for implementing tracking algorithms.

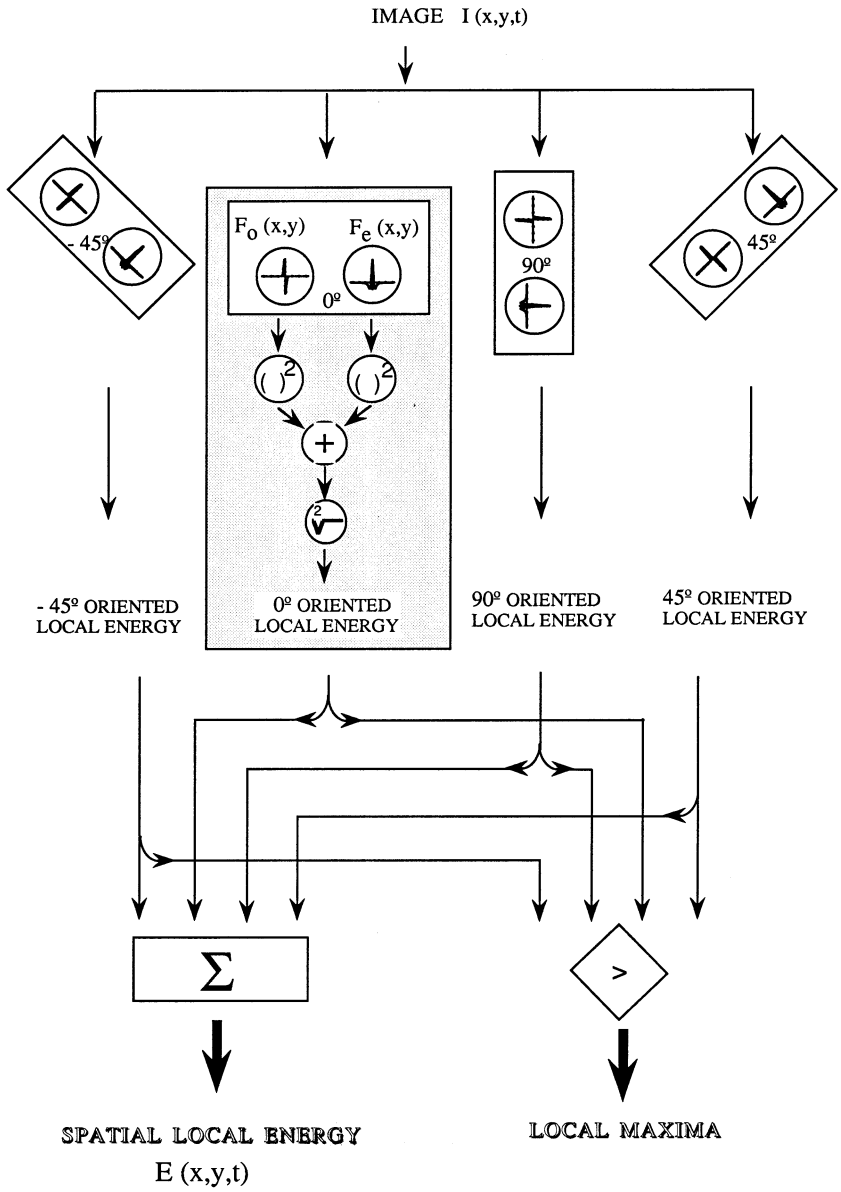

Fig. 1. Computation of Spatial Local Energy. The image $I(x, y, t)$ was convolved with pairs of even $f_e(x, y)$ and odd $f_o(x, y)$ symmetric linear spatial filters related via the Hilbert Transform. Four pair filters oriented along $0, 45, 90$ and $-45°$ were used. For each orientation, the local energy was calculated as the square-root of the sum of the square of the even and odd filters output. Features were marked by selecting the most responsive energy operator for the various orientations at each point (see text). In parallel, the oriented energy functions were added together to obtain the total local energy.

2.2. First stage: parallel pathway for feature detection

As with the static implementation of the model, the spatial features were located at each scale by searching for maxima of the local energy functions. The maxima were marked separately for each frame, using the following procedure: for each pixel, the most responsive operator amongst the four oriented filters was chosen, and the pixel was marked as a feature only if the point was a local maximum along the direction orthogonal to the orientation of the chosen filters. This strategy has been widely tested in many synthetic and natural images producing valid and reliable results [61,64] and it detects a more complete set of features than simply searching for local maxima in the total sum of all energy at different orientations. It also allows features to be classified in orientation and type (line, edge, etc.), that could be of advantage if the output of the model were used to categorise and identify the objects in the image, as in many robotics applications.

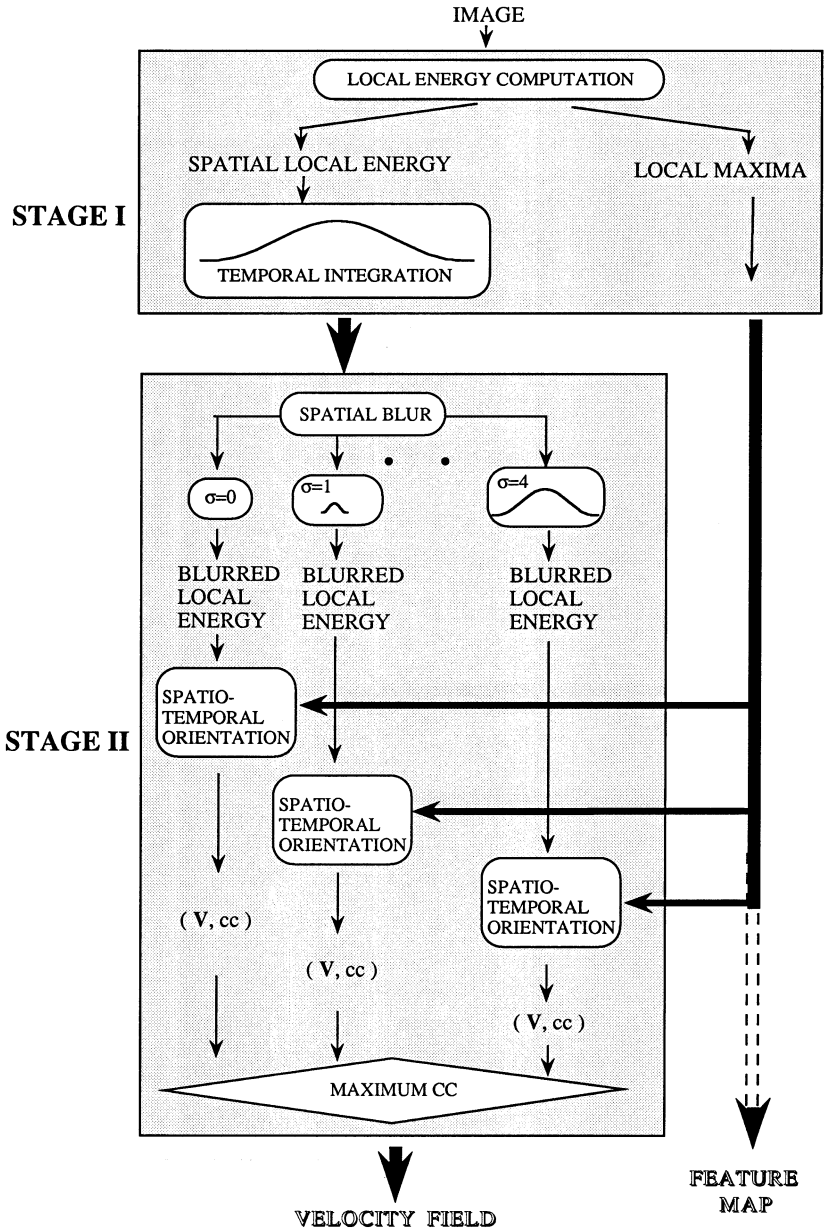

Fig. 2. General description of the algorithm. First Stage: The outputs of the front filters, illustrated in Fig. 1, were used to calculate the spatial local energy functions and to locate spatial features. The sum of all local energy functions was then convolved with a temporal Gaussian. Second Stage: The resulting spatio-temporal energy was blurred with four spatial Gaussian operators of different size. For each blur size, the velocity of the features is calculated from the spatio-temporal orientation of the total local energy at the point of local maximum (Hessian computation). At each point, the algorithm assigns the velocity estimate with the highest reliability.

2.3. First stage: parallel pathway for temporal integration

Temporal integration is a necessary process to compute motion. The temporal integration constant will determine much of the performance in velocity determination, like the selectivity to the speed and the robustness of the algorithm to under-sampling.

To detect simultaneously both stationary objects and those moving at high velocities, we used a low-pass rather than a band-pass temporal filter, simulating the action of the putative sustained temporal channel of the human visual system [65–67]. The motivation behind this choice is mainly biological, to study how well motion can be detected by a sustained not-directionally tuned system, such as the one operating in the primate visual system.

Temporal integration was achieved by convolving the total local energy of the various input frames with a temporal Gaussian function[1] (Fig. 2), with a time con-

[1] A Temporal impulse function described by a Gaussian is not casual. We verified the generality of the result using a casual filter given by: $f(t) = e^{0.5t} - e^{0.9t}$ where t is time (no. of frames). The results obtained using the two filters were very comparable.

stant of 1.5 frames. Temporal integration following the non-linear stage of the energy function offers two main advantages. Firstly, features can be detected and localised with high precision, independently of stimulus velocity (whereas integration preceding the localisation stage would blur moving stimuli, causing imprecision of localisation). Secondly (and more importantly), it allows optimal integration of temporal information between features that could change luminance profiles over time. For example, changing position and/or intensity of the light source in the scene could change the luminance profile of a moving border from a blurred step with frontal illumination to a blurred ramp or roof [64] with angular illumination. Temporal integration applied directly to the luminance profile would reduce the signal, but not if applied to the energy operator, that is invariant to such changes.

2.4. Second-stage: motion determination

The fourth step of the model is the computation of the velocity of each feature (Fig. 2), for which we developed an algorithm that follows the evolution of the feature over time. The previous parallel stages give as outputs the spatial position of each feature and the overall energy integrated over time, that contains the trajectory information necessary for the analysis. The instantaneous velocity of a travelling isolated feature point (not subject to the window problem) is, by definition, the local tangent to the feature trajectory. As features presumably are not created or destroyed from one frame to the next, the most probable matches between frames are features that have the same amplitude and shape of local maxima of the energy. Thus a moving feature will produce a temporal ridge of the energy function corresponding to its trajectory. Feature velocity can then be evaluated from the direction of this ridge, that corresponds to the direction along which the integral of the energy over consecutive frames is maximal. Biologically, we can imagine many oriented spatio-temporal receptive fields that compute the integral of the local energy: the receptive field that responds maximally will have the same orientation of the tangent to the trajectory and hence will provide the velocity of that particular feature point.

This method, although intuitive and biologically plausible, would in practice be complex and cumbersome to implement. An easier and more general solution can be obtained by analysing the local curvature of the energy function with differential geometry. In the previous example the maximally responding spatio-temporal elongated field corresponds to the direction along which the energy has less variation and more constant curvature. For the ideal translating point, the direction of zero curvature will be parallel to the orientation of the most responsive filter: curvatures along all other orientations will be significantly different from zero. Locally any continuous 3D function, such as the energy function, can be approximated by a parabolic equation characterised by the three principal curvature axes. We computed the three characteristic curvatures from the second order derivative matrix (the Hessian) of the local energy at each local maximum:

$$H = \begin{bmatrix} \frac{\partial^2 E(x,y,t)}{\partial x^2} & \frac{\partial^2 E(x,y,t)}{\partial x\,\partial y} & \frac{\partial^2 E(x,y,t)}{\partial x\,\partial t} \\ \frac{\partial^2 E(x,y,t)}{\partial y\,\partial x} & \frac{\partial^2 E(x,y,t)}{\partial y^2} & \frac{\partial^2 E(x,y,t)}{\partial y\,\partial t} \\ \frac{\partial^2 E(x,y,t)}{\partial t\,\partial x} & \frac{\partial^2 E(x,y,t)}{\partial t\,\partial y} & \frac{\partial^2 E(x,y,t)}{\partial t^2} \end{bmatrix}$$

After suitable rotation of the spatio-temporal bases (x, y, t), the Hessian matrix assumes a diagonal form. The corresponding eigenvalues l_0, l_1 and l_2 are the three characteristic curvatures at that point.

Usually (but not necessarily, given the non-maximum suppression procedure), the points where the Hessian are computed are points of local maxima and all three curvatures are negative. If one curvature, say l_0, is close to zero and much smaller in absolute value than the other two, the energy function at this point is ridge-shaped, clearly elongated along only one direction. The velocity is uniquely determined and given by:

$$V = \left[\frac{V_x}{V_t}, \frac{V_y}{V_t}\right]$$

where V_x, V_y and V_t are the components of the eigenvector corresponding to the minimum eigenvalue l_0. An example of such a simple case is a dot moving at constant velocity as illustrated before.

In many cases the energy is not ridge-shaped at the feature points, and no curvature is significantly close to zero or much smaller than the others. For example, a flashing dot will produce an energy function with all three curvatures equal. A three-dimensional space of solutions for the velocity are all equally possible, so no reliable velocity can be associated with the feature. For other images the local energy is ridge-shaped only at certain points, such as at the endings of a small bar when analysed with small scale operators. For the central points of the bar the energy is locally planar. For these points two of the three curvatures are equal and both close to zero: an infinite two-dimensional set of possible velocities, all lying on a plane, are possible. The latter case is a typical example of the aperture problem (for review see Ref. [24]).

To solve the aperture problem we employ detectors of various spatial dimensions. In principle the spatial scale that matches the particular length of the contour is the optimal size for the operator. However, this will also cause heavy integration of motion signals over the

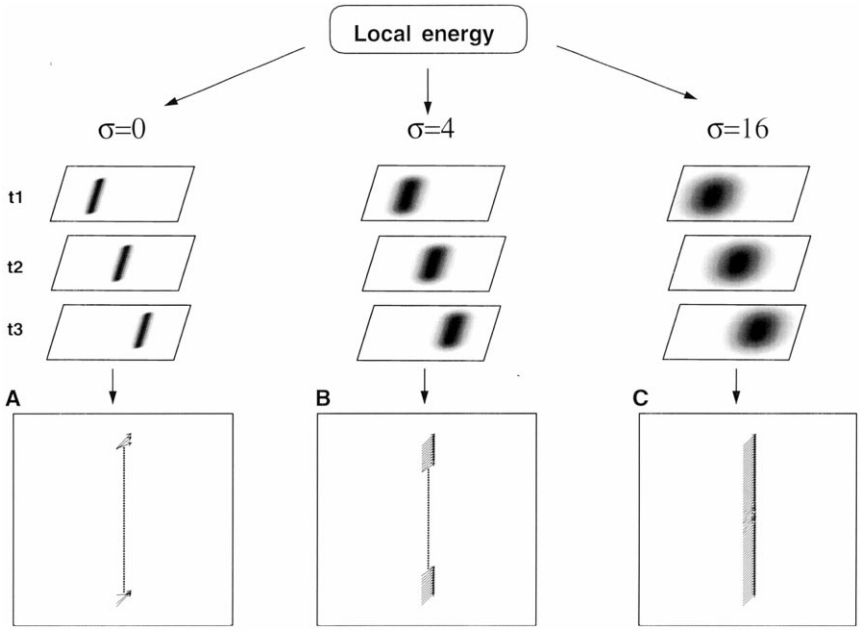

Fig. 3. Velocity field of a bar moving obliquely upwards and rightward at 45° orientation. Bar speed is 1.4 pixels/frame. Bar size: 1 pixel wide, 64 pixels long. Image size 128 × 128 pixels; medium scale front-end filters. (A), (B) and (C) show the velocity maps obtained for no blur, 4 pixel and 16 pixels Gaussian blurs respectively. The arrows indicate the direction of motion with the length proportional to speed. The filled circles show the feature points with a negative curvature contrast. Blurring the total energy in the space domain with progressively larger Gaussians ((B), (C)), allows the resolution of the motion of the central points of the bar.

same area, that can be a problem, especially for transparent motion. To optimise this trade off, we apply multiple scale filters before the computation of the velocity, but after the filtering front-end stage. We integrate spatially the local energy using four Gaussians of S.D. of 2, 4, 8 and 16 pixels (see Fig. 2). The shape and steepness of a ridge (the spatio-temporal elongation of the local energy surface at the point of local maximum) depends both on the particular image under analysis and on the size of the spatial blur applied to the energy. When only one type of motion is present, the spatio-temporal orientation biases of the energies will not be affected by the spatial blur. When several different motions are simultaneously present in the integration area, a heavy blur will reduce the energy orientation bias, but a light spatial blur may increase it, depending on the particular motion signals and on the local contour elongation.

2.5. Curvature contrast

Integration of local motion signals is essential to accomplish many motion tasks, and the demands for integration depend on nature of the task. Here we propose a new method of integration based on contrast evaluation, that has proven useful in many areas such as Michelson or RMS luminance contrast, and RMS cone contrast. We define a reliability index of energy elongation as the Weber curvature fraction (curvature contrast, CC):

$$CC = \sin(l_1 \cdot l_2) \cdot \sqrt{\frac{(|l_1| - |l_0|) \cdot (|l_2| - |l_0|)}{l_0^2}}$$

where l_0, l_1 and l_2 are the curvatures ranked in absolute value from the lowest (l_0) to the highest (l_2) and sin is the sine function. The index is very large when l_0 approaches zero and both l_1 and l_2 are much $> l_0$. This is the typical case of a small spatial feature (relative to the front-stage filter size) that moves at constant velocity. In general, a large CC implies a large distance between curvatures and hence a more elongated energy surface, with less ambiguity about the direction of velocity. The contrast is zero when either the two smallest curvatures are equal or all three curvatures are equal: in these cases, as previously discussed, the velocity cannot be uniquely determined. The contrast is negative if the medium or the largest curvature is positive: in this case the point corresponds to a saddle, not to a local maximum and it is not classified as a feature. All velocities associated with zero or negative contrasts are very unreliable and are thus rejected by our analysis. The curvature contrast is a good way to reveal the shape of the local energy surface: we used it to select the best velocity between the estimates at all spatial blurs, by assigning to each feature the velocity with the largest index (Fig. 2).

Figs. 3 and 4 show how this procedure can resolve the aperture problem and evaluate correctly the velocity of a thin bar moving obliquely rightward at 45°. Fig. 3 shows three successive frames of the total energy and the relative velocity map computed by the model in the

case of zero spatial blur (A), of a blur of 4 pixels (B), and of 16 pixels (C) (for simplicity the results of intermediate blurs are not shown). Velocity is represented at each labelled feature point by an arrow whose length and direction depict the velocity of motion. The dots represent the features for which the computed velocity has a negative curvature contrast, and hence considered unreliable. All points of the bar are correctly localised since they have been determined before the application of spatial blur. The number of features with a reliable velocity increases by increasing the spatial blur, as expected.

Fig. 4 shows how the velocity field is synthesised from the various maps at different blurs. Fig. 4 plots the curvature contrast for each pixel of the bar of Fig. 3, at the various spatial blurs. The velocity at the apices of the bar (pixels 0 and 60) is determined more reliably at small blurs, while the velocity of the central pixels (between 15 and 55) are determined more reliably with the largest blur, as expected from aperture considerations. For each pixel, the algorithm chooses the velocity associated with the largest contrast, hence the smallest uncertainty. The algorithm can determine correctly the velocity for all pixels. This result is particularly interesting, considering that the bar is longer than half the picture size, but the front-end linear filters were very small and the temporal integration brief (practically only two frames). For longer bars, the curvature contrast of even the largest blur (16 pixels) became negative for the central points. For these cases, larger blur would be necessary to resolve the velocity of all points. In the extreme case of an infinitely long bar, that the visual system perceives as moving in the direction orthogonal to its orientation, the model would never assign a reliable measure of velocity, since two curvatures will always be equally small. To solve these cases additional constraints need to be introduced, such as a default rule for a velocity locally orthogonal to the feature orientation.

Although the assumption of translation of a rigid object under constant illumination would require at least one eigenvalue to be zero, such a strong constraint was not introduced. We relaxed considerably this assumption and also accepted velocity estimates associated with large curvatures along the direction of motion, provided that the curvature was small compared with the remaining two (i.e. high curvature contrasts). This offers the possibility of evaluating successfully the motion of non-rigid objects, accelerating motion and the motion of objects that change contrast over time (see Section 3).

The input image was analysed in parallel with three front-end filters of different size, yielding three energy functions. How information from different spatial scales is combined to obtain a single description of the features of the image is still an open problem in vision research. Given that scale integration is not a central issue for the present model, we adopted the more general scheme, using the same strategy used to combine outputs from different spatial blurs: features that had spatial correspondence through scales were assigned the velocity value corresponding to the most reliable velocity (according to our definition) through all the possible front-end spatial scales and blurs. If there was no correspondence in position between features, they were all represented in the final velocity flow, with their most reliable velocity. For most of the simulations reported in this paper the results at various scales were kept separate, to help the reader to disentangle the contribution of each scale analysis.

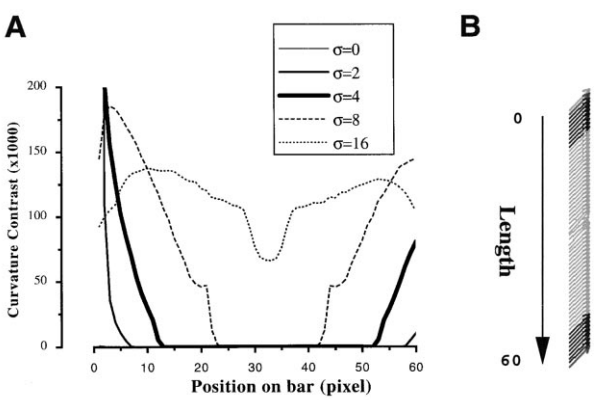

Fig. 4. Synthesis of velocity maps of the bar from the different Gaussian blurs described in Fig. 3. Curvature contrast as function of pixel position along the bar for different spatial blurs. For each pixel the algorithm chooses the velocity associated with the largest curvature contrast. The velocity of central points was detected correctly at the largest blur, while that of the extremities was detected correctly without blur.

3. Applications

All the synthetic stimuli were created with HIPS image-processing software [68] and simulations and statistics were run on a Silicon Graphics Iris Workstation. Real images were recorded by standard video camera (50 Hz, interlaced) and digitised at 256×256 pixels and 256 grey levels, using a video grabber board on a PC computer.

3.1. Tolerance to speed

Fig. 5 illustrates the performance of the algorithm in evaluating the velocity of a thin (1 pixel) bar drifting over a wide range of speeds from 0.25 to 16 pixels/frame. The motion of a thin bar drifting at high velocity is subject to the spatio-temporal under-sampling, resulting in a stroboscopic motion difficult to evaluate, especially for brief integration periods. This algorithm

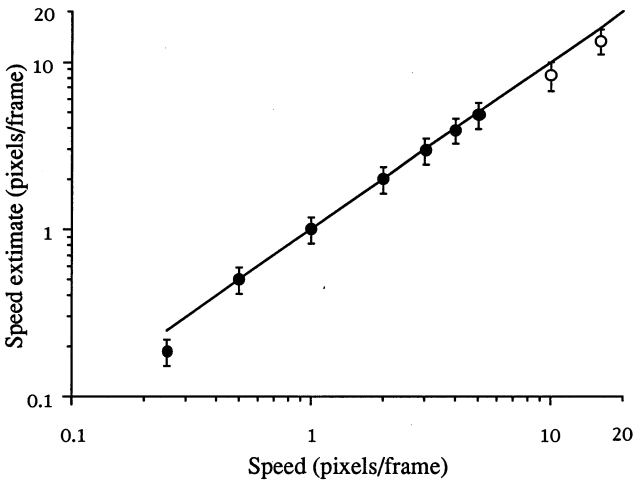

Fig. 5. Velocity estimates, averaged over all bar points, of a bar moving horizontally at various speeds. The bar is 32 pixels long and 1 pixel wide, the image is 128 × 128 pixels. The filled symbols plot the results for the small scale front-end filter, open symbols for the medium scale.

computed correctly the speed and the direction of motion of the line over a wide range of speeds. For speeds < 5 pixels/frames, the small spatial scale front-end operators were sufficient to evaluate the motion, even with the spurious signals introduced by under-sampling. For larger velocities, the small operators detected the spurious signals, while the larger scales still responded veridical to the motion.

3.2. Tolerance to noise contamination

One of the main goals of the algorithm was to obtain high precision in localisation of features and a good estimate of velocity in the presence of noise. The robustness of the algorithm to noise corruption was tested with a thin bar, 1/4 of the picture long (16 pixels), drifting rightward at constant speed of 1 pixel/frame under various levels of transparent Gaussian dynamic noise. Fig. 6(A) shows one frame of the test stimulus at a signal-to-noise ratio of 0.24 (ratio between the S.D. of the noise and the signal). As a reference point of performance, we measured the psychophysical threshold for judging the direction of motion of the bar. The input sequence was vignetted in time with a Gaussian of time constant equal to that of the temporal integration of the model. In these conditions our S/N threshold is about 0.07. Although the S/N ratio of the test stimulus (Fig. 6(A)) was only about three times above the psychophysical threshold, the model clearly detects the bar, amongst all the other features (Fig. 6(B)). The average curvature contrast for the pixels belonging to the bar is about 34000, while that for the noise is about 250, indicating a clear spatio-temporal elongation of the energy function for the signal. A further noise reduction could be achieved by gating the output using the curvature contrast. The image in Fig. 6(C) shows the velocity vector with a CC > 100 and in Fig. 6(D) a CC > 800: many noise points have been eliminated, while all points belonging to the bar are still present after a threshold of 800.

Fig. 7 shows the quantitative results of the performance of the algorithm for various amounts of noise contamination, for the high (Fig. 7(A), (B) and (C)) and medium scales (Fig. 7(D), (E) and (F)), corresponding to 1/4 and 1/8 the sampling frequency. The detection and the velocity estimation were very similar (for both scales of analysis) and reasonably good, considering that no CC threshold was used. For signals about twice the psychophysical detection threshold, the highest scale reduced the detection probability to 50% and gave an erroneous estimate of motion direction of 45°. However, the performance of the middle scale was sufficient to detect correctly signals that were close to the psychophysical threshold.

3.3. Motion transparency

Many optic flow algorithms need to integrate over space to evaluate velocity, losing the capacity to discriminate the motion of distinct objects that fall within the region of spatial integration. In all tests reported so far, the output of the larger Gaussian blur operator usually estimated the correct velocity. We then tested the algorithm in tasks requiring high spatial resolution. A good example is given by two or more non opaque objects moving over one-another in different directions, such as the random dot fields in Fig. 8(A) and (B). The dots in both fields move at the same speed (1 pixel/frame) but in opposite directions, as indicated by the arrows. Fig. 8(C) shows the velocity field obtained using the highest spatial scale at the front-end. Although the algorithm was allowed to assess velocity from the full set of possible spatial blurs, most of the velocity vectors were derived from the minimal spatial integration (no blur). The algorithm localised the dots well and calculated the correct velocity for most of them. It failed only at those points for which even the human visual system cannot give a correct answer, since these dots were colliding.

The perception of transparent motion depends on the density of dots: the greater the density the weaker the transparency effect. Fig. 9(A) shows the proportion of points for which the model evaluated correctly the velocity, as a function of dot density. Velocity is considered correct if within 2% of correct speed and 22° of correct direction. The performance of the model decreased roughly linearly with increasing dot density. However, even at densities $< 1\%$ more than half of the points were correctly classified. (Note that the density reported here is the true density, i.e. number of dots

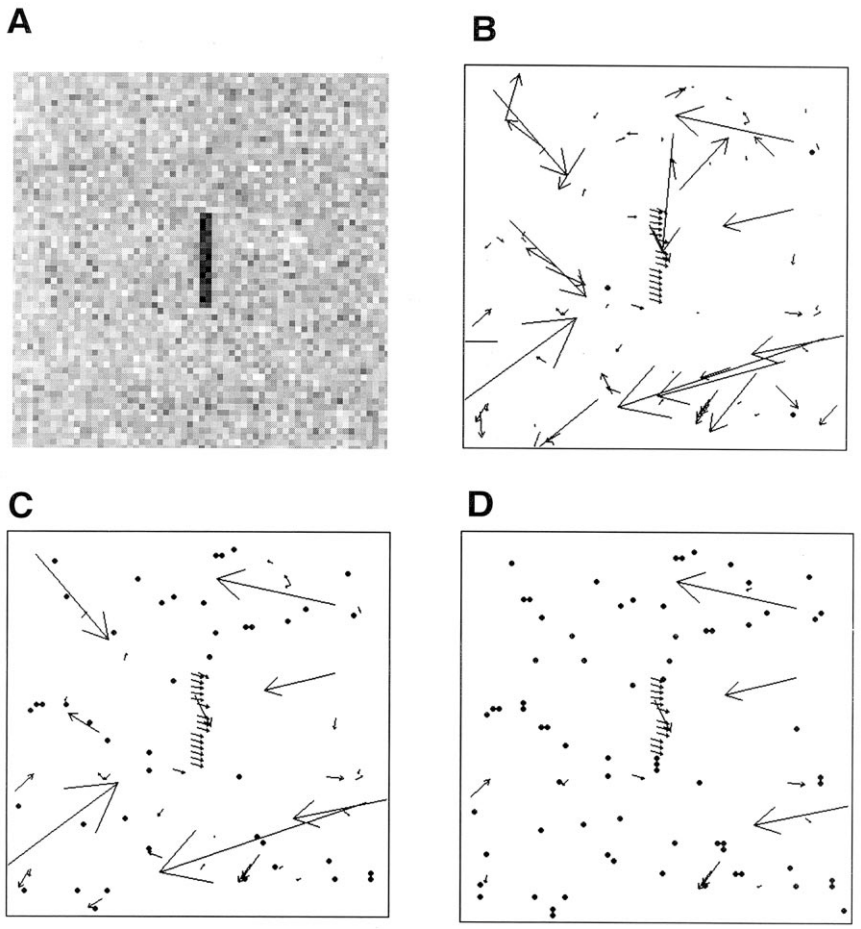

Fig. 6. Motion of a bar embedded in Gaussian noise. (A) One frame of a bar, 16 pixels long and 2 pixels wide, translating rightward at speed of 1 pixel/frame. Image size is 64×64 pixels. The signal to noise ratio is 0.24. (B), (C) and (D) Velocity maps produced by the algorithm using the highest scale front-end filters with a threshold of the curvature contrast equal to 0, 100 and 800 respectively.

over total number of pixels, 64×64 in this case). The psychophysical threshold to perceive transparent motion was about 10%, if the stimulus is presented within the same Gaussian temporal window as used for the simulations. At this density, the algorithm can detect correctly the velocity of 5% of dots. Similar results were also obtained for random dot field drifting at higher speeds (3 and 6 pixels/frames).

To increase the difficulty of the task, we also ran tests where the field was composed of vertical lines rather than random dots. Fig. 9(B) shows the performance of the algorithm when the lines were 4 pixels long. The algorithm identified correctly the position and the velocity of the lines and the performance was even better than for random dot field. A similar performance was obtained for lines with a length of 8 pixels (for stimuli of 64×64 pixels). It is important to note that for the line patterns the velocity was evaluated at intermediate spatial blurs, while for the dot patterns they were evaluated always at the smallest integration. The algorithm successfully recruits information from the more appropriate sized operator and chose, for the line, the operator that was large enough to overcome the aperture problem, while minimising integration between opposite direction of motion.

3.4. Motion capture

The complementary phenomenon of motion transparency is motion capture, where nearby motion signals influence each other producing illusory motion, like the induced motion of a dynamic noise superimposed on a drifting grating [69], or the motion of a stationary large disk superimposed on a field of translating random dots [70]. Again the perception of a capture phenomenon indicates the action of an integration of motion signals over a large region that embraces both the translating and the stationary stimuli: an integration condition opposite to that required to detect many examples of transparent motion. We tested the capability of the algorithm to simulate the illusory effect using a low contrast stationary disk that is perceived as following the translation of the superimposed random dot field (Fig. 10(A)), especially when seen at a distance or in

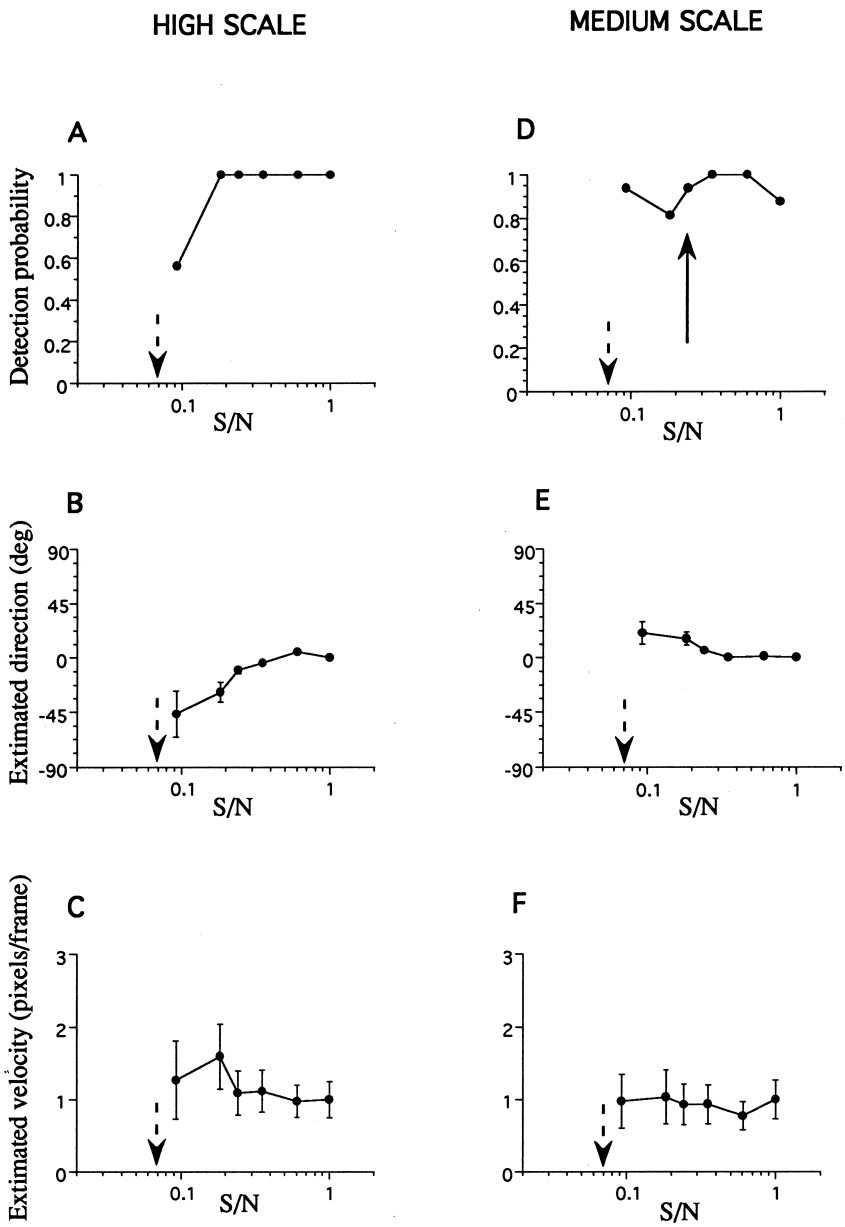

Fig. 7. Performance of the algorithm in detecting the bar of Fig. 6 at various signal to noise ratios. The results obtained with the small spatial scale are shown on the left, with the medium spatial scale on the right. (A) and (D) Probability detection, measured as the proportion of bar pixels correctly marked, as a function of signal to noise ratio. (B) and (E) Average estimates of velocity direction over all bar pixels, as a function of signal to noise ratio. (C) and (F) Average estimates of bar speed as a function of signal to noise ratio. The signal to noise ratio is measured as the ratio of the signal and noise S.D., S.E. indicated by bars. The arrows indicate the image and the results shown in Fig. 6. The psychophysical detection threshold is around 0.07 (see text for details)

peripheral vision. The results, illustrated in Fig. 10(B), show that more than 50% of the points corresponding to the contour of the disk move coherently with the dots, for the remaining contour point the motion is undetermined at this particular scale of analysis. It is important to note that the density and velocity of the random dot field of Figs. 8 and 10 are exactly the same. However, in one case the algorithm predicts transparent motion and in the other capture without any other a priori knowledge of the image or changing any parameter of the algorithm. In the transparent motion, a travelling dot generates a strong energy elongation along its trajectory, balancing the spread of the elongation generated by the other dots moving in the opposite direction. The stationary border is too weak to balance such contribution and the capture takes place. The curvature contrast is sensitive enough to detect the inducing elongation effect over the stationary border.

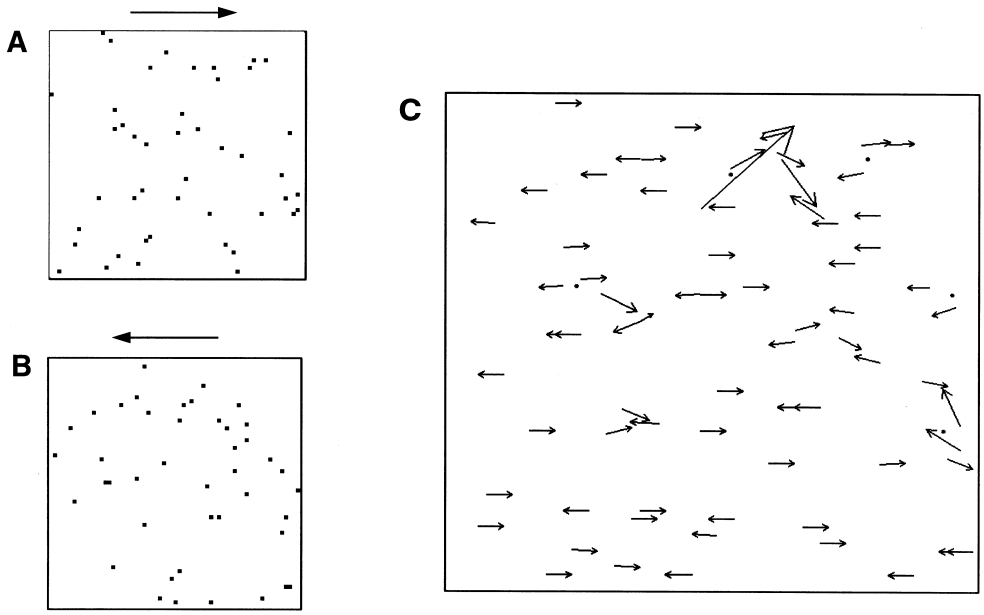

Fig. 8. Motion transparency. (A) and (B) show two frames of two random dot kinematograms translating in opposite directions (see arrows): the test stimulus is the sum, frame by frame, of the two. The speed of every point of each kinematogram is equal to 1 pixel/frame. (C) Velocity field obtained using small scale front-end filters.

3.5. Complex motion

The initial idea of determining the trajectory of a moving object by the direction of zero curvature relies on the assumption that motion can be approximated locally by translation of a rigid object. However, using the maximum curvature contrast to synthesise the velocity from the various spatial blur maps (rather than an absolute measure like the minimal curvature) allows us to relax considerably this assumption and open the possibility of detecting complex motion of objects that deform or change contrast. The curvature contrast is a measure of relative elongation of the energy function: velocity of features associated with high curvatures can still be calculated with precision, as long as the elongation along the direction of motion is more pronounced than in the other two directions.

3.5.1. Rotation, expansion and deformation

Rotating or expanding patterns (such as optic flow-fields) provide examples of violation of rigid translation. Fig. 11(A) shows a bar rotating around its centre at constant angular speed of 0.1 rad/frame, and Fig. 11(B) shows the output of the algorithm. Speed decreases correctly from the extremities to the centre of the bar and the direction of velocity is close to perpendicular to the bar as expected. The model derived the velocity field from the lowest spatial blur.

Fig. 11(C) shows an example of expanding motion, produced by a ring that increased in diameter at a velocity of 1 pixel/frame. The algorithm correctly located the ring (Fig. 11(D)) and determined a velocity equal in modulus for each point, that was always perpendicular to the ring.

For the rotation and expansion of the previous examples, the velocity direction is locally perpendicular to the border, and this could improve the motion detection. We tested more directly the ability of the algorithm to detect deformation motion by testing it with an ellipse whose vertical axis decreases at a rate of 0.27 pixel/frames while the horizontal axis remains constant at 14 pixels (Fig. 11(E)). With the exception of few points, the algorithm successfully detected the complex motion (Fig. 11(F)), particularly surprising given the heavy under-sampling of the input image (see the pixelation of Fig. 11(E)).

3.5.2. Illusory deformation by rotary motion

In some cases motion along a complex trajectory creates perceptual illusions. The visual system may solve erroneously the intrinsic under-constrained problem of velocity integration along area or border, such as in the Barber-pole illusion or a rotating logarithmic spiral [24] or translation of quasi-linear profiles [30,31]. We tested the performance of the algorithm on the spiral-illusion. Fig. 12(A) shows one frame of a logarithmic spiral rotating with constant angular velocity 0.16 rad/frame. When the endings of the spiral arms are concealed, the stimulus does not appear to rotate but to contract or to expand depending on the direction of rotation (clockwise or anti-clockwise). The real motion is illustrated in Fig. 12(B) with the output of the algorithm in Fig. 12(C). The detected motion is very similar to the illusory perception and different from the

real motion, except at the extremities. The endings of the spiral arms move along a rotary trajectory, while the inner points move along an expanding trajectory. Points located at intermediate positions follow the combination of the two motions. Also in this case, the most reliable velocity estimate came from the largest Gaussian blur, that integrates velocity over a distance comparable to the diameter of the spiral.

3.5.3. Second-order motion

The human visual system can extract coherent motion information from random dynamic images designed to have no luminance modulation along the perceived trajectory [48]. Fig. 13(A) shows a typical example of these stimuli: a drifting bar comprising a slice of dynamic random noise of same mean luminance as the background. The bar is drifting obliquely upwards in the direction of 45° at constant speed of 1.4 pixels/frame. Given that a different random noise pattern is generated at each frame, velocity signals derived from the analysis of grey levels and sensed by a linear system would result in a random flow field. Chubb and Sperling [48] measured this and more sophisticated stimuli and advanced the idea of a non-linear spatial mechanism (a full-wave spatial rectification) operating before the motion analysis. Our algorithm, based on a second-order spatial non linearity, performs well with this type of 'non-Fourier' or 'second-order' motion. Fig. 13(B) shows an example of the simulation result: the algorithm successfully marks as features more than 80% of the pixels of the bar and calculates their correct velocity (arrows). The performance for the detected pixels is comparable to that obtained for the luminance modulated bar (see Fig. 3(C)) even if the task here is more difficult. Also in this case the algorithm recruits velocity information mainly from the largest Gaussian blur.

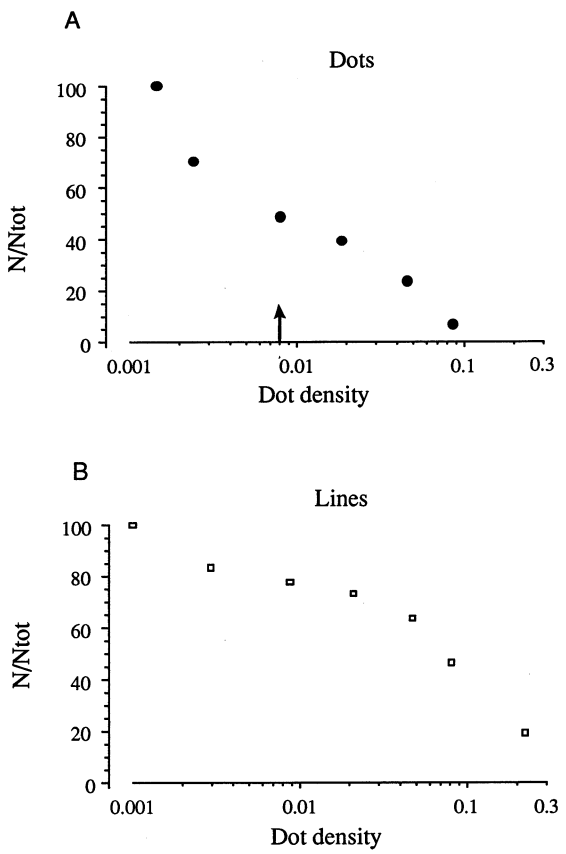

Fig. 9. Performance of the algorithm on motion transparency. (A) Percentage of points of the RDK for which the algorithm computed correctly the velocity in modulus and direction as a function of points density. The arrow shows the density corresponding to the example show in Fig. 8. The algorithm performance dropped to 5% at a density of 0.1 corresponding to the psychophysical threshold. (B) Results for a 4 pixels long bar moving in transparency. For (A) and (B), the density was calculated by dividing the number of pixels belonging to all bars or dots by the total number of image pixels (64×64).

3.6. Real images

A strong test for optical flow algorithms is how well they perform with real images, since they contain various sources of noise due to image quality and illumination. Fig. 14 illustrates results obtained for a digitised film. On the left are three consecutive frames (t1, t2 and t3) of the motion of three postal parcels drifting to the left on a conveyer belt, at the average velocity of 5 pixels/frame. The output of the model is shown on the right, using filters 8 pixels wide (the image is 256×256 pixels). All relevant features were correctly localised, both those associated with edges, such as the contours of the parcels, and those associated with lines, such as strings and markings. For the majority of features, velocity was calculated correctly and not influenced by spatial orientation: the velocity of features moving parallel to the orientation of their borders was detected as well as when the direction was orthogonal to borders. Only for a very few feature points did the velocity remain undetermined (dots), being associated with a negative curvature. To solve all the velocity fields, an analysis with lower front-end scales could be combined with the higher scales analysis, or a larger spatial blur could be employed. Alternatively, as the undetermined points are intermingled with points associated with correct velocity values, their velocity could easily be derived with a constraint on the continuity of velocity change along the border [1,14].

4. Discussion

We have presented an efficient algorithm that successfully locates salient features in a moving scene and computes their velocity. Similar to earlier edge-detection-tracking models [13,14], our algorithm first ex-

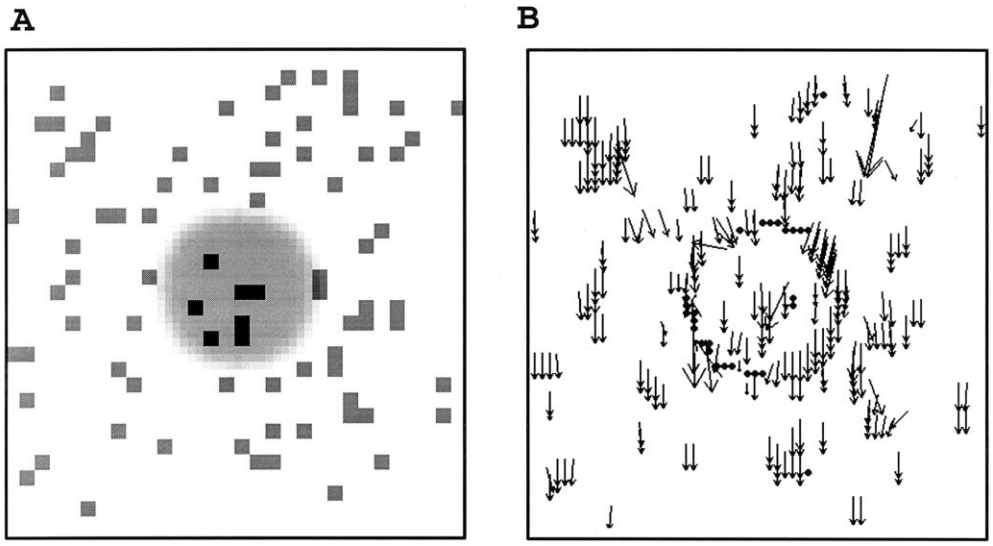

Fig. 10. Simulation of the motion capture effect. (A) A random dot field is translating at a velocity of 1 pixel/frame over a stationary blurred disk. The disk appear to move with the dots. (B) Velocity field obtained using small scale front-end filters. The border of the disk is detected to move coherently with the dots.

tracts visually salient features and then computes their velocity by tracking their energy over time. The feature-tracking approach to the analysis of motion has been somewhat neglected over the last few decades, except for a few recent studies that are mainly concerned with artificial and robot vision [17,18]. One reason may be the common belief in the visual community that feature-tracking is biologically implausible, given the correspondence problem. However, recent experimental evidence suggests that feature-tracking may be used by the human visual system [54,55] and the motion of 2D features can dramatically influence the motion perception of many ambiguous stimuli [29–31,33]. Here we show that feature-tracking models can fulfil many demanding tasks and simulate many aspects of human motion perception. Our evidence, together with the success achieved in robot vision by other types of feature tracking models, question whether the feature-tracking scheme really is implausible.

As for stationary images, features are detected by locating the peaks of the spatial energy functions, frame by frame. Local energy maxima have been shown to correspond to different types of visual features, such as borders, specularity, shadows, bars and combinations of them [56]. The organisation of the feature map corresponds closely to the structure perceived by human observers, and predicts many visual illusions [59,61,71]. One of the two parallel stages of the present model constructs a similar feature map for moving images. The other parallel stage performs a temporal integration of the energy and a subsequent evaluation of feature velocity by measuring the orientation in space-time of the energy peaks. This orientation corresponds to the direction along which the features move: the more ridge-shaped the energy surface is, the more reliable the velocity.

4.1. Comparison with current computational feature-tracking approaches

From the original idea of Marr and Ulmann of tracking zero-crossings over time, feature tracking models of motion have developed considerably [17,18,26]. The most successful models are those that take advantage of a priori knowledge of the contour, and make a Bayesian prediction of its evolution over time [19,20]. These models are very efficient and can easily be implemented in real time, but do not generalise well, nor simulate human vision (that does not rely on a priori knowledge). Anandan's model resolves the correspondence problem between two-frame motion using a correlation function and multiple scale analysis, with the assumption of a gradual variation of the velocity signal from the low to high spatial scales. Waxman's approach, that has some properties similar to ours, first detects the features, then substitutes them with a Gaussian profile obtaining a simple function suitable for resolving the correspondence problem. Both approaches perform a very good localisation and analysis of the form of the object in motion. However, these two approaches are not general enough to handle a wide variety of situations in motion analysis. In particular both have great difficulty in resolving transparency and/or capture. This is because Anandan's model recruits information from low scales (where the transparency signal is most confused) to high scales, and Waxman's model uses the same Gaussian function, independently of the contrast and the spatial frequency

content of the local feature. Both models blur information that should be kept separate to disentangle the transparent surfaces. In addition, Waxman's model limits the analysis to the component orthogonal to the local contour, so does not address the aperture problem. Both approaches are fundamentally based on the assumption of local translation, hence may find difficulty in handling deforming motion.

4.2. Comparison with models of the human motion system

Nearly all the current models of motion perception have a first filtering stage followed by a second stage where velocity is calculated by comparing the output of the various front-end filters. Let us examine and discuss first the differences in the properties of the front-end filters.

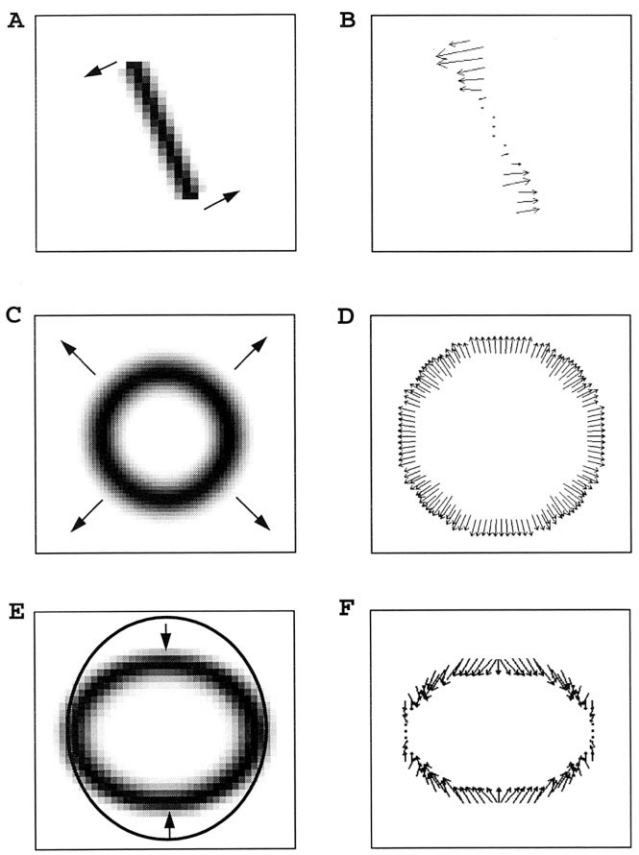

Fig. 11. Rotation, expansion and deformation. (A) and (B) A bar rotating around its centre, with angular velocity of 0.1 rad/frame and associated velocity map. The size of the image: 64×64 pixels. Bar length is 20 pixels. Velocity decreases from the extremities, approaching zero for the central point. (C) and (D) An annulus enlarging at a velocity of 1 pixel/frame and associated velocity map. (E) and (F) A deforming ellipse and the associated velocity map. All simulations were obtained using the smallest scale front-end filters.

4.2.1. First stage filtering properties

Many of the current computational models of optic flow [7–10,25] are based on the idea of elongated spatio-temporal receptive field operating at early stage of motion analysis ([38,39], for review see Refs. [41,42]). The models by Heeger [8] and Grzywacz and Yuille [9] use a battery of spatio-temporal operators in quadrature phase, each pair tuned to different spatial and temporal frequencies and spatio-temporal energy functions are computed from the response of each pair of operators. Despite the use of the same non-linear computations and similar nomenclature (spatio-temporal energy vs local energy), all these models are quite different from the present one. They use the energy computation to evaluate the power or the amplitude of the spatio-temporal frequency content of the stimulus for the various filters. Furthermore, to improve the frequency localisation of the power spectra, Heeger used very narrow band filters. We computed the energy function only in the spatial domain, using an infinite broad-band temporal filter. The aim was not to evaluate the distribution of the power spectrum of the stimulus, but the phase congruency in space. In this respect our approach is more similar to that of Fleet and Jepson [10] that also measures phase spread, by different means. Perhaps a simple example can illustrate this point. Two drifting sinusoidal gratings of close spatial and temporal frequency are not seen independently but generate beats, and their group velocity is perceived. Heeger's and Grzywacz and Yuille's models will never be able to detect the group velocity, but will simply detect a velocity average of the two components. Our approach, like the models of Fleet and collaborators [10,72] and of Chubb and Sperling [48], will succeed because the major features of the stimulus, corresponding to the point where the phases of the two gratings are equal, will move accordingly to the group velocity.

Beats are a particular case of contrast-modulated stimuli. One of the important features of the present model is that it analyses successfully such stimuli and more generally stimuli in which the light source changes position and/or intensity during motion. These kinds of stimuli cannot be perceived by models that perform a standard motion analysis directly on the image using oriented spatio-temporal linear filters or gradient algorithms [7,8,43,46,47]. Contrast-modulated stimuli and more generally second-order stimuli do not have any spatio-temporal correlation along the trajectory, so models with linear spatio-temporal filters at the front-end eliminate the motion information of this stimulus, that cannot be recovered by further processing. In real life, many stimuli pose the same difficulties as second-order stimuli, such as the motion of objects under variable illumination. The luminance profile of a moving border, that receives illumination from different angles along its trajectory, changes over time to assume

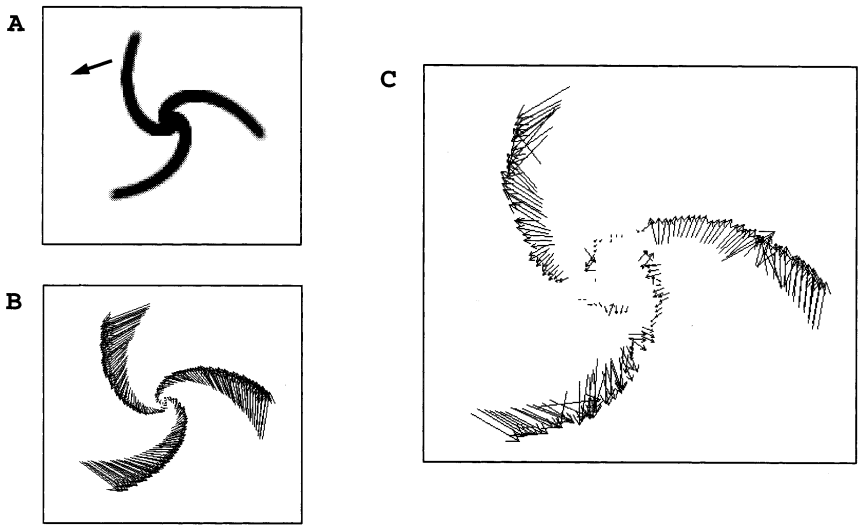

Fig. 12. Illusory deformation. (A) and (B) A logarithmic spiral rotating around its center, in the direction indicated by the arrow and its associated real velocity field. Image size 128×128 pixels, angular velocity 0.16 rad/frame. (C) Velocity map obtained from the small scale front-end filters and the largest Gaussian blur. At the extremities, the velocity is perpendicular to the radius, corresponding to rotation, while for the inner points a radial component (expansion) is present.

a continuum of shapes from a blurred step for frontal illumination to a blurred ramp or roof for angular illumination. Again, calculating the velocity of such stimuli poses problems for algorithms based on mathematical constraints such as constant luminance or constant gradient. However, the energy operator is invariant for such profiles and is therefore a useful operator to apply to determine the optic flow of the border.

In the last decade theoretical frameworks and psychophysical evidence supported the existence of a parallel motion system capable of perceiving second-order motion or motion of changing features [48,50,51,73]. Our model is similar in some respect to Chubb and Sperling's [48] model for non-Fourier stimuli. They argued the necessity of introducing a spatial non-linearity to detect the motion of non-Fourier stimuli. Their model uses full-wave rectification, applied after a front-end linear filtering obtained with separable band-pass spatial and temporal operators. We used the squaring non-linearity that is the core of the local energy model proposed by Morrone and Burr [56] for feature extraction from stationary images. Full wave rectifiers, rather than the squaring non-linearity associated with the computation of local energy, is not an efficient feature detector: it has a very low noise tolerance and generates spurious features. The other element conferring a good performance of the local energy is the use of a 2D base (even and odd spatial operators) to approach what is intrinsically a 2D problem: the detection of a variety of luminance profiles ranging from lines to edges. Chubb and Sperling's model uses only one type of operators and will face problems when detecting features that change from edge to line over time.

The stimuli that distinguish best between the action of the two models are those eliciting transparent motion. For example, two overlapping square-waves drifting in opposite direction are perceived as transparent. The present model would detect the border of the square-waves and would easily follow them over time. The only time where it could fail in detecting the two motions is at collision time. A rectification applied to the input image, before or after a temporal filtering, will produce a pattern containing component aligned along a stationary direction and a pure temporal modulation directions. Chubb and Sperling's model will not predict the presence of the two opposing velocities seen in transparency but the sum of the flickering gratings generated from the various square-waves components.

Other differences between our algorithm and that of Chubb and Sperling are in the temporal filtering. Chubb and Sperling used a front-end band-pass temporal filter, while our model uses a low-pass temporal filter. The use of a temporal band-pass filter with a prominent differentiating component was introduced by Chubb and Sperling to detect some subtype of second-order motion produced by difference in temporal modulation, such as drifting patch of random texture flickering at different rates. However, given that the temporal integration of the present model follows rather than precedes the spatial non-linearity, and that several blurs are used to calculate velocity, such images do not pose a particular problem for our algorithm, suggesting that a sustained system may detect a wide type of motion, including those originated from pure differential temporal modulation.

The feature-tracking model presented here achieves for second-order motion stimuli performances com-

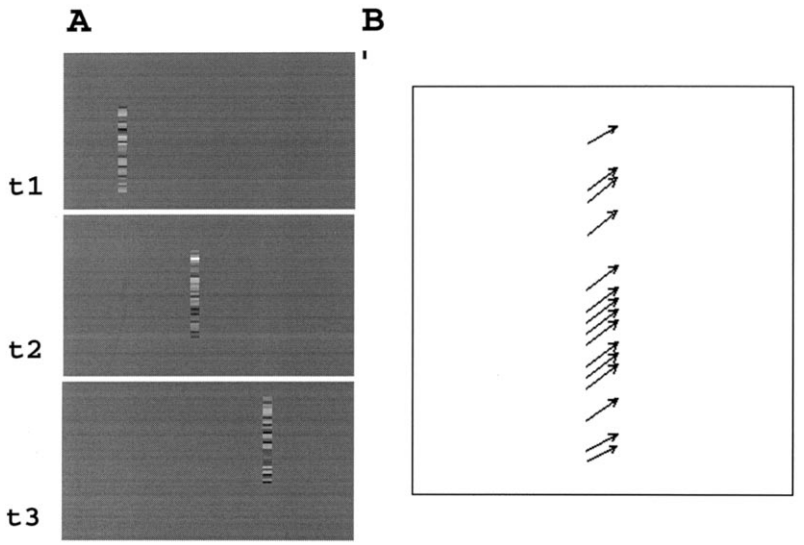

Fig. 13. Second-order motion (A) Three frames taken at three successive instants (t_1, t_2, t_3) of a dynamic noise bar translating obliquely rightward at 45° with a speed of 1.4 pixels/frame. The bar pixels change luminance randomly from frame to frame keeping average luminance constant. The image is 64×64 pixels and the bar length 20 pixels, bar width 1 pixel. (B) Output of the model using small scale front-end filters.

parable to those obtained for luminance modulated features, suggesting that these two kind of stimuli may share a common nature and may be processed by the same mechanism. Recent psychophysical evidence from our laboratory [74,75] indicates that the perception of transparent motion of both luminance modulated and non-Fourier stimuli can be explained adequately by considering the action of a feature-tracking mechanism. We would like to advance the idea that second-order motion is not analysed by a specialised mechanism, but is handled by a general feature-tracking motion system. Such a system could play a fundamental role in the perception of visual motion.

4.2.2. Second-stages properties

There are several major differences between the various models in how the velocity is derived from the output of the front-end filters. In approximating motion as local translation, Heeger [8] determines velocity by evaluating the prevailing orientation of the spatio-temporal image power spectra, using the energy transform. The main orientations of the energy function are then evaluated by the relative strengths of the responses of all the energy mechanisms over space. The work of Grzywacz and Yuille [9] is similar but considers only those mechanisms with receptive fields all centred on the pixel under analysis. Restricting the comparison to each pixel has the obvious advantages of detecting simultaneously the presence of different motions on the scene, and of relaxing the initial local translation hypotheses. In this respect our model shares some similarities with that of Grzywacz and Yuille: both use only a local comparison and both use locally, signals derived from various scales. However, one major difference is that, in our approach, the multiple scale operators are not applied directly to the image, but to the output of the front-stage filters. This offers the advantage of detecting interference between motions arising from signals at different spatial bands (that could annul each other in Grzywacz and Yuille's model), of detecting phenomena of cooperativity and capture (given that each local motion is spread over the surrounding region with a positive increase of energy) and solving the aperture problem. To deal with all these unresolved problems, Yuille and Grzywacz's [27] proposed a third stage of analysis, the motion coherence theory. This theory, as many others [1,14], tackles the aperture problem by restricting the analysis to the velocity components orthogonal to the border and attempts to recover lost information by subsequent integration. Unlike the strategy introduced by Hildreth [14,24] that performed a weighted integration of the orthogonal velocity component along the contour of the object, the coherence theory imposes a smoothing of the velocity field over a Gaussian region with the constraint that the measured and constructed velocity fields should be as close as possible for each estimate. Both Hildreth's and Yuille and Grzywacz's theories successfully resolve the aperture problem for simple images, but the integration constraints limit their performance in detecting motion discontinuity and transparency. An additional problem for both theories is that an erroneous estimate of the normal component will propagate to the all surrounding velocity estimates.

Here, to solve simultaneously all these problems, but without introducing a priori knowledge of the visual input, we developed a different strategy. A multiple and parallel estimate of velocity is obtained from the differ-

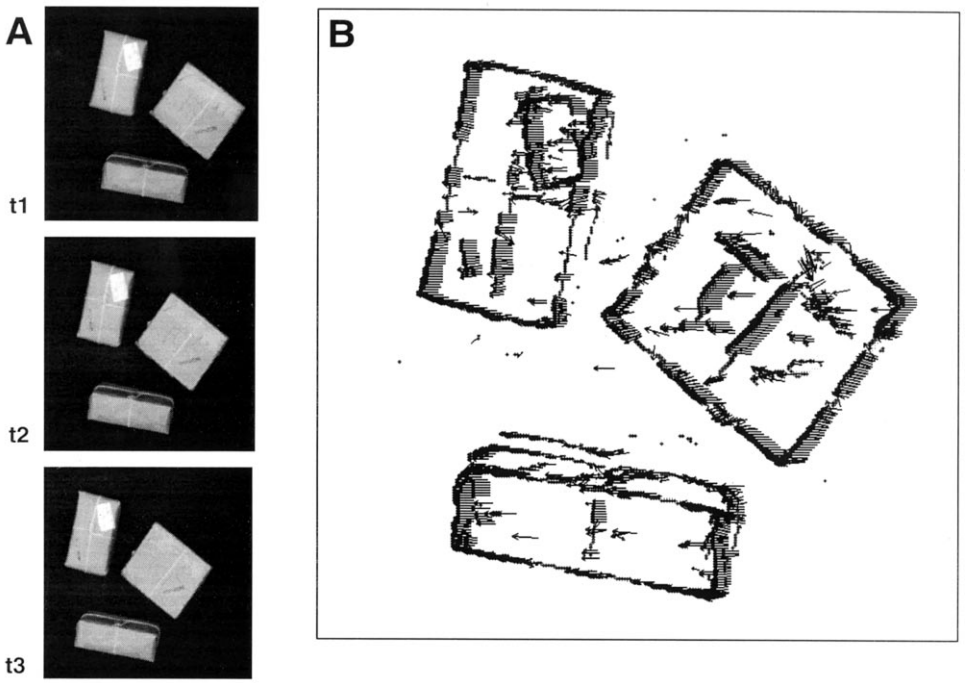

Fig. 14. Real images. (A) Three successive frames (t_1, t_2, t_3) of a test sequence containing three postal parcels on a conveying belt moving at constant velocity (around 5 pixels/frame). (B) Velocity map calculated by the model using medium scale filters. The image was 256×256 pixels. The features were correctly localised and, for the majority of points, both modulus and direction of the velocity were computed correctly.

ent blurs of the energy function by measuring locally its spatio-temporal orientation, with a confidence estimate of the orientation evaluation using the Weber curvature contrast. The higher the contrast, the more reliable the estimate of velocity. Contrast measures are widely used by the biological visual system to extract visual information, such as Michaelson contrast for luminance, Weber fraction for cone response, etc.. The same concept can usefully be extended to other dimensions of a visual stimulus, such as orientation. The divisive cross-orientation inhibition between neurones tuned to different orientations [76–78] could be instrumental in achieving an orientation contrast, that simulates many properties of visual primary neurones. Here we propose a similar contrast concept for spatio-temporal orientation: the product of the Weber fractions of the medium and high curvature, with respect to the minimum curvature.

An embryonic idea of relative elongation is present in the Reichardt detector [44,45]. The generalised Reichardt detector [47] is very similar to the spatio-temporal energy detector (see Refs. [7,79]), with the addition of a subtractive opponent stage between the response of two energy detectors tuned to opposite directions of motion. Schematically, Reichardt detectors measure the relative elongation of the energy between two opposite directions of motion. The curvature contrast measures the elongation along any arbitrary direction and includes a normalisation factor to allow for the detection of a weak differential change of velocity, as in the accelerating motion and deforming objects.

Elongation in the spatio-temporal domain was given by the characteristic curvatures using differential geometry. However, the mathematical operation of curvature evaluation could be easily implemented by known biological circuitry. Second-stage units that collect the stimulus response over long spatial (or cortical) ranges could be successfully employed to measure elongation. A contrast measure of the relative response of the units integrating along different trajectory could give an estimate of the curvature contrast. Units of this type have been proposed for the human visual system for the detection of piece-wise spatial signals over long contours [80,81]. Second-stage units integrating motion signals over large regions have also been demonstrated for both the human and the mammalian visual system [82–84]. It is plausible that second-stage units integrating spatio-temporal feature signal along various trajectory may also function in biological visual systems.

The concept of curvature contrast allowed us to relax greatly the assumption of rigidity of moving objects. The model has performed successfully with objects that accelerate, rotate or deform. Many models that rely on the assumption of constant luminance or energy profile over time, such as the gradient models [1,3–5], have difficulties in dealing with these phenomena, although they are very common in every day visual experience. In addition, the model is able to simulate the illusory motion associated with the recruitment of motion signals over long contours, such as the motion of a

rotating spiral that appears to contract or expand [24]. A similar illusion was used by Nakayama and Silverman [30,31] to address the question of whether the human visual system integrates along contours rather than isotropically over regions, and provided evidence for both types of integration, although integration along contours seems to prevail in many conditions. Recent psychophysical evidence [33] shows that isolated single points can influence the Barber-pole illusion and that the diffusion of unambiguous motion to ambiguous contour motion does not require a spatial overlap constraint and the spread may take place between different windows. The present algorithm performing an isotropic integration of the energy function surrounding the feature pixel, predicts all these perceptual effects.

From the above discussion it is clear that a wide range of models for motion detection are currently available and that all models are able to analyse some kind of motion signals. However, all of them fail for certain tasks, such as transparency, capture, second-order motion, beating, deformation or the inability to localise spatial structure. The present approach may perform some tasks less efficiently than other models (such as the detection of texture or of gradual images containing poorly defined features), but it is unique in its ability to detect simultaneously second-order motion, transparent and captured motion, deformations, analyse the form of the moving object and resolve the aperture problem. It is also important to stress that its performance in many tasks simulated qualitatively and quantitatively human motion perception. The definition of curvature contrast is the major innovation that confers such generality to the model, suggesting that the human visual system may use a similar concept to analyse motion.

Acknowledgements

We thank Dr David Burr for helpful discussion during this research and comments on the manuscript. This research was supported by the targeted grant Robotica 94.00926.pf67 from the Italian C.N.R.. Del Viva was supported by a PhD fellowship from the University of Pisa department of Physiology and Biochemistry.

References

[1] Horn BKP, Schunck BG. Determining optical flow. Artif Intell 1981;17:185–203.
[2] Nagel HH. Displacement vectors derived from second-order intensity variations in image sequences. Comput Vis Graph Image Process 1983;21:85–117.
[3] Verri A, Girosi F, Torre V. Mathematical properties of the two-dimensional motion field: from singular points to motion parameters. J Opt Soc Am 1989;6:698–712.
[4] Srinivasan MV. Generalized gradient schemes for the measurement of two-dimensional image motion. Biol Cybern 1990;63:421–31.
[5] Sobey P, Srinivasan MV. Measurement of optical flow by a generalized gradient scheme. J Opt Soc Am 1991;A8:1488–98.
[6] Johnston A, McOwan PW, Buxton H. A computational model of the analysis of some first-order and second-order motion patterns by simple and complex cells. Proc R Soc Lond 1992;B250:297–306.
[7] Adelson EH, Bergen JR. Spatio-temporal energy models for the perception of motion. J Opt Soc Am 1985;A22:284–99.
[8] Heeger DJ. Model for the extraction of image flow. J Opt Soc Am 1987;A4:1455–71.
[9] Grzywacz N, Yuille A. A model for the estimate of local image velocity by cells in the visual cortex. Proc Roc Soc Lond 1990;B239:129–61.
[10] Fleet DJ, Jepson AD. Computation of component image velocity from local phase information. Int J Comp Vis 1990;5:77–104.
[11] Verri A, Girosi F, Torre V. Differential techniques for optical flow. J Opt Soc Am 1990;7:912–922.
[12] Thompson WB, Mutch KM, Berzins, V. Edge detecton in optical flow fields. Proceedings of the National Conference on Artificial Intelligence (AAAI82), 1982;26–29.
[13] Marr D, Ullman S. Directional selectivity and its use in early visual processing. Proc R Soc Lond 1981;B211:151–80.
[14] Hildreth EC. The computation of the velocity field. Proc R Soc Ser B 1984;221:189–220.
[15] Marr D. Vision. San Francisco, CA: Freeman, 1982.
[16] Marr D, Hildreth EC. Theory of edge detection. Proc R Soc Lond 1980;B207:187–217.
[17] Waxman AM, Wu J, Bergholm F. Convected activation profiles and the measurement of visual motion. Proc IEEE CVPR, Ann Arbor, 1988;717–723.
[18] Blake A, Curwen R, Zisserman A. A framework for spatio-temporal control in the tracking of visual contours. Int J Comput Vis 1993;11(2):127–45.
[19] Isard M, Blake A. Contour tracking by stochastic propagation of conditional density. Proceedings of the European Conference on Computer Vision. Cambridge UK, 1996;343–356.
[20] Reynard D, Wildenberg A, Blake A, Marchant J. Learning dynamics of complex motions from image sequences. Proceedings of the European Conference on Computer Vision. Cambridge UK, 1996;357–368.
[21] Fennema CL, Thompson WB. Velocity determination in scenes containing several moving objects. Comput Graph Image Process 1979;9:301–15.
[22] Poggio T, Torre V, Koch C. Computational vision and regularization theory. Nature 1985;317:314–9.
[23] Wuerger S, Shapley R, Rubin N. On the visually perceived direction of motion by Hans Wallach: 60 years later. Perception 1996;25:1317–67.
[24] Hildreth HC. The neural computation of the velocity field. In: Cohen B, Bodis-Wollner I, editors. Vision and the brain. New York: Raven Press, 1990:139–64.
[25] Jacobson L, Wechsler H. Derivation of optical flow using a spatio-temporal frequency approach. Comput Vis Graph Image Process 1987;38:29–65.
[26] Anandan P. A computational framework and an algorithm for the measurement of visual motion. Int J Comput Vis 1989;2:283–310.
[27] Yuille A, Grzywacz N. A computational theory for the perception of coherent visual motion. Nature –74.
[28] Braddick O. Segmentation versus integration in visual motion processing. Trends Neurosci 1993;16:263–8.

[29] Rubin N, Hochstein S. Isolating the effect of one-dimensional motion signals on the perceived direction of moving two-dimensional images. Vis Res 1993;33:1385–96.
[30] Nakayama K, Silverman GH. The aperture problem I Perception of nonrigidity and motion direction in translating sinusoidal lines. Vis Res 1988;28:739–46.
[31] Nakayama K, Silverman GH. The aperture problem II Spatial integraton of velocity information along contours. Vis Res 1988;28:747–53.
[32] Ben-Av MB, Shiffrar M. Disambiguating velocity estimates across image space. Vis Res 1995;35:2889–95.
[33] Shiffrar M, Li X, Lorenceau J. Motion integration across differing image features. Vis Res 1995;35:2137–46.
[34] Lorenceau J, Shiffrar M. The role of terminators in motion integration across contours. Vis Res 1992;32:263–73.
[35] Mingolla E, Todd J, Norman JF. The perception of globally coherent motion. Vis Res 1992;32:1015–31.
[36] Wilson HR, Kim J. Perceived motion in the vector sum direction. Vis Res 1994;34:1835–42.
[37] Adelson EH, Movshon JA. Phenomenal coherence of moving visual patterns. Nature 1982;300:523–5.
[38] Burr DC. Human vision in space and time. Proceedings IUPS XV: 510.04 1983.
[39] Burr DC. Spatial and temporal selectivity of the human visual system. Neurosci Letters 1984; Suppl 15 S10.
[40] Burr DC, Ross J, Morrone MC. Seeing objects in motion. Proc R Soc Lond B 1986;B227:249–65.
[41] Burr DC, Ross J. Visual processing of motion. Trends Neurosci 1986;9:304–6.
[42] Burr DC, Ross J. Visual analysis during motion. In: Arbib MA, Hansen AR, editors. Vision, Brain and Co-operative processes. Boston: MIT Press, 1987.
[43] Watson AB, Ahumada AJ. A model of human visual-motion sensing. J Opt Soc Am 1985;A22:322–42.
[44] Reichardt W. Autocorrelation, a principle for evaluation of sensory information by the central nervous system. In: Rosenblith W, editor. Sensory communications. New York: John Wiley, 1961.
[45] Borst A, Egelhaaf M. Principles of visual motion detection. Trends Neurosci 1989;12:297–306.
[46] van Santen JPH, Sperling G. A temporal covariance model of motion perception. J Opt Soc Am 1984;A1:451–73.
[47] van Santen JPH, Sperling G. Elaborated Reichardt detectors. J Opt Soc Am 1985;A22:300–21.
[48] Chubb C, Sperling G. Drift-balanced random stimuli: a general basis for studying non-Fourier motion perception. J Opt Soc Am 1988;A5:1986–2007.
[49] Cavanagh P, Mather G. Motion: the long and short of it. Spat Vis 1989;4:103–29.
[50] Turano K, Pantle A. On the mechanism that encodes the movement of contrast variations. Vis Res 1989;29:207–21.
[51] Solomon JA, Sperling G. Full-wave and half-wave rectification in second-order motion perception. Vis Res 1995;34:2239–57.
[52] Lu ZL, Sperling G. The functional architecture of human visual motion perception. Vis Res 1995;35:2697–722.
[53] Vaina LM, Cowey A. Impairment of the perception of second-order motion but not first-order motion in a patient with unilateral focal brain damage. Proc R Soc 1996; B263:1225–32.
[54] Morgan MJ. Spatial filtering precedes motion detection. Nature 1992;355:344–6.
[55] Morgan MJ, Mather G. Motion discrimination in two-frame sequences with differing spatial frequency content. Vis Res 1994;34:197–208.
[56] Morrone MC, Burr DC. Feature detection in human vision: a phase dependent energy model. Proc R Soc Lond 1988;B235:221–45.
[57] Morrone MC, Burr DC. A model of human feature detection based on matched filters. In: Dario P, Sandini G, Aebischer P, editors. Robots and Biological Systems: Towards a new Bionics? Berlin: Springer-Verlag, 1993:43–64.
[58] Morrone MC, Owens R. Feature detection from local energy. Pattern Recognit Lett 1987;1:103–13.
[59] Burr DC, Morrone MC. A non-linear model of feature detection. In: Pinter RB, Nabet B, editors. Non-linear Vision. Boca Raton: CRC Press, 1992:309–28.
[60] Owens RA, Venkatesh S, Ross J. Edge detection is a projection. Pattern Recognit Lett 1989;9:233–44.
[61] Morrone MC, Burr DC. Capture and transparency in coarse-quantized images. Vis Res 1997;37:2609–2629.
[62] Anderson SJ, Burr DC. Receptive field length and width of human motion detector units: spatial summation. J Opt Soc Am 1991;A8:1330–9.
[63] Anderson SJ, Burr DC, Morrone MC. The two-dimensional spatial and spatial frequency properties of motion sensitive mechanisms in human vision. J Opt Soc Am 1991;A8:1340–51.
[64] Perona P, Malik J. Detecting and localising edges composed of steps, peaks and roofs. In: Osaka: Japan, editors. Proc Int Conf Comput Vis. New York: 1990;52–57.
[65] Kulikowski JJ, Tolhurst DJ. Psychophysical evidence for sustained and transient neurones in the human visual system. J Physiol Lond 1973;232:149–62.
[66] Burr DC. Temporal summation of moving images by the human visual system. Proc R Soc 1981;B211:321–39.
[67] Anderson SJ, Burr DC. Spatial and temporal selectivity of the human motion detection system. Vis Res 1985;25:1147–54.
[68] Landy MS, Cohen Y, Sperling G. HIPS: a Unix-based image processing system. Comput Vis Graph Image Process 1984;25:331–47.
[69] Ramachandran VS, Anstis SM. Displacement thresholds for coherent apparent motion in random dot-patterns. Vis Res 1983;23:1719–24.
[70] Ramachandran VS, Mc Lain K. Motion capture and induced motion can generate changes in perceived size and orientation. Soc Neurosci 1996;22:270.
[71] Ross J, Morrone MC, Burr DC. The conditions for the appearance of Mach bands. Vis Res 1989;29:699–715.
[72] Fleet DJ, Langley K. Computational analysis of non-Fourier motion. Vis Res 1994;34:3057–79.
[73] Wilson HR, Ferrera VP, Yo C. Psychophysically motivated model for two-dimensional motion perception. Vis Neurosci 1992;9:79–97.
[74] Del Viva MM, Morrone MC. Motion Capture and phase congruency. Invest Ophtalmol Vis Sci 1996;37:744.
[75] Del Viva MM, Morrone MC. The conditions under which opposing motion is seen in transparency or as flicker. Perception 1997;26:74.
[76] Morrone MC, Burr DC, Maffei L. Functional significance of cross-orientational inhibition: part I Neurophysiological evidence. Proc Roy Soc Lond 1982;B216:335–54.
[77] Morrone MC, Burr DC. Evidence for the existence and development of visual inhibition in humans. Nature 1986;321:235–7.
[78] Bonds AB. Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. Vis Neurosci 1989;2:41–55.
[79] Emerson RC, Bergen JR, Adelson EH. Directionally selective complex cells and the computation of motion energy in cat visual cortex. Vis Res 1992;32:203–18.

[80] Field DJ, Hayes A, Hess RF. Contour integration by the human visual system: evidence for a local 'association field'. Vis Res 1993;33:173–93.
[81] Moulden B. Collator units: second-stage orientation filters. In: Ciba Foundation, editors. Higher-order Processing in the Visual System. New York: Wiley, 1994.
[82] Morrone MC, Burr DC, Vaina LM. Second-stage detectors integrate motion along radial and circular trajectories. Nature 1995;376:507–9.
[83] Duffy CJ, Wurtz RH. Sensitivity of MST neurons to optic flow stimuli II Mechanisms of response selectivity revealed by small-field stimuli. J Neurophysiol 1991;65:1346–59.
[84] Graziano MSA, Andersen RA, Snowden RJ. Tuning of MST neurons to spiral motions. J Neurosci 1994;14:54–67.