A cortical architecture on parallel hardware for motion processing in real time

Karl Pauwels Laboratorium voor Neuro- en Psychofysiologie, K.U. Leuven, Leuven, Belgium

Norbert Krüger Cognitive Vision Laboratory, The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, Denmark

Markus Lappe Psychological Institute II and Otto Creutzfeldt Center for Cognitive and Behavioral Neuroscience, Westf. Wilhelms-University, Münster, Germany

Florentin Wörgötter Bernstein Center for Computational Neuroscience, Department for Computational Neuroscience III, Physikalisches Institut-Biophysik, Georg-August-Universität, Göttingen, Germany

Marc M. Van Hulle Laboratorium voor Neuro- en Psychofysiologie, K.U. Leuven, Leuven, Belgium

Walking through a crowd or driving on a busy street requires monitoring your own movement and that of others. The segmentation of these other, independently moving, objects is one of the most challenging tasks in vision as it requires fast and accurate computations for the disentangling of independent motion from egomotion, often in cluttered scenes. This is accomplished in our brain by the dorsal visual stream relying on heavy parallel-hierarchical processing across many areas. This study is the first to utilize the potential of such design in an artificial vision system. We emulate large parts of the dorsal stream in an abstract way and implement an architecture with six interdependent feature extraction stages (e.g., edges, stereo, optical flow, etc.). The computationally highly demanding combination of these features is used to reliably extract moving objects in real time. This way—utilizing the advantages of parallel-hierarchical design—we arrive at a novel and powerful artificial vision system that approaches richness, speed, and accuracy of visual processing in biological systems.

Keywords: dorsal visual stream, optical flow, binocular disparity, egomotion, object motion, graphics processing unit

Citation: Pauwels, K., Krüger, N., Lappe, M., Wörgötter, F., & Van Hulle, M. M. (2010). A cortical architecture on parallel hardware for motion processing in real time. *Journal of Vision, 10*(10):18, 1–21, http://www.journalofvision.org/content/10/10/18, doi:10.1167/10.10.18.

Introduction

The survival of many animals relies fundamentally on their ability to extract the movement patterns of all moving objects in a scene. This is vital for predator and prey in the wilderness, but even in our more orderly urban settings also drivers and pedestrians must be able to correctly interpret motion in daily traffic. However, motion analysis has to date remained a major challenge for artificial vision systems as it pairs a very demanding computational problem with the need for accuracy, density, and—especially—speed. Any useful movement analysis must be performed online and in real time! To be fast, image analysis to date is often restricted to sparse feature maps (e.g., edge maps; Canny, 1986; Lowe, 2004), but reliable motion estimates are hard to obtain on limited features. This is due to a prevalent correspondence problem: Frame-to-frame matching of such restricted, sparse, image structures is often impossible and leads to wrong motion estimates. Full, pixel-parallel computations would offer a solution. The computational load for this, however, is tremendous. The difficulty lies in the disentangling of egomotion from independent motion in dense images. Imagine, for example, two cars following each other with the same speed. In the trailing car, raw optical flow signals (Gibson, 1950; Koenderink & van Doorn, 1987) are of little use, since the stationary scene rather than the leading vehicle seems to move, and other information is required to extract the motion and the outlines of the car in front. The optical flow has to be parsed into components arising from egomotion and independent motion by subtracting the egomotion induced flow from the total flow (Rushton, Bradshaw, & Warren,

doi: 10.1167/10.10.18 Received December 3, 2009; published August 18, 2010 ISSN 1534-7362

2007; Rushton & Warren, 2005; Warren & Rushton, 2007, 2008, 2009a, 2009b). This apparently simple subtraction reveals itself as a difficult multi-faceted computational problem, not the least because virtually all algorithms for egomotion estimation rely on a stationarity assumption of the scene, which is violated by independently moving objects such as humans or cars. The mutual dependence between egomotion and independent motion can be overcome by an iterative approach, but this requires on the order of several tens of billion calculations per image frame spread out across the analysis of many different visual features.

In the current study, we solve this problem by utilizing fine-grained, pixel-parallel, and thus dense processing of many neuron-like entities on a hardware architecture matched to the algorithm. Thereby, we emulate processing stages and interactions along the visual areas of the dorsal pathway (Orban, 2008). The detailed biophysics of this system are not relevant for this problem. Therefore, computations take the shape of filtering operations similar to those suggested for cortical cells (Burt & Adelson, 1983; Daugman, 1985; Granlund, 1978), which are ideally suited for implementation on a graphics processing unit (GPU) hardware architecture (Wong, Leung, Heng, & Wang, 2007).

Related work

The problem of segmenting independent motion from egomotion has received extensive study from the psychophysics, neuroscience, computational neuroscience, robotics, and computer vision communities. We briefly review a few representative studies from each of these domains.

Psychophysics

The work by Gogel (1982) examined the role of extra-retinal information in the segmentation problem. Gogel's "theory of phenomenal geometry" (Gogel, 1990) describes the geometry of perceived shape and perceived movement of objects in terms of perceptual factors, highlighting the importance of perceived distance. According to Gogel (1990), perceived distance is important even in the absence of visual distance cues. In this case, equidistance and specific distance tendencies are used for its determination. Wexler, Panerai, Lamouret, and Droulez (2001) also looked at the role of extra-retinal information and discovered differences between moving and stationary observers when presented with the same visual stimulus, indicating that action influences depth perception. Wexler, Lamouret, and Droulez (2001) further showed that the visual system prefers stationary (in a world-fixed reference frame) over rigid interpretations. Dyde and Harris (2008) manipulated retinal and extra-retinal egomotion cues available to the observer and found an effect of errors in perceived egomotion on the judgment of object movement. Royden and Hildreth (1996) and Warren and Saunders (1995) also looked for interactions between egomotion and object motion and found that moving objects cause a small bias in the perceived heading direction when the object crosses the observer's path, indicating that the human brain does not segment egomotion from object motion in these circumstances. This misjudgment has been assigned a possible role of aiding in evading or intercepting animate objects. A heading model using motion-opponent operators was proposed to explain these findings (Royden, 2002). The work from Brenner (1991) and Brenner and van den Berg (1996) suggested that the judgment of object velocity depends neither on distance information nor on detailed knowledge of egomotion, but rather on the movement of the most distant structures in the scene, both in monocular and binocular situations. Warren and Rushton (2009a) pointed out that the stimuli used in these studies provide only limited information regarding egomotion and instead demonstrated that depth information (monocular or binocular) is important. They showed that the judgment of object movement is poorer when depth order is specified by parallax alone but approaches the performance achieved with binocular depth information when additional monocular cues are added. Rushton and Warren (2005) proposed the "flow-parsing hypothesis" according to which the brain parses retinal motion into the components due to egomotion and object motion by globally subtracting the optical flow due to egomotion. Specific tests ruled out a strictly local processing mechanism such as local motion contrast (Warren & Rushton, 2009b). Further evidence was obtained for the existence of the flow-parsing mechanism in the assessment of object trajectory in 3D (Warren & Rushton, 2007) and 2D (Warren & Rushton, 2008) scenes as well as during visual search for moving objects (Rushton et al., 2007).

The work by Rushton and Warren provides a psychophysical justification for our system, which also employs the flow-parsing mechanism as a purely visual solution for the segmentation problem.

Neuroscience

In macaque monkey, the medial superior temporal visual area (MST), further distinguished in dorsal and ventral parts (Komatsu & Wurtz, 1988), has been found to be of particular importance for optical flow processing. Neurons in the dorsal part (MSTd) have large receptive fields that are sensitive to optical flow components and are believed to be involved in the analysis of egomotion (Duffy & Wurtz, 1995). They are selective to heading, even in the presence of simulated eye movements without

extra-retinal input (Bremmer, Kubischik, Pekel, Hoffmann, & Lappe, 2010). Neurons in the ventrolateral part (MSTl) respond to small moving stimuli and appear to be specialized for the analysis of object motion in the scene (Eifuku & Wurtz, 1999; Tanaka, Sugita, Moriya, & Saito, 1993). In more recent work, Logan and Duffy (2006) found strong interactions between optical flow and independently moving object signals in MSTd providing support for the egomotion misjudgments described by Royden and Hildreth (1996).

Our model includes both an egomotion stage that corresponds to MSTd and an independent motion stage that is selective to object motion in world-centered coordinates, similar to MSTl cells (Ilg, Schumann, & Their, 2004).

Computational neuroscience

Zemel and Sejnowski (1998) trained an unsupervised neural network to develop a compressed representation of MT neuron activity while presented with scenes involving egomotion and object motion. The model cells obtained a tuning similar to that of MST neurons and represented a motion hypothesis of a scene element in observer-centered coordinates. The cell responses could be used to estimate heading or independent motion, albeit the latter only in simple situations since distance information was not considered in the experiments. More recently, Browning, Grossberg, and Mingolla (2009) combined various stages of the visual system into a model for navigation and obstacle avoidance. Their model responds to differential motion signals that can result either from stationary objects at a depth discontinuity or from truly independently moving objects. Since the model does not use distance information (e.g., from disparity), it cannot discriminate between these two causes.

Unlike the proposed model, both these models are severely restricted in the types of independent motion they can detect. The inclusion of distance information and the global flow-parsing mechanism allow us to overcome these limitations.

Robotics and computer vision

Many approaches have been developed in the robotics community toward the problem of Simultaneous Localization and Mapping (SLAM; Durrant-Whyte & Bailey, 2006). SLAM techniques detect and use a sparse set of environmental landmarks to build and update maps of the environment while at the same time keeping track of the robot's current location. Initially, these techniques relied on sensors other than cameras (e.g., laser rangefinders), but more recently also visual SLAM techniques have been developed (Davison, Reid, Molton, & Stasse, 2007). The recent inclusion of visual odometry techniques (Nister, Naroditsky, & Bergen, 2006) in visual SLAM (Williams & Reid, 2010) demonstrates a further convergence between the robotics and computer vision domains. Visual odometry techniques track a large number of features over a small number of frames and use global optimization methods (bundle adjustment) to obtain accurate egomotion estimates. Both visual SLAM and visual odometry can operate on monocular and stereo input but, originally, neither considered the presence of moving objects, although this had received study in nonvisual SLAM (Wang, Thorpe, Thrun, Hebert, & Durrant-Whyte, 2007). In visual SLAM, this problem has recently been approached by tracking objects of which a 3D model is available (Wangsiripitak & Murray, 2009). In a similar fashion, visual odometry has been combined with appearance-based object detection and tracking (Ess, Leibe, Schindler, & Van Gool, 2009; Leibe, Schindler, Cornelis, & Van Gool, 2008). Both extensions greatly increase the robustness but require prior knowledge of the moving targets' appearance or 3D shape. The proposed motion-based system is not subject to this restriction.

More closely related to the work presented here are computer vision techniques that do not consider the navigation aspect and focus entirely on the segmentation problem. A distinction can be made between methods that detect independent motion by removing a global component due to egomotion (as in the flow-parsing approach) and methods that identify clusters of consistent (3D rigid) motion.

Clustering methods are usually computationally expensive and limited to a small number of objects that each occupy a large part of the image. One of the first monocular clustering techniques was proposed by Adiv (1985). This method first clusters the optical flow into regions consistent with rigidly moving planar surfaces and then groups these regions into objects with a common rigid body motion. Motion clustering has also been demonstrated on the basis of sparse stereo feature sets (Demirdjian & Horaud, 2000; Wang & Duncan, 1996). Closely related is the work on model selection, which jointly considers clustering the features, selecting the motion model (with complexity depending on the available evidence), and estimating the model parameters (Schindler, Suter, & Wang, 2008; Torr, 1998).

Thompson and Pong (1990) suggested a number of different approaches (monocular, active, binocular) for independent motion detection rather than motion clustering. One of the binocular techniques they propose proceeds by evaluating the depth/flow constraint (Equation A16) for a number of points and signaling independent motion in case of inconsistency. A number of techniques rely on component motion (normal flow) rather than pattern motion (optical flow) to simplify the correspondence problem. These methods typically require either known egomotion, or additional assumptions or cues (e.g., disparity) to enable egomotion estimation. Examples are the technique by Nelson (1991), which

assumes known camera motion or animate (rapidly changing) independent motion, the technique by Sharma and Aloimonos (1996), which assumes active control of the camera, and the technique by Argyros, Trahanias, and Orphanoudakis (1998). Franke and Heinrich (2002) achieved real-time performance in driving situations by using a simple optical flow algorithm and assuming a fixed translational component of egomotion. There are also numerous detection techniques that model the scene in terms of deviations from a dominant plane (Irani & Anandan, 1998; Sawhney, Guo, & Kumar, 2000; Yuan, Medioni, Kang, & Cohen, 2007). These "plane + parallax" approaches greatly simplify correspondence and egomotion estimation but assume that a plane is present and remains dominant for the duration of the sequence. Ogale, Fermuller, and Aloimonos (2005) took an alternative approach and used occlusions as a static (ordinal) depth cue, thereby increasing the number of motion classes that can be detected by a monocular system. Detecting and properly assigning occlusions is, however, not trivial. Also noteworthy is the recent work on scene flow, which adds the temporal gradient of disparity to the feature set. The disparity gradient is typically quite noisy and powerful regularization mechanisms are required. Wedel, Rabe, Meissner, Franke, and Cremers (2009) proposed such a method that combines scene flow with graph-cut-based segmentation for independently moving object detection.

Methods

Processing proceeds along several stages (Figure 1A), where some stages are standard; others, however, had to be newly developed and represent original contributions to the solution of the motion analysis problem.

There is no one-to-one mapping between model stages and cortical areas. For clarity, Figure 1 and this section are organized according to the processing stages in the system rather than the cortical areas to which they correspond. We next discuss each stage in turn and indicate the cortical areas that are most likely involved in the processing. We do not provide details on the algorithms here but rather refer to the Appendices. To each subsection corresponds a section in Appendix A that provides a more detailed description of the algorithms involved. The different stages and the model as a whole have been realized using a hybrid of conventional hardware and massively parallel graphics hardware. This is discussed in Appendix B.

Gabor pyramid

A stream of stereo images, here from a movie taken during driving a car, is filtered by a pyramid of quadrature-pair (complex) Gabor wavelets of different orientation and spatial frequency (six frequencies, eight orientations). Panel 1 in Figure 1B shows the real-valued response of a horizontal filter (f = 1/16 cyc/pixel) for the left image of the image pair in Figure 1A.

Gabor filters have been used successfully to model the binocular receptive field properties of visual cortical cells (Daugman, 1985; Ohzawa, DeAngelis, & Freeman, 1990; Pollen & Ronner, 1981) and serve as the front end in energy models for stereo and motion computation (Adelson & Bergen, 1985; Ohzawa et al., 1990). This first stage of the model performs part of the processing that occurs in the retina, LGN, and cortical area V1 simple cells.

Binocular disparity

Binocular disparity (Panel 2) is calculated using phase-based methods (Sanger, 1988). Phase differences between the left and right images are pooled across orientation and propagated from coarser to finer scales to increase the dynamic range (Sabatini et al., 2010). In Panel 2, the stereo disparity ranges from −60 to 0 pixels (see scale bar).

Phase-based methods are exactly equivalent to energy methods at the final steps where disparity values are made explicit (Qian & Mikaelian, 2000). Predictions from the disparity energy model agree well with the receptive field properties of binocular simple and complex cells observed in quantitative physiological experiments (Ohzawa, DeAngelis, & Freeman, 1997). The distributed representation of disparity imposes large memory requirements on energy methods. To satisfy the real-time requirement, we instead use the more efficient phase-difference method. The coarse-to-fine control scheme and orientation pooling are compatible with an energy model with position- and phase-shift components (Chen & Qian, 2004).

Disparity is used as a source of depth information in our model and can be replaced or enhanced by other sources (Warren & Rushton, 2009a). We therefore use a simple measure of absolute disparity, such as observed in cortical area V1. We do not require relative disparity (V2 and beyond), disparity gradients (MT/V5 and MST), or higher order disparity (ventral stream; Orban, 2008). If we also consider the cue integration discussed further on, then our disparity signal is likely situated in areas MT/V5 and MSTl.

Optical flow

Optical flow (Panel 3) is also calculated on the basis of the phase of the Gabor filter responses. The temporal phase gradient now serves the role of the binocular phase difference in the disparity stage. We use a highly efficient algorithm that derives component velocity from the temporal evolution of spatial phase in time and integrates component velocities into pattern velocity by means of an

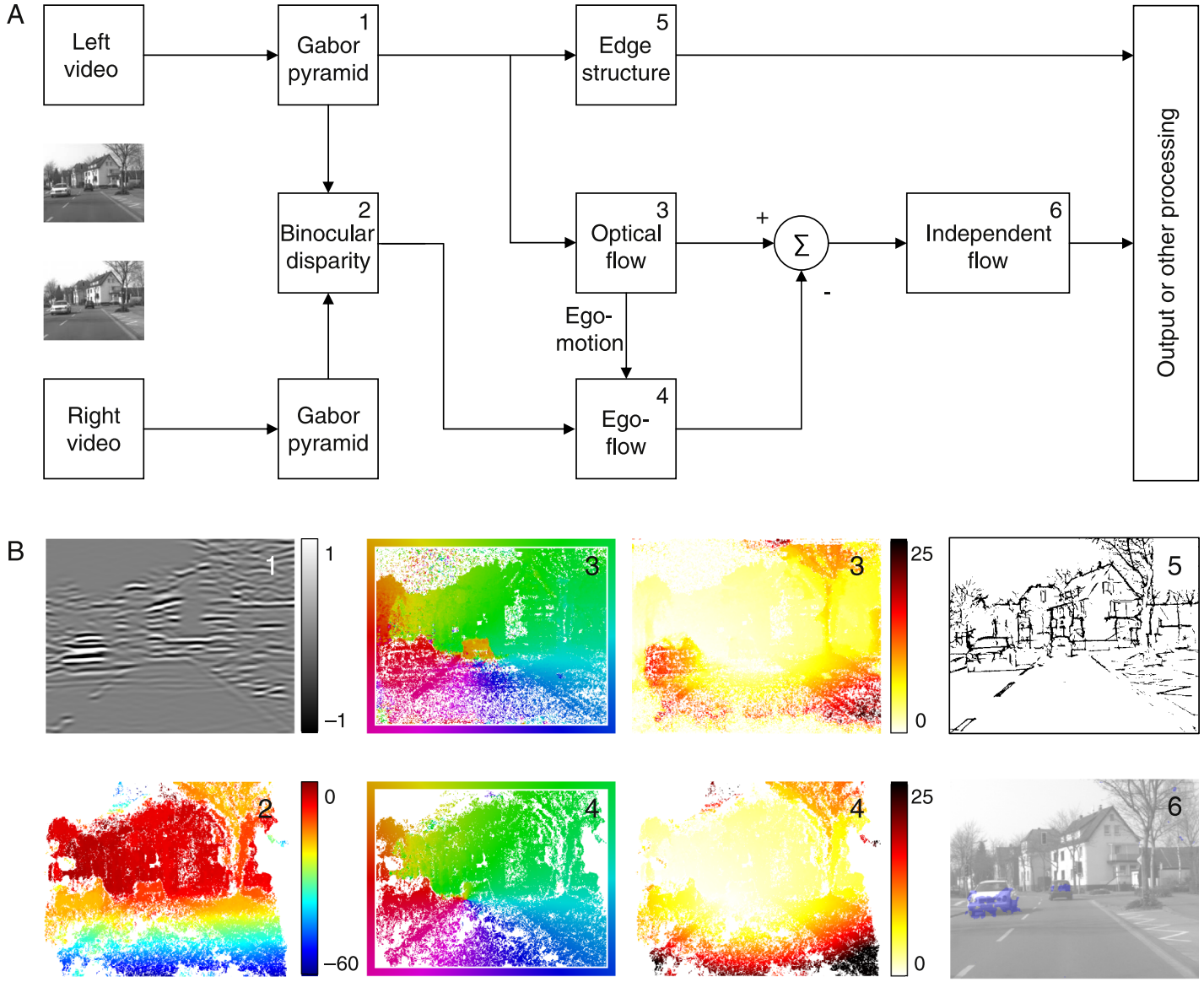

Figure 1. (A) Algorithmic procedure and flow diagram for the processing of motion. (B) Results of the different processing stages. Numbers refer to the boxes in (A).

intersection-of-constraints procedure (Gautama & Van Hulle, 2002). As in the disparity stage, a coarse-to-fine control scheme is used to increase the dynamic range (Pauwels & Van Hulle, 2009). The orientation and amplitude of the optical flow vectors are encoded separately in Figure 1B. The legend frame around Panels 3 (left) and 4 (left) depicts all vectors from the image center to the respective frame pixel location, for example red represents that particular flow vector that points from the image center to the location where the frame has this color (for red, this is a leftward pointing flow vector). Panels 3 (right) and 4 (right) provide the amplitude information, i.e., the vector length of the flow vectors. Panel 3, thus, represents a mixture of egomotion induced radial flow and other flow components, which originate from independently moving objects.

The motion energy model (Adelson & Bergen, 1985) is the basis of many computational models of areas V1 and MT/V5 and relies on amplitude information extracted from quadrature-pair spatiotemporal band-pass filters. Since amplitude information is highly sensitive to contrast changes, this class of models requires explicit normalization mechanisms. The details thereof and the nonlinearities involved are the basis of many modeling studies (Perrone & Krauzlis, 2008a; Rust, Mante, Simoncelli, & Movshon, 2006; Simoncelli & Heeger, 1998). Unlike amplitude, phase is stable under small deviations from image translation that typically occur with projections of 3D scenes, such as dilations, rotations, and changes in contrast and scale (Fleet & Jepson, 1993). Contours of constant phase can be tracked using the phase gradient and used to derive component velocity (Fleet & Jepson,

1990). Unlike Fleet and Jepson (1990), we rely on spatial instead of spatiotemporal filtering (Gautama & Van Hulle, 2002). This reduces the number of required filters since only the spatial frequency domain needs to be tiled (see also Figure A1B).

The optical flow stage presented here corresponds to areas V1 and MT/V5 since both component and pattern motion are extracted. As in the disparity stage, we do not account for surround interactions to extract, e.g., velocity gradients, nor do we include interactions with disparity estimation at this stage.

Edge structure

The edge structure (Panel 5) is extracted on the basis of a robust phase congruency method that identifies features at points where the Fourier components are maximally in phase (Kovesi, 1999).

Phase congruency is a central component of the Local Energy Model (Morrone & Burr, 1988), which was introduced to explain human feature detection. Selectivity to phase congruency has been observed in V1 and multiple higher visual cortical areas in both the dorsal and ventral streams (Henriksson, Hyvärinen, & Vanni, 2009; Mechler, Reich, & Victor, 2002; Perna, Tosetti, Montanaro, & Morrone, 2008).

Egomotion and ego-flow

Binocular disparity and optical flow are used to calculate the ego-flow (Panel 4); those flow components that result only from egomotion and are removed by the flow-parsing mechanism. To this end, translation direction (heading) and rotational velocity are derived from the (monocular) optical flow using a Gauss–Newton optimization procedure (Zhang & Tomasi, 2002). The optical flow vectors associated with independently moving objects whose heading differs from that of the observer are often strong outliers in the egomotion estimation. We therefore use an iterative procedure to gradually exclude them from the estimation process. The novelty of this procedure stems from the fact that we perform analysis for the first time on dense flow fields, which is only possible due to parallel hardware processing. The egomotion can now be combined with the depth information from binocular disparity to obtain the ego-flow. Due to driving forward mostly radial flow is observed in Panel 4.

Many computational models have been proposed to explain cortical egomotion estimation, such as the population heading-map model (Lappe & Rauschecker, 1993), the template model (Perrone, 1992), and the motion-opponent model (Royden, 1997). Besides the different core mechanisms, these models differ in terms of the constraints they impose on eye movements (e.g., fixation), and the degree to which they incorporate extra-retinal signals. Nonetheless, they all operate on the same differential formulation of the motion problem we use here (Longuet-Higgins & Prazdny, 1980). This is different from the discrete formulation used more commonly in computer vision methods for matching sparse feature sets (Hartley & Zisserman, 2004). Two important differences between our model and the computational models mentioned above are the outlier removal stage, which avoids the bias due to the presence of independent motion, and the use of a normalized error function (Zhang & Tomasi, 2002), which avoids the well-known heading bias toward the center of the scene (Royden, 1997).

In accordance with a large amount of experimental evidence, the egomotion stage of our model most likely corresponds to area MSTd. It is less clear where ego-flow is situated, but its computation possibly involves the reciprocal connections between MT and MST (Maunsell & Van Essen, 1983) and the disparity information available in both these regions.

Independent flow segments

In accordance with the flow-parsing mechanism, the ego-flow is subtracted from the optical flow. Next, a 3D translation is estimated on the basis of the residual flow and disparity estimates from a small region around the pixel (see Appendix A for details). Regions where the measurements are consistent with a 3D translation are shown in Panel 6 and are indicative of moving objects in the scene. Recent studies also point toward world-centered criteria for judging object motion during visually guided egomotion from optical flow (Matsumiya & Ando, 2009). Note that this procedure is different from that of Thompson and Pong (1990), which evaluates the inconsistency (rather than the consistency) of the optical (rather than the residual) flow and disparity with the depth/flow constraint.

Computational models that employ the flow-parsing mechanism have yet to be proposed. Warren and Rushton (2009b) identified MST (dorsal and ventrolateral) as a potential candidate for flow parsing and suggested a Perrone (1992)-like template model to perform the subtraction (Warren & Rushton, 2009a). Perrone and Krauzlis (2008b) also presented a vector subtraction mechanism (using extra-retinal information) and pointed out that an actual subtraction should occur downstream from MT.

Area MSTl is a good candidate for the independent flow segment stage of the model. Its cells show a clear preference for small, moving stimuli (Tanaka et al., 1993) and encode object motion in world-centered coordinates (Ilg et al., 2004). In addition, they have large receptive fields (useful for spatial consistency checking)

	Time (in ms)	fps	
320 × 256	33	30	GPU
	2929	0.34	CPU
640 × 512	47	21	GPU
	10,007	0.10	CPU

Table 1. Processing times (ms) and frames per second (fps) comparing GPU and CPU performances for different image sizes.

that receive input from MT and are selective to disparity (Eifuku & Wurtz, 1999). They thus have access to all the information required by this final stage of the model.

Results

Table 1 shows the processing times achieved with our parallel implementation and demonstrates that the complete set of dense feature maps, including the final motion analysis stage, operates in video real time up to an image size of 640 × 512 pixels. These processing times are discussed in more detail in Appendix B.

High speed is combined with high precision of the motion estimates. Uncalibrated outdoor settings, however, cannot be used to demonstrate this. Hence, we used indoor scenes with precisely known parameters for egomotion as well as independent motion. The benchmark scenes used in this section are available as auxiliary material. A scenario similar to Figure 1 has been chosen roughly to scale, where two "cars" move in the same direction but with different speed on a "road," while the camera moves exactly in the same way as the slower "car." Note that this co-motion makes movement measurement a very difficult problem.

A sample stereo image pair is shown in Figures 2A and 2B, with arrows in Figure 2A depicting object (red) and camera (green) motion. Only the cars on the "road" are moving. The bottom car and the camera move at 2.54 mm/frame and the top car moves at 5.08 mm/frame. A stop-motion technique was used to obtain highly precise object and camera motion. The stereo camera pair (a Point Grey BumbleBee2) was attached to an articulated industrial robot arm (Staubli RX-60) with six degrees of freedom to enable repositioning with submillimeter precision after each frame. Highly accurate object motion was obtained by moving the cars on a prototyping printed circuit board.

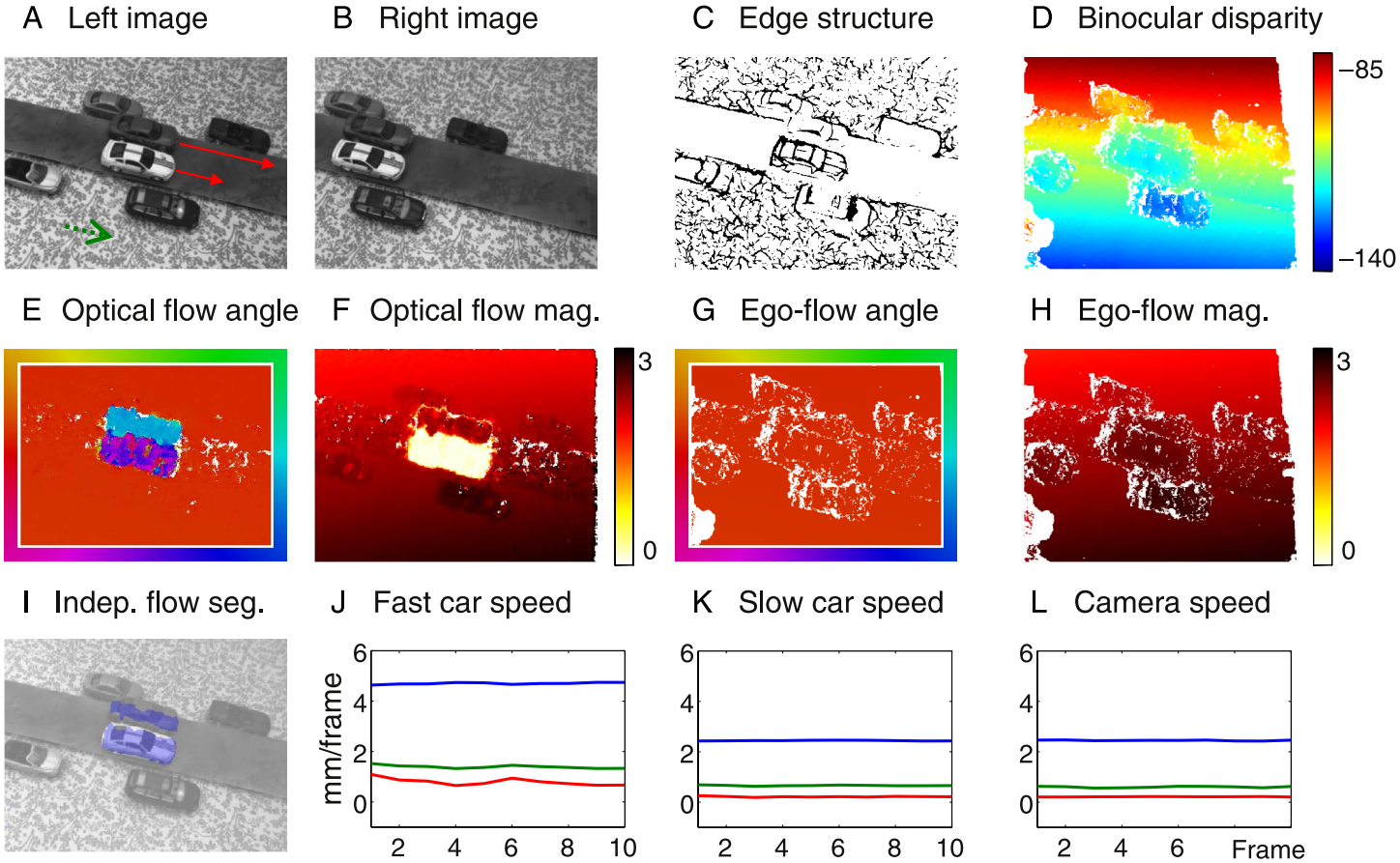

Figure 2. Low-level visual cues and independent motion detection and description results obtained on the benchmark sequence. (A, B) A magnified section of the fifth image pair, with corresponding cues and detection results in (C)–(I). The red and green arrows in (A) depict, respectively, the cars and camera motion. (J–L) The horizontal (blue), vertical (green), and in-depth (red) translation speeds.

Speed	Camera	Slow car	Fast car
x	2.45	2.46	4.73
y	0.59	0.66	1.37
z	0.22	0.20	0.73
Mag.	2.53	2.55	4.98
Real	2.54	2.54	5.08

Table 2. Measuring motion estimation accuracy using indoor scenes. Values give the different velocity components for the *x*-, *y*-, and *z*-axes as well as the magnitude (mag.) in mm/frame. The real velocity magnitude is given in the last row.

Figure 2 contains, for the fifth frame of the sequence, the image pair (A, B), the low-level visual cue edge structure (C), binocular disparity (D), and optical flow (E, F) together with the ego-flow (G, H) and the independently moving segments (I). Note that the results shown here correspond to an enlarged part of the original images used.

The planar scene structure is clearly visible in the disparity and optical flow magnitude. The optical flow magnitude of the slowest car is equal to zero as a result of the camera's tracking motion. The two largest detected independently moving segments correspond to the moving cars. For this particular frame, the estimated camera and moving objects' translation speeds are shown in Table 2. The largest magnitude error is obtained for the fast car and is equal to 0.1 mm/frame. The estimated camera rotation was essentially zero (the largest magnitude found was 1.4×10^{-4} radians/frame).

The cars were tracked across the sequence by selecting those segments that maximally overlap from one frame to the next. The camera and moving cars' translation speeds estimated on frames one to ten are shown in Figures 2J, 2K, and 2L. These estimates remain almost constant across the sequence. As expected, the camera and slow car have almost identical translation speeds, and the fast car's speed is approximately double that of the camera. The largest variability can be observed in the translation component along the line of sight (in-depth). This can be expected since it is difficult to estimate the convergence or divergence component of the optical flow field with high precision on the basis of measurements obtained from a small spatial region.

In summary, we find that both moving objects are detected with a density of more than 50% as compared to all pixels that belong to the objects and magnitude (mag.) of the speed is calculated with an accuracy of better than 98% (Table 2). This combination of accuracy and speed is unprecedented and currently not reached by any other artificial vision system.

To demonstrate that the egomotion algorithm is accurate in the presence of both camera translation and camera rotation, we increased the complexity of the benchmark sequence by introducing a 3D camera rotation. Since camera rotations do not depend on the scene structure, they can be simulated by warping the image sequence. Using the discrete formulation (Hartley & Zisserman, 2004), we applied an arbitrary but constant 3D rotation, equal to $\boldsymbol{\omega} = (-0.002, -0.004, -0.006)^{\mathrm{T}}$ radians/frame, to

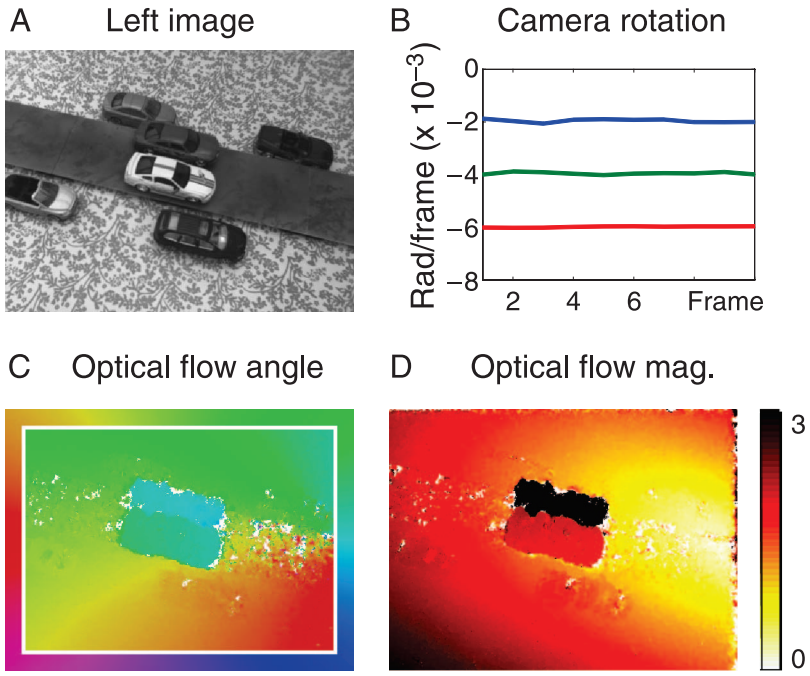

Figure 3. Optical flow and camera motion obtained on the benchmark sequence after the addition of a simulated 3D rotation. (A) A magnified section of the fifth image (identical to Figure 2A). (B) The estimated rotational velocity along the horizontal (blue), vertical (green), and line-of-sight (red) axes. (C, D) The optical flow angle and magnitude.

Figure 4. Detection results in real-world driving situations involving cars, a bike, and pedestrians.

the sequence (see Appendix A for a description of the egomotion parameters) while keeping the fifth frame unaltered. Since this rotation is added on top of the already present camera translation, the resulting sequence is highly complex (see the optical flow in Figures 3C and 3D). Nevertheless, the camera rotation is extracted with high precision, as shown in Figure 3B. It is not possible to extract disparity and independent motion in this way, since the right camera needs to be translated (and this cannot be simulated) to maintain the rectification required by the disparity algorithm.

Some additional detection results obtained in real-world driving situations are shown in Figure 4 (A and B courtesy of Vaudrey, Rabe, Klette, & Milburn, 2008 and available at http://www.mi.auckland.ac.nz/EISATS).

Figure 4A contains an approaching and receding car and a bike approaching from the right. Figure 4B illustrates a dangerous situation where a driver initiates a turn while a car approaches. Figures 4C and 4D contains pedestrians crossing a road (the van on the right is stationary). The camera is moving in all four examples.

Discussion

A number of different approaches toward the motion segmentation problem have been proposed by the robotics and machine vision communities. These approaches can be categorized in terms of the number of features used and the number of frames these features are retained. SLAM approaches usually extract around ten features each frame and store them indefinitely, whereas visual odometry approaches track hundreds of features over a small number of frames. Our system is extreme in this sense since it relies on tens or even hundreds of thousands of features (optical flow) that are not retained at all. Although the lack of a temporal global optimization step renders the system less suitable for navigation problems, the enormous feature redundancy increases the robustness of the egomotion component to the presence of independent motion. The short time span also enables rapid detection of highly variable independent motion. The motion processing system presented here is therefore complementary to SLAM and odometry approaches. It can help avoid the erroneous inclusion of moving objects in the feature set, or even direct the methods toward these regions so that the objects can be tracked and their trajectories identified (Pugeault, Pauwels, Pilz, Van Hulle, & Krüger, in press). In a similar fashion, interactions with appearance-based (cf. ventral stream) object detection methods have already proven useful in real-world situations (Ess et al., 2009; Leibe et al., 2008). Some psychophysical studies highlight the importance of target location (cf. landmarks) as a cue for the guidance of locomotion (Rushton, Harris, Lloyd, & Wann, 1998), which also points toward the involvement of different mechanisms. These and other interactions provide many interesting further topics of study.

Our model differentiates itself from motion segmentation methods from the computer vision literature, by borrowing more elements from neuroscience and psychophysics studies. In particular, the energy model for low-level features, the differential motion constraint for egomotion, and the MSTl-like templates for world-centered object motion are not commonly used in technical systems. Only minimal abstractions, such as removing population codes and biases, have been introduced to achieve high accuracy in real time in real-world complex scenes. This high performance and the central role of the flow-parsing mechanism (Rushton et al., 2007; Rushton & Warren, 2005; Warren & Rushton, 2007, 2008, 2009a, 2009b) is also the main difference between our model and the alternative computational models of motion processing (Browning et al., 2009; Zemel & Sejnowski, 1998).

It is important to note that the here adopted modular pixel-parallel approach allows extending this system (possibly by the use of additional GPU cores) to render

other, additional image processing combinations. For example, one could now combine independent flow segments with edge information by an AND-like operation to overlay object outlines (Pugeault et al., in press), or use the edge structure to guide spatial regularization mechanisms (cf. Markov random field methods) that can greatly improve the quality of the segmentation (Wedel et al., 2009). Other possible extensions are the inclusion of spatial surround interactions (such as observed in MT/V5) and motion-in-depth mechanisms (Rokers, Cormack, & Huk, 2009), which both require regularization mechanisms for noise reduction (Wedel et al., 2009).

During the last decades, an improved understanding of the processing principles in the ventral and dorsal visual cortical pathways has led to more and more attempts for transferring these principles to model systems (Fukushima, 1980; Grossberg, 2003; Lappe, Bremmer, Pekel, Thiele, & Hoffmann, 1996; LeCun et al., 1989; Mishkin, Ungerleider, & Macko, 1983; Riesenhuber & Poggio, 1999; Serre, Wolf, Bileschi, Riesenhuber, & Poggio, 2007; Wallis & Rolls, 1997). One of the first serious attempts to combine cortical processing with parallel hardware in a large-scale system for proving the power of neuronal inter-area processing for image analysis was performed by Poggio and co-workers (Little, Poggio, & Gamble, 1988; Poggio, Gamble, & Little, 1988).

All these approaches are either highly tailored toward a certain problem or they remain focused on individual components and are thereby limited. Furthermore, neuronal approaches usually focus on processing in the ventral object recognition stream, the dorsal motion processing stream has received far less attention, and efficient parallel motion processing models do not yet exist. This is due to the fact that up to now real-time image processing has remained a very demanding technical task with substantial scientific challenges. The here shown realization of a pixel-parallel, neuronal implementation is the first artificial vision system of this size and complexity, which operates in real time and achieves dense and very reliable estimates for a total of six feature representations concluding in the analysis of independent motion. It is important to note that the highly parallel and dense processing architecture presented here is not just a qualitative extension of older systems. To solve the independent motion analysis problem in a model-free way (hence, without further assumptions on the image structure), high-speed, high-density, and massive inter-feature interactions are needed at the same time. Systems that lack one of these aspects remain fundamentally impaired and the solution to this problem cannot be obtained from a gradual extension of existing approaches. Many problems in the analysis of low-level visual information (so-called "early vision") are similarly structured and require, thus, holistic approaches like the one taken here. Given the increasing power of parallel hardware and its new and efficient programming tools, it is highly likely that systems of this kind will therefore in the near future dominate early vision. This would represent a major breakthrough as complex cognitive vision problems, which vitally rely on reliable early vision, would become more easily addressable this way.

Appendix A

System modules

This first appendix describes the different algorithmic procedures employed in our system and their specific modification for parallel use. This study rests on the successful transfer of neuronal mechanisms to abstract models, which can be implemented on parallel hardware for real-time operation.

The main scientific contribution is the synthesis of a highly complex system with many interleaved components. This could not be achieved without detailed analysis of the internal interaction processes and the seeking for novel solutions so that all components can communicate highly efficiently in a large network. To this end, several subcomponents had to be specifically invented and/or modified allowing for real-time processing within high-density maps. In general, this leads to increasing levels of abstraction of the original neuronal models.

The system consists of a total of six subcomponents, depicted in Figure 1, which will be described in the following.

Gabor pyramid

Following models of cortical simple cells (Daugman, 1985; Ringach, 2002), all the algorithms used in the different modules critically rely on Gabor phase to detect features and to establish correspondences. For a specific orientation, θ, the spatial phase at pixel location $\mathbf{x} = (x, y)^{\mathrm{T}}$ is extracted using 2D complex Gabor filters:

$$f(\mathbf{x}) = e^{-\frac{x^2+y^2}{2\sigma^2}} e^{j\omega_0(x\cos\theta + y\sin\theta)}, \quad \text{(A1)}$$

with peak angular frequency ω_0 and spatial extension σ. We use a total of eight evenly distributed orientations in our implementation. The peak frequency is doubled from one scale to the next. To accommodate this, the filters span an octave bandwidth: $B = \omega_0/3$. With a cutoff frequency equal to half the amplitude spectrum, the spatial extension is then equal to $\sigma = \sqrt{(2\log(2)/B}$.

At the highest frequency, we have $\omega_0 = \pi/2$ rad/pixel, which results in a spatial extension $\sigma = 2.25$ pixels. The lower frequency responses are obtained by applying the same filters to an image pyramid that is constructed by

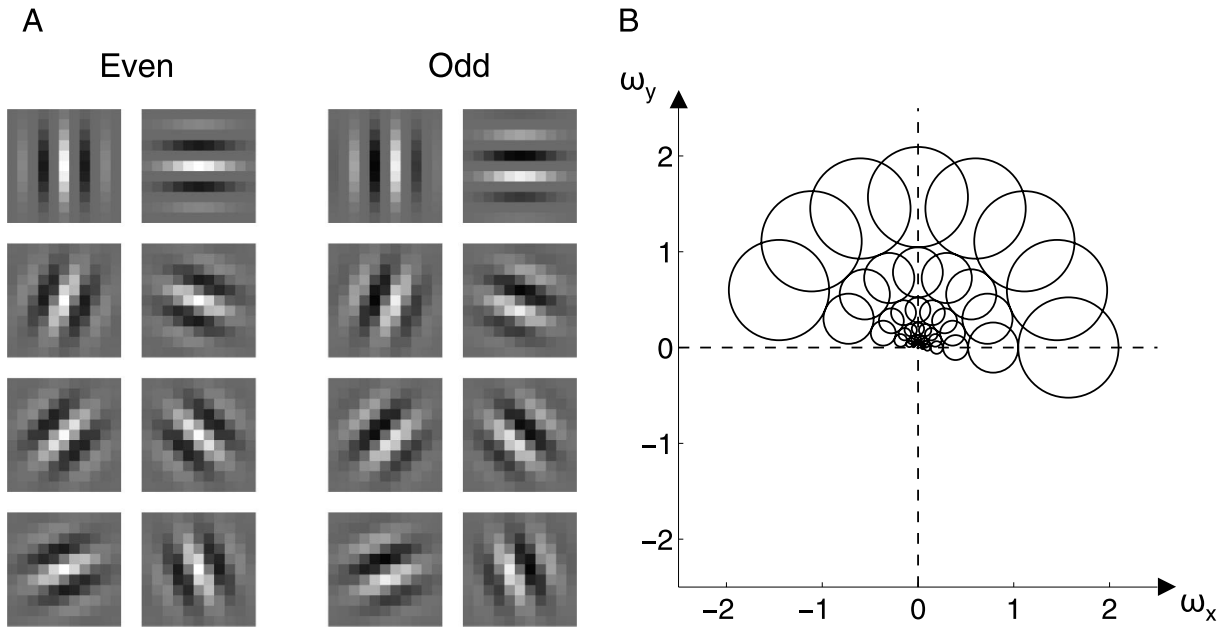

Figure A1. (A) Even and odd filters used to extract spatial phase at eight orientations. The responses can be obtained on the basis of 24 1D convolutions with 11-tap filters. (B) Frequency domain coverage of the filter-bank when applied to the image pyramid. The circles correspond to the cutoff frequency at half the amplitude spectrum.

repeatedly blurring the images with a Gaussian kernel and subsampling (Burt & Adelson, 1983). We use a total of six scales. This is a technically very feasible way to approximate the multitude of neuronal responses from neurons with different receptive field sizes. The spatial filter kernels are 11 × 11 pixels in size and separable. Since some of the responses can be reused, all eight even and odd filter responses can be obtained on the basis of only 24 1D convolutions. The spatial filter kernels are shown in Figure A1A and their frequency domain coverage when applied to the image pyramid is shown in Figure A1B.

The filter responses, obtained by convolving the image, $I(\mathbf{x})$, with the oriented filter from Equation A1 can be written as

$$Q(\mathbf{x}) = (I * f)(\mathbf{x}) = \rho(\mathbf{x})e^{j\phi(\mathbf{x})} = C(\mathbf{x}) + jS(\mathbf{x}). \quad \text{(A2)}$$

Here $\rho(\mathbf{x}) = \sqrt{C(\mathbf{x})^2 + S(\mathbf{x})^2}$ and $\phi(\mathbf{x}) = \text{atan2}(S(\mathbf{x}), C(\mathbf{x}))$ are the amplitude and phase components, and $C(\mathbf{x})$ and $S(\mathbf{x})$ are the responses of the quadrature filter pair (Pollen & Ronner, 1981). The * operator depicts convolution.

Binocular disparity

Binocular disparity is computed on the basis of the phase difference between the left and right images. Assuming rectified images, a stereo disparity estimate can then be obtained from each oriented filter response (at orientation θ) by projecting the phase difference on the epipolar line (the horizontal). In this way, multiple disparity estimates are obtained at each location. We robustly combine these estimates using the median:

$$\delta(\mathbf{x}) = \underset{\theta}{\text{median}} \left(\frac{\left[\phi_\theta^L(\mathbf{x}) - \phi_\theta^R(\mathbf{x})\right]_{2\pi}}{\omega_0 \cos\theta} \right), \quad \text{(A3)}$$

where the $[]_{2\pi}$ operator depicts reduction to the $]-\pi;\ \pi]$ interval. As discussed in more detail in the next section, a coarse-to-fine control scheme is used to integrate the estimates over the different pyramid levels. Unreliable estimates are removed by running the algorithm from left to right and from right to left and looking for mutual consistency.

Optical flow

By exploiting the conservation property of local phase measurements (phase constancy; Fleet & Jepson, 1990), optical flow can be computed from the temporal evolution of equi-phase contours $\phi(\mathbf{x}, t) = c$. Differentiation with respect to t yields

$$\nabla\phi \cdot \mathbf{u} + \psi = 0, \quad \text{(A4)}$$

where $\nabla\phi$ is the spatial and ψ is the temporal phase gradient, and $\mathbf{u}$ is the optical flow. Under a linear phase model, the spatial phase gradient can be substituted by the radial frequency vector, $(\omega_0\cos\theta, \omega_0\sin\theta)^T$ (Fleet, Jepson,

& Jenkin, 1991). In this way, the component velocity, $\mathbf{c}_\theta(\mathbf{x})$, at pixel $\mathbf{x}$ and for filter orientation θ, can be estimated directly from the temporal phase gradient, $\psi_\theta(\mathbf{x})$:

$$\mathbf{c}_\theta(\mathbf{x}) = -\frac{\psi_\theta(\mathbf{x})}{\omega_0}(\cos\theta, \sin\theta)^{\mathrm{T}}. \quad \text{(A5)}$$

At each location, the temporal phase gradient is obtained by fitting a linear model to the (unwrapped) spatial phase across five frames (Gautama & Van Hulle, 2002):

$$\phi_\theta(\mathbf{x}, t) = a + \psi_\theta(\mathbf{x})t. \quad \text{(A6)}$$

The reliability of each component velocity is measured by the mean squared error (MSE) of the linear fit. Provided a minimal number of component velocities at pixel $\mathbf{x}$ are reliable, they are integrated into a full velocity by means of an intersection-of-constraints procedure:

$$\mathbf{v}^*(\mathbf{x}) = \underset{\mathbf{v}(\mathbf{x})}{\operatorname{argmin}} \sum_{\theta \in O(\mathbf{x})} \left(|\mathbf{c}_\theta(\mathbf{x})| - \mathbf{v}(\mathbf{x})^{\mathrm{T}} \frac{\mathbf{c}_\theta(\mathbf{x})}{|\mathbf{c}_\theta(\mathbf{x})|} \right)^2, \quad \text{(A7)}$$

where $O(\mathbf{x})$ is the set of orientations that correspond to reliable component velocities.

As in the disparity stage, a coarse-to-fine control scheme is used to extend the dynamic range of the algorithm in an efficient way. This is illustrated in Figure A2.

The control strategy starts at the top of the pyramid, level k. Using the optical flow estimate obtained at that resolution, $\mathbf{v}^k$, the phase estimate at the next higher resolution, ϕ^{k-1}, is warped in such a way that the estimated motion is removed (Bergen, Anandan, Hanna, & Hingorani, 1992):

$$p^{k-1}(\mathbf{x}, t) = \phi^{k-1}(\mathbf{x} - 2 \cdot \mathbf{v}^k(\mathbf{x}) \cdot (3 - t), t). \quad \text{(A8)}$$

Since the optical flow estimate has been computed at the lower resolution, it needs to be doubled first. The factor $(3 - t)$ ensures that each pixel in the five frame sequence (t = 1, 2, …, 5) is warped to its corresponding location in the center frame (t = 3). Bilinear interpolation is used to perform subpixel warps. The warped phase is then used to compute the residual motion. This process is repeated until the pyramid level corresponding to the original image resolution is reached.

The algorithm also relies on a built-in stabilization mechanism to deal with undesired camera jitter (Pauwels & Van Hulle, 2009).

Edge structure

The Local Energy Model (Morrone & Burr, 1988) postulates that features are perceived at points in an image where the Fourier components are maximally in phase. This occurs for step edges, line and roof edges, and Mach bands. Using the filter-bank responses, a measure of phase congruency can be obtained for each orientation by scaling the local energy by the sum of the amplitudes obtained at each scale k:

$$PC_1(\mathbf{x}) = \frac{E(\mathbf{x})}{\sum_k \rho_k(\mathbf{x})}, \quad \text{(A9)}$$

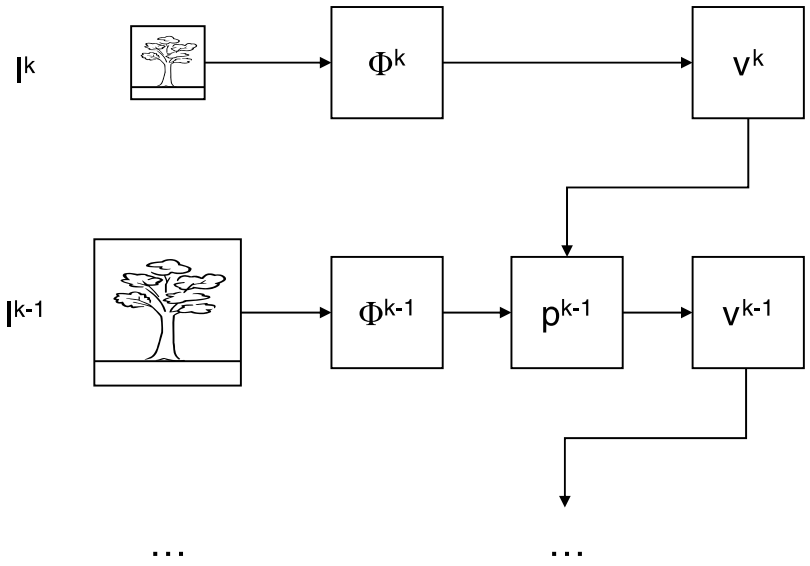

Figure A2. Coarse-to-fine control strategy used to increase the dynamic range of the optical flow algorithm. The optical flow estimates obtained at a lower resolution are used to pre-warp the phase at the next higher resolution so that the estimated motion is removed.

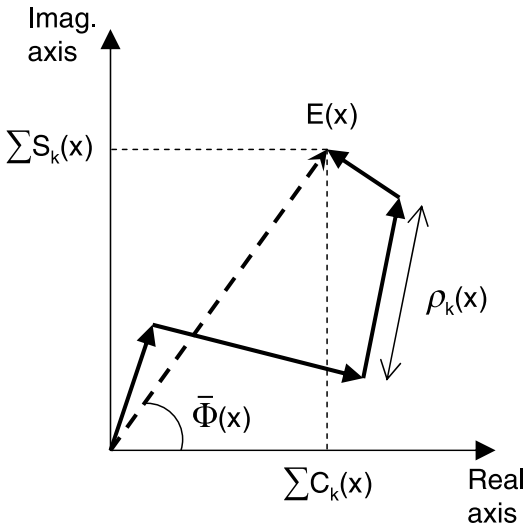

Figure A3. Polar diagram showing the relationship between the local energy and the response amplitudes summed across scale. Both will be equal only if the phase remains constant across scale.

where the local energy $E(\mathbf{x}) = \sqrt{(\sum C_k(\mathbf{x}))^2 + (\sum S_k(\mathbf{x}))^2}$. These different components are illustrated in Figure A3. It is clear from this figure that $PC_1(\mathbf{x})$ will reach a maximum value of unity if all components have the same phase.

The measure from Equation A9 has a number of shortcomings that make it unsuitable for real-world applications. We use the measure proposed by Kovesi (1999), which includes a number of extensions:

$$PC_2(\mathbf{x}) = \frac{\sum_k W(\mathbf{x}) \lfloor \rho_k(\mathbf{x}) \Delta\Phi(\mathbf{x}) - T \rfloor}{\sum_k \rho_k(\mathbf{x}) + \varepsilon}, \quad \text{(A10)}$$

where

$$\Delta\Phi(\mathbf{x}) = \cos(\phi_k(\mathbf{x}) - \bar{\phi}(\mathbf{x})) - |\sin(\phi_k(\mathbf{x}) - \bar{\phi}(\mathbf{x}))|. \quad \text{(A11)}$$

Here $W(\mathbf{x})$ is a factor that weights for frequency spread (derived from the distribution of response amplitudes), $\bar{\phi}(\mathbf{x})$ is the mean phase angle (see Figure A3), T is a threshold to remove noisy components with low energy values, and ε is a small factor to avoid division by zero. The operator $\lfloor\ \rfloor$ converts negative values to zero and leaves positive values unchanged. Subtracting the magnitude of the sine of the phase deviation from the cosine improves localization.

A classical moment analysis is then performed to integrate the different phase congruency estimates across orientation:

$$a = \sum (PC(\theta)\cos(\theta))^2, \quad \text{(A12)}$$

$$b = 2\sum (PC(\theta)\cos(\theta))(PC(\theta)\sin(\theta)), \quad \text{(A13)}$$

$$c = (PC(\theta)\sin(\theta))^2, \quad \text{(A14)}$$

where $PC(\theta)$ is the phase congruency value determined at orientation θ. The maximum moment, M, is then used as a measure for the presence of an edge:

$$M = \frac{1}{2}\left(c + a + \sqrt{b^2 + (a-c)^2}\right). \quad \text{(A15)}$$

Although these extensions are computationally expensive, they are entirely local and well suited for a GPU implementation.

Egomotion and ego-flow

In a dense setting, egomotion estimation is problematic since the correspondences (optical flow) are less accurate than in a sparse framework (interest points). However, a very large number of such correspondences are available and, since the camera position only changes slightly from one image to the next, an instantaneous-time model can be used, which simplifies the estimation.

Assuming, without loss of generality, a focal length equal to unity, the optical flow due to egomotion in a static environment can be described as follows (Longuet-Higgins & Prazdny, 1980):

$$\mathbf{u}(\mathbf{x}) = d(\mathbf{x})A(\mathbf{x})\mathbf{t} + B(\mathbf{x})\boldsymbol{\omega}, \quad \text{(A16)}$$

where $\mathbf{t} = (t_x, t_y, t_z)^{\mathrm{T}}$ is the translational velocity, $\boldsymbol{\omega} = (\omega_x, \omega_y, \omega_z)^{\mathrm{T}}$ is the rotational velocity of the moving observer, $d(\mathbf{x})$ is the inverse depth, and

$$A(\mathbf{x}) = \begin{bmatrix} -1 & 0 & x \\ 0 & -1 & y \end{bmatrix}, \quad \text{(A17)}$$

$$B(\mathbf{x}) = \begin{bmatrix} xy & -1-x^2 & y \\ 1+y^2 & -xy & -x \end{bmatrix}. \quad \text{(A18)}$$

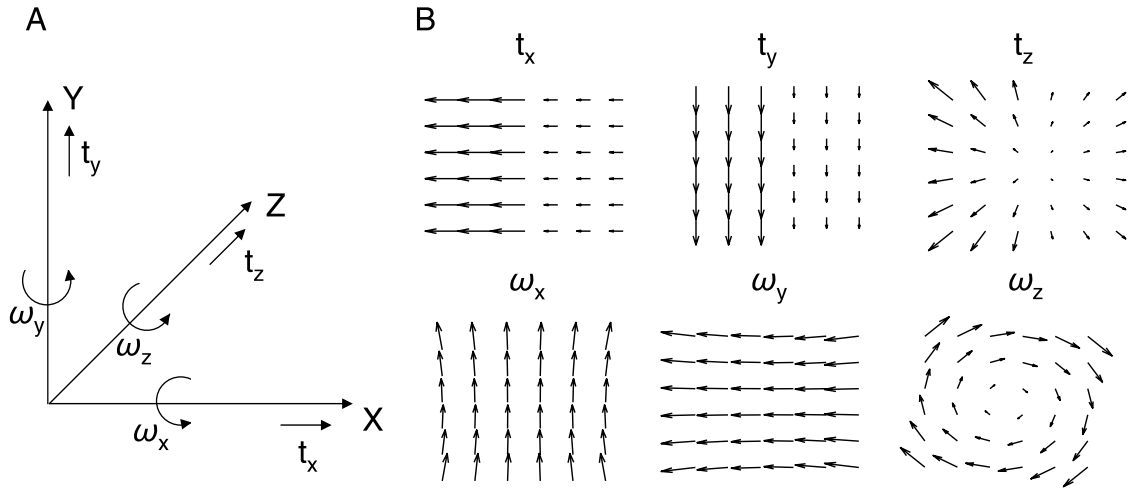

Figure A4. (A) Egomotion and (B) optical flow in a static environment.

The relationship between egomotion and optical flow in a static environment is illustrated in Figure A4. The right panel shows the optical flow fields obtained when each one of the egomotion parameters in turn differs from zero while all the others are equal to zero. A very simple scene is used that consists of two frontoparallel planes at different depths, with the left part closer to the observer than the right part.

To derive the egomotion parameters from the optical flow in general situations where both translations and rotations are present, we use the algorithm by Zhang and Tomasi (2002). This algorithm uses a Gauss–Newton procedure to estimate the egomotion parameters as follows:

$$(\mathbf{t},\boldsymbol{\omega}) = \underset{\mathbf{t},\boldsymbol{\omega}}{\operatorname{argmin}} \sum_{\mathbf{x}} \left[\tau(\mathbf{x},\mathbf{t})^{\mathrm{T}}(\mathbf{u}(\mathbf{x}) - B(\mathbf{x})\boldsymbol{\omega})\right]^2, \qquad \text{(A19)}$$

where

$$\tau(\mathbf{x},\mathbf{t}) = \frac{1}{\|A(\mathbf{x})\mathbf{t}\|}([A(\mathbf{x})\mathbf{t}]_y, -[A(\mathbf{x})\mathbf{t}]_x)^{\mathrm{T}}. \qquad \text{(A20)}$$

These different components are illustrated in Figure A5. The constraints represent the normalized, orthogonal deviations from the epipolar lines and the estimates obtained from Equation A19 minimize the least-squares image-reprojection error (Oliensis, 2005). Since algorithms that operate on this error function obtain the most accurate parameter estimates, they are commonly referred to as "optimal" (unbiased and minimal variance of the estimates; Chiuso, Brockett, & Soatto, 2000; Oliensis, 2005; Zhang & Tomasi, 2002). The algorithm by Zhang and Tomasi (2002) is consistent, which means that arbitrarily accurate estimates of the motion parameters are possible with more and more samples. Highly accurate estimates are required here to obtain a precise mapping between optical flow and binocular disparity in the independent motion detection stage. Independently moving objects are effective outliers in the egomotion estimation. We therefore use an iteratively reweighted least-squares procedure to gradually remove them from the estimation process.

Due to the nonlinearity of the problem ($\mathbf{t}$ and $\boldsymbol{\omega}$ appear as a product in Equation A19) and the robust estimation procedure, the estimation process is sensitive to local minima. We reduce this sensitivity by evaluating a number of initializations ($n = 32$) in parallel. The median of the residuals is used to evaluate the quality of the estimates and to determine the outlier rejection threshold. We use an approximation to the median that is more suitable for parallel implementation.

After estimating optical flow, heading (translation direction), and rotation, the relative inverse depth from motion, $d_M(\mathbf{x})$, can be computed in the following way (Zhang & Tomasi, 2002):

$$d_M(\mathbf{x}) = \frac{(\mathbf{u}(\mathbf{x}) - B(\mathbf{x})\boldsymbol{\omega})^{\mathrm{T}} A(\mathbf{x})\mathbf{t}}{\|A(\mathbf{x})\mathbf{t}\|^2}. \qquad \text{(A21)}$$

Note that $d_M(\mathbf{x})$ is relative due to the scale ambiguity that occurs in monocular egomotion estimation ($d(\mathbf{x})$ and $\mathbf{t}$

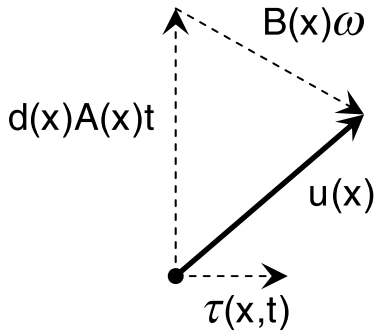

Figure A5. Optical flow and its components due to egomotion.

appear as a product in Equation A16). For parallel cameras with baseline b and unity focal length (again, without loss of generality), binocular disparity, $\delta(\mathbf{x})$, is also related to inverse depth: $\delta(\mathbf{x}) = -b/z$. Consequently, the mapping between the two can be easily obtained in a robust fashion:

$$S = \underset{\mathbf{x}}{\text{median}} \left(\frac{d_M(\mathbf{x})}{\delta(\mathbf{x})} \right). \tag{A22}$$

This accounts for the egomotion speed and the baseline. Using this mapping, the relative inverse depth from disparity, $d_D(\mathbf{x})$, becomes $d_D(\mathbf{x}) = S\delta(\mathbf{x})$. We then use this measure to compute the ego-flow, the flow that would have been observed in an entirely static environment:

$$\mathbf{u}_{\text{ego}}(\mathbf{x}) = d_D(\mathbf{x})A(\mathbf{x})\mathbf{t} + B(\mathbf{x})\boldsymbol{\omega}. \tag{A23}$$

Independent flow segments

For the detection of independent motion, we propose a cue combination procedure that only relies on the quality of fit to a 3D motion model, and not on the particular values estimated. Under the simplifying assumption that the object undergoes a strict 3D translation, $\mathbf{t}^I$, the following model should be able to explain the differences between the optical flow and the ego-flow:

$$\mathbf{u}(\mathbf{x}) - \mathbf{u}_{\text{ego}}(\mathbf{x}) = d_D(\mathbf{x})A(\mathbf{x})\mathbf{t}^I. \tag{A24}$$

For each pixel, we estimate this 3D translation in the least-squares sense using all the reliable measurements in a small spatial region (11×11) around the pixel. This is computationally very intensive but, due to its similarity with convolutions, well suited for a GPU implementation.

By using a 3D translation model, we also allow for looming. The inclusion of disparity in the model enforces mutual consistency on the three visual cues (optical flow, egomotion, and disparity), while still allowing for arbitrary complex object structure. In many situations, this model will be overly complex and a simple 2D translation model (and frontoparallel structure) may suffice. Due to this inherent ambiguity, we do not rely on the estimated parameters, but rather on the quality of the model fit. We use a measure related to the coefficient of determination. This measure ensures that detectability is not affected by the magnitude of the independent motion vectors and enables the detection of both near and far moving objects. The moving segments can then be identified by thresholding this measure and labeling the connected components.

Provided that a sufficiently large moving segment has been found, its motion parameters are determined by estimating the model from Equation A24 on the basis of all the data within the segment.

Appendix B

Creating a complex, interactive, and parallel system

The entire system has been implemented using NVIDIA's CUDA framework (Lindholm, Nickolls, Oberman, & Montrym, 2008), which facilitates the development of hybrid CPU/GPU applications. The main bottleneck in such hybrid applications is the memory bandwidth between CPU and GPU, which is about 50 times smaller than the bandwidth inside the GPU. It is therefore crucial to perform all data-intensive processing on the GPU, while minimizing transfers to and from the CPU. In our system, the CPU manages the complex control logic and performs simple computations that cannot be done efficiently on the GPU. An overview of the proposed system is shown in Figure B1.

The discussion is organized in five sections: the Gabor pyramid construction, the computation of the low-level vision cues (edge structure, optical flow, and disparity), the estimation of egomotion, the detection of independent flow segments, and finally, the processing times.

Gabor pyramid

After the CPU has captured the images, the stereo image pair is transferred onto the GPU. The transformation from a grayscale image into our multi-resolution, multi-orientation Gabor wavelet representation represents more than a 20-fold increase in data (16 filter responses per pyramid pixel). This fact has long prevented the use of such a rich image representation in real-time vision systems. The GPU architecture, however, is ideally suited for this. Since the wavelet expansion can be performed on-chip, it benefits from the high internal bandwidth. The construction of the Gabor wavelet representation relies extensively on 2D separable convolution operations that are highly data-parallel and thus well matched to the GPU architecture.

In this stage, the GPU is also responsible for subsampling the images and for appropriately combining and interleaving the 1D convolution responses. The final Gabor pyramid is stored exclusively in GPU memory.

Edge structure, optical flow, and disparity

The computations required to obtain the edge measure from Equation A15 can be performed for each pixel independently. The GPU's texture units are used to bilinearly interpolate the filter-bank responses from coarser scales to the image resolution. Note that these texture units are an additional computational resource, separate from the main computation units (the streaming processors).

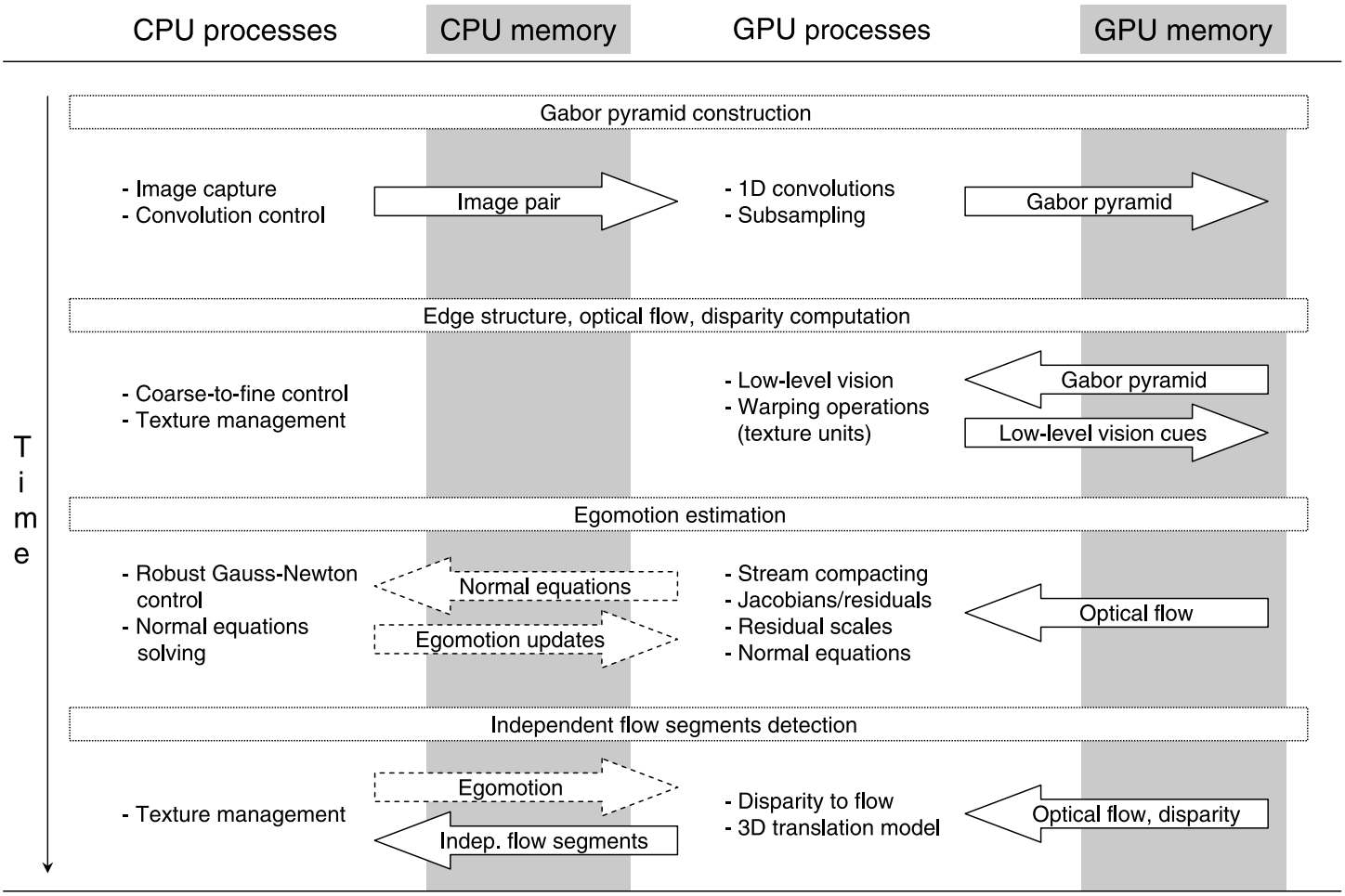

Figure B1. System overview illustrating the sequence of processes running on the CPU and GPU. Solid and dashed arrows depict large and small data transfers, respectively.

The optical flow and disparity processes share the left image Gabor wavelet responses. They also rely heavily on the GPU's texture units. Contrary to the extraction of edge structure, the Gabor pyramid is gradually propagated from coarse to fine scales. The filter responses used at a certain pixel location depend on the optical flow or disparity computed at the previous pyramid level. Although this warping transformation cannot be predicted, spatial locality is high, and the texture cache enables a very high bandwidth in this crucial data-intensive stage of the algorithms. The texture units also perform bilinear interpolation here, which greatly increases the precision of the estimates. After the warping operations, the remaining computations are entirely local.

In this low-level vision stage, the CPU is responsible for controlling the pyramid traversal and for appropriately assigning GPU memory to the texture units.

Egomotion

The robust nonlinear egomotion formulation suffers from local minima. A global optimization strategy based on the evaluation of many initializations is well suited for a GPU implementation because different initializations can be evaluated in parallel.

The GPU first selects a subset of the reliable optical flow estimates in parallel (stream compacting). Next, the Gauss–Newton procedure is started in close cooperation with the CPU. The GPU is responsible for the data-intensive parts. These are the computation of the Jacobians and residuals, robustly estimating the scale of the residuals and composing the normal equations.

The most time-consuming step is the composition of the normal equations. The scale estimates are used here to determine the weights. Data reduction and vector outer product computations are combined in this stage to control the bandwidth usage.

The normal equations are very compact for this problem and only 20 values for each initialization need to be transferred to the CPU. The equations are then solved on the CPU and the egomotion updates are used to start the next iteration on the GPU. After a pre-defined number of iterations (typically 30), the egomotion estimate with the smallest median absolute residual is selected as the final solution.

	GPU		CPU	
Resolution	320 × 256	640 × 512	320 × 256	640 × 512
Gabor filtering	3	5	150	608
Edges/flow/disparity	7	14	1406	6926
Egomotion	19	20	1003	1003
Indep. flow seg.	4	8	370	1470
Total	33	47	2929	10,007
Fps	30	21	0.34	0.10

Table B1. Computation times (in ms).

Independent flow segments

The mapping between inverse depth from motion and disparity (Equation A22) is first determined on the GPU. The cue combination then requires the solution of a linear least-squares problem at each pixel. Since the normal equations for this problem consist of only nine unique values, it is possible to both compose and solve them on the GPU. In a fashion similar to a nonseparable 2D convolution, the data are gathered from an 11 × 11 region surrounding the pixel. The high degree of local data reuse is handled efficiently by the texture cache.

Processing times

The processing times obtained for low- and high-resolution video using the hybrid CPU/GPU implementation and a CPU-only implementation (standard C++ running on a single core) are shown in Table B1. An Intel Xeon 2.5 GHz was used as CPU and an NVIDIA GeForce GTX 280 as GPU. The hybrid system operates at 30 frames per second (fps) for 320 × 256 and at 21 fps for 640 × 512 resolutions, whereas the CPU-only system requires approximately 3 and 10 s, respectively, to process a single frame.

In the Gabor filtering stage, we observe 50- and 120-fold speedups. The increase in speedup at the high resolution indicates that the GPU is not yet fully saturated at the low resolution. As compared to the filtering stage, a much higher number of operations per data point are required to compute edge structure, optical flow, and binocular disparity. The GPU can use both its stream processors and texture units and for this reason very large speedups are observed over the CPU-only implementation.

The egomotion times reported here were obtained with a configuration involving 32 initializations, 30 iterations, and 10,000 samples. An equal number of samples were used at both resolutions. At these settings, egomotion estimation is about fifty times faster on the GPU.

Finally, the independent flow segments stage is again computationally intensive with a high degree of data locality. With our texture-based implementation, we can achieve 90- and 180-fold speedups over the CPU-only implementation.

Acknowledgments

This work was funded by the European Projects ECOVISION (IST-2001-32114), DRIVSCO (IST-2002-016276), and EYESHOTS (IST-2007-217077). F.W. acknowledges funding from the BMBF BCCN W3-Project. The help from Dirk Kraft in making the benchmark sequence is gratefully acknowledged.

Commercial relationships: none.
Corresponding author: Marc M. Van Hulle.
Email: marc@neuro.kuleuven.be.
Address: Laboratorium voor Neuro- en Psychofysiologie, K.U. Leuven, O&N II Herestraat 49-bus 1021, 3000 Leuven, Belgium.

References

Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 2,* 284–299.

Adiv, G. (1985). Determining three-dimensional motion and structure from optical flow generated by several moving objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 7,* 384–401.

Argyros, A. A., Trahanias, P. E., & Orphanoudakis, S. C. (1998). Robust regression for the detection of independent 3D motion by a binocular observer. *Real-Time Imaging, 4,* 125–141.

Bergen, J., Anandan, P., Hanna, K., & Hingorani, R. (1992). Hierarchical model-based motion estimation. *Proceedings of the Second European Conference on Computer Vision, Italy, 588,* 237–252.

Bremmer, F., Kubischik, M., Pekel, M., Hoffmann, K. P., & Lappe, M. (2010). Visual selectivity for heading in monkey area MST. *Experimental Brain Research, 200,* 51–60.

Brenner, E. (1991). Judging object motion during smooth pursuit eye-movements: The role of optic flow. *Vision Research, 31,* 1893–1902.

Brenner, E., & van den Berg, A. V. (1996). The special role of distant structures in perceived object velocity. *Vision Research, 36,* 3805–3814.

Browning, N. A., Grossberg, S., & Mingolla, E. (2009). Cortical dynamics of navigation and steering in natural scenes: Motion-based object segmentation, heading, and obstacle avoidance. *Neural Networks, 22,* 1383–1398.

Burt, P. J., & Adelson, E. H. (1983). The Laplacian pyramid as a compact image code. *IEEE Transactions on Communications, 31,* 532–540.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 8,* 679–698.

Chen, Y. H., & Qian, N. (2004). A coarse-to-fine disparity energy model with both phase-shift and position-shift receptive field mechanisms. *Neural Computation, 16,* 1545–1577.

Chiuso, A., Brockett, R., & Soatto, S. (2000). Optimal structure from motion: Local ambiguities and global estimates. *International Journal of Computer Vision, 39,* 195–228.

Daugman, J. G. (1985). Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two dimensional visual cortical filters. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 2,* 1160–1169.

Davison, A. J., Reid, I. D., Molton, N. D., & Stasse, O. (2007). MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29,* 1052–1067.

Demirdjian, D., & Horaud, R. (2000). Motion-egomotion discrimination and motion segmentation from image-pair streams. *Computer Vision and Image Understanding, 78,* 53–68.

Duffy, C. J., & Wurtz, R. H. (1995). Response of monkey MST neurons to optic flow stimuli with shifted centers of motion. *Journal of Neuroscience, 15,* 5192–5208.

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous localization and mapping: Part I. The essential algorithms. *IEEE Robotics & Automation Magazine, 13,* 99–108.

Dyde, R. T., & Harris, L. R. (2008). The influence of retinal and extra-retinal motion cues on perceived object motion during self-motion. *Journal of Vision, 8*(14):5, 1–10, http://www.journalofvision.org/content/8/14/5, doi:10.1167/8.14.5. [PubMed] [Article]

Eifuku, S., & Wurtz, R. H. (1999). Response to motion in extrastriate area MSTl: Disparity sensitivity. *Journal of Neurophysiology, 82,* 2462–2475.

Ess, A., Leibe, B., Schindler, K., & Van Gool, L. (2009). Robust multiperson tracking from a mobile platform. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31,* 1831–1846.

Fleet, D. J., & Jepson, A. D. (1990). Computation of component image velocity from local phase information. *International Journal of Computer Vision, 5,* 77–104.

Fleet, D. J., & Jepson, A. D. (1993). Stability of phase information. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 15,* 1253–1268.

Fleet, D. J., Jepson, A. D., & Jenkin, M. R. M. (1991). Phase-based disparity measurement. *CVGIP-Image Understanding, 53,* 198–210.

Franke, U., & Heinrich, S. (2002). Fast obstacle detection for urban traffic situations. *IEEE Transactions on Intelligent Transportation Systems, 3,* 173–181.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics, 36,* 193–202.

Gautama, T., & Van Hulle, M. M. (2002). A phase-based approach to the estimation of the optical flow field using spatial filtering. *IEEE Transactions on Neural Networks, 13,* 1127–1136.

Gibson, J. J. (1950). *The perception of the visual world.* Boston: Houghton Mifflin.

Gogel, W. C. (1982). Analysis of the perception of motion concomitant with a lateral motion of the head. *Perception & Psychophysics, 32,* 241–250.

Gogel, W. C. (1990). A theory of phenomenal geometry and its applications. *Perception & Psychophysics, 48,* 105–123.

Granlund, G. H. (1978). In search of a general picture processing operator. *Computer Graphics and Image Processing, 8,* 155–173.

Grossberg, S. (2003). How does the cerebral cortex work? Development, learning, attention, and 3D vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews, 2,* 47–76.

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

Henriksson, L., Hyvärinen, A., & Vanni, S. (2009). Representation of cross frequency spatial phase relationships in human visual cortex. *Journal of Neuroscience, 29,* 14342–14351.

Ilg, U. J., Schumann, S., & Their, P. (2004). Posterior parietal cortex neurons encode target motion in world-centered coordinates. *Neuron, 43,* 145–151.

Irani, M., & Anandan, P. (1998). A unified approach to moving object detection in 2D and 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 577–589.

Koenderink, J. J., & van Doorn, A. J. (1987). Facts on optic flow. *Biological Cybernetics, 56,* 247–254.

Komatsu, H., & Wurtz, R. H. (1988). Relation of cortical areas MT and MST to pursuit eye movements: 1. Localization and visual properties of neurons. *Journal of Neurophysiology, 60,* 580–603.

Kovesi, P. (1999). Image features from phase congruency. *Videre, 1,* 1–26.

Lappe, M., Bremmer, F., Pekel, M., Thiele, A., & Hoffmann, K. P. (1996). Optic flow processing in monkey STS: A theoretical and experimental approach. *Journal of Neuroscience, 16,* 6265–6285.

Lappe, M., & Rauschecker, J. P. (1993). A neural network for the processing of optic flow from egomotion in man and higher mammals. *Neural Computation, 5,* 374–391.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., et al. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation, 1,* 541–551.

Leibe, B., Schindler, K., Cornelis, N., & Van Gool, L. (2008). Coupled object detection and tracking from static cameras and moving vehicles. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30,* 1683–1698.

Lindholm, E., Nickolls, J., Oberman, S., & Montrym, J. (2008). NVIDIA Tesla: A unified graphics and computing architecture. *IEEE Micro, 28,* 39–55.

Little, J. J., Poggio, T., & Gamble, E. B. (1988). Seeing in parallel—The vision machine. *International Journal of Supercomputing Applications and High Performance Computing, 2,* 13–28.

Logan, D. J., & Duffy, C. J. (2006). Cortical area MSTd combines visual cues to represent 3-D self-movement. *Cerebral Cortex, 16,* 1494–1507.

Longuet-Higgins, H. C., & Prazdny, K. (1980). The interpretation of a moving retinal image. *Proceedings of the Royal Society of London B: Biological Sciences, 208,* 385–397.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60,* 91–110.

Matsumiya, K., & Ando, H. (2009). World-centered perception of 3D object motion during visually guided self-motion. *Journal of Vision, 9*(1):15, 1–13, http://www.journalofvision.org/content/9/1/15, doi:10.1167/9.1.15. [PubMed] [Article]

Maunsell, J. H. R., & Van Essen, D. C. (1983). The connections of the middle temporal visual area (MT) and their relationship to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience, 3,* 2563–2586.

Mechler, F., Reich, D. S., & Victor, J. D. (2002). Detection and discrimination of relative spatial phase by V1 neurons. *Journal of Neuroscience, 22,* 6129–6157.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision—2 cortical pathways. *Trends in Neurosciences, 6,* 414–417.

Morrone, M. C., & Burr, D. C. (1988). Feature detection in human vision: A phase dependent energy model. *Proceedings of the Royal Society of London B: Biological Sciences, 235,* 221–245.

Nelson, R. C. (1991). Qualitative detection of motion by a moving observer. *International Journal of Computer Vision, 7,* 33–46.

Nister, D., Naroditsky, O., & Bergen, J. (2006). Visual odometry for ground vehicle applications. *Journal of Field Robotics, 23,* 3–20.

Ogale, A. S., Fermuller, C., & Aloimonos, Y. (2005). Motion segmentation using occlusions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27,* 988–992.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science, 249,* 1037–1041.

Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1997). Encoding of binocular disparity by complex cells in the cat's visual cortex. *Journal of Neurophysiology, 77,* 2879–2909.

Oliensis, J. (2005). The least-squares error for structure from infinitesimal motion. *International Journal of Computer Vision, 61,* 259–299.

Orban, G. A. (2008). Higher order visual processing in macaque extrastriate cortex. *Physiological Reviews, 88,* 59–89.

Pauwels, K., & Van Hulle, M. M. (2009). Optic flow from unstable sequences through local velocity constancy maximization. *Image and Vision Computing, 27,* 579–587.

Perna, A., Tosetti, M., Montanaro, D., & Morrone, M. C. (2008). BOLD response to spatial phase congruency in human brain. *Journal of Vision, 8*(10):15, 1–15, http://www.journalofvision.org/content/8/10/15, doi:10.1167/8.10.15. [PubMed] [Article]

Perrone, J. A. (1992). Model for the computation of self-motion in biological systems. *Journal of the Optical Society of America A, Optics, Image Science, and Vision, 9,* 177–194.

Perrone, J. A., & Krauzlis, R. J. (2008a). Spatial integration by MT pattern neurons: A closer look at pattern-to-component effects and the role of speed tuning. *Journal of Vision, 8*(9):1, 1–14, http://www.journalofvision.org/content/8/9/1, doi:10.1167/8.9.1. [PubMed] [Article]

Perrone, J. A., & Krauzlis, R. J. (2008b). Vector subtraction using visual and extraretinal motion signals: A new look at efference copy and corollary discharge theories. *Journal of Vision, 8*(14):24, 1–14, http://www.journalofvision.org/content/8/14/24, doi:10.1167/8.14.24. [PubMed] [Article]

Poggio, T., Gamble, E. B., & Little, J. J. (1988). Parallel integration of vision modules. *Science, 242,* 436–440.

Pollen, D., & Ronner, S. (1981). Phase relationships between adjacent simple cells in the visual cortex. *Science, 212,* 1409–1411.

Pugeault, N., Pauwels, K., Pilz, F., Van Hulle, M. M., & Krüger, N. (in press). A three level architecture for model-free detection and tracking of independently moving objects. *Proceedings of the International Conference on Computer Vision Theory and Applications, Angers, France*.

Qian, N., & Mikaelian, S. (2000). Relationship between phase and energy methods for disparity computation. *Neural Computation, 12,* 279–292.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience, 2,* 1019–1025.

Ringach, D. (2002). Spatial structure and symmetry of simple cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology, 88,* 455–463.

Rokers, B., Cormack, L. K., & Huk, A. C. (2009). Disparity- and velocity-based signals for three-dimensional motion perception in human MT. *Nature Neuroscience, 12,* 1050–1055.

Royden, C. S. (1997). Mathematical analysis of motion-opponent mechanisms used in the determination of heading and depth. *Journal of the Optical Society of America A: Optics, Image Science, and Vision, 14,* 2128–2143.

Royden, C. S. (2002). Computing heading in the presence of moving objects: A model that uses motion-opponent operators. *Vision Research, 42,* 3043–3058.

Royden, C. S., & Hildreth, E. C. (1996). Human heading judgments in the presence of moving objects. *Perception & Psychophysics, 58,* 836–856.

Rushton, S. K., Bradshaw, M. F., & Warren, P. A. (2007). The pop out of scene-relative object movement against retinal motion due to self-movement. *Cognition, 105,* 237–245.

Rushton, S. K., Harris, J. M., Lloyd, M. R., & Wann, J. P. (1998). Guidance of locomotion on foot uses perceived target location rather than optic flow. *Current Biology, 8,* 1191–1194.

Rushton, S. K., & Warren, P. A. (2005). Moving observers, relative retinal motion and the detection of object movement. *Current Biology, 15,* R542–R543.

Rust, N. C., Mante, V., Simoncelli, E. P., & Movshon, J. A. (2006). How MT cells analyze the motion of visual patterns. *Nature Neuroscience, 9,* 1421–1431.

Sabatini, S. P., Gastaldi, G., Solari, F., Pauwels, K., Van Hulle, M., Díaz, J., et al. (2010). A compact harmonic code for early vision based on anisotropic frequency channels. *Computer Vision and Image Understanding, 114,* 681–699.

Sanger, T. D. (1988). Stereo disparity computation using Gabor filters. *Biological Cybernetics, 59,* 405–418.

Sawhney, H. S., Guo, Y. L., & Kumar, R. (2000). Independent motion detection in 3D scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 1191–1199.

Schindler, K., Suter, D., & Wang, H. (2008). A model-selection framework for multibody structure-and-motion of image sequences. *International Journal of Computer Vision, 79,* 159–177.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29,* 411–426.

Sharma, R., & Aloimonos, Y. (1996). Early detection of independent motion from active control of normal image flow patterns. *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics, 26,* 42–52.

Simoncelli, E. P., & Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research, 38,* 743–761.

Tanaka, K., Sugita, Y., Moriya, M., & Saito, H. A. (1993). Analysis of object motion in the ventral part of the medial superior temporal area of the macaque visual cortex. *Journal of Neurophysiology, 69,* 128–142.

Thompson, W. B., & Pong, T. C. (1990). Detecting moving objects. *International Journal of Computer Vision, 4,* 39–57.

Torr, P. H. S. (1998). Geometric motion segmentation and model selection. *Philosophical Transactions of the Royal Society of London A, Mathematical, Physical, and Engineering Sciences, 356,* 1321–1340.

Vaudrey, T., Rabe, C., Klette, R., & Milburn, J. (2008). Differences between stereo and motion behaviour on synthetic and real-world stereo sequences. *Proceedings of the 23rd International Conference on Image and Vision Computing, New Zealand, 1*, 1–6.

Wallis, G., & Rolls, E. T. (1997). Invariant face and object recognition in the visual system. *Progress in Neurobiology, 51,* 167–194.

Wang, C. C., Thorpe, C., Thrun, S., Hebert, M., & Durrant-Whyte, H. (2007). Simultaneous localization, mapping and moving object tracking. *International Journal of Robotics Research, 26,* 889–916.

Wang, W. D., & Duncan, J. H. (1996). Recovering the three-dimensional motion and structure of multiple moving objects from binocular image flows. *Computer Vision and Image Understanding, 63,* 430–446.

Wangsiripitak, S., & Murray, D. W. (2009). Avoiding moving outliers in visual SLAM by tracking moving objects. *Proceedings of the International Conference on Robotics and Automation, Kobe, Japan, 1,* 375–380.

Warren, P. A., & Rushton, S. K. (2007). Perception of object trajectory: Parsing retinal motion into self and object movement components. *Journal of Vision, 7*(11):2, 1–11, http://www.journalofvision.org/content/7/11/2, doi:10.1167/7.11.2. [PubMed] [Article]

Warren, P. A., & Rushton, S. K. (2008). Evidence for flow-parsing in radial flow displays. *Vision Research, 48,* 655–663.

Warren, P. A., & Rushton, S. K. (2009a). Perception of scene-relative object movement: Optic flow parsing and the contribution of monocular depth cues. *Vision Research, 49,* 1406–1419.

Warren, P. A., & Rushton, S. K. (2009b). Optic flow processing for the assessment of object movement during ego movement. *Current Biology, 19,* 1555–1560.

Warren, W. H., & Saunders, J. A. (1995). Perceiving heading in the presence of moving objects. *Perception, 24,* 315–331.

Wedel, A., Rabe, C., Meissner, A., Franke, U., & Cremers, D. (2009). Detection and segmentation of independently moving objects from dense scene flow. *Proceedings of the 7th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, Bonn, Germany, 5681,* 14–27.

Wexler, M., Lamouret, I., & Droulez, J. (2001). The stationarity hypothesis: An allocentric criterion in visual perception. *Vision Research, 41,* 3023–3037.

Wexler, M., Panerai, F., Lamouret, I., & Droulez, J. (2001). Self-motion and the perception of stationary objects. *Nature, 409,* 85–88.

Williams, B., & Reid, I. (2010). On combining visual SLAM and visual odometry. *Proceedings of the International Conference on Robotics and Automation, Anchorage, Alaska, 1,* 3494–3500.

Wong, T. T., Leung, C. S., Heng, P. A., & Wang, J. Q. (2007). Discrete wavelet transform on consumer-level graphics hardware. *IEEE Transactions on Multimedia, 9,* 668–673.

Yuan, C., Medioni, G., Kang, J. M., & Cohen, I. (2007). Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 29,* 1627–1641.

Zemel, R. S., & Sejnowski, T. J. (1998). A model for encoding multiple object motions and self-motion in area MST of primate visual cortex. *Journal of Neuroscience, 18,* 531–547.

Zhang, T., & Tomasi, C. (2002). On the consistency of instantaneous rigid motion estimation. *International Journal of Computer Vision, 46,* 51–79.