Using the properties of primate motion sensitive neurons to extract camera motion and depth from brief 2-D monocular image sequences

John A. Perrone[1][0000−0001−7540−3240], Michael J. Cree[2][0000−0002−2973−2710], and Mohammad Hedayati[1,2][0000−0002−7266−0986]

[1] School of Psychology, University of Waikato, Hamilton 3240, New Zealand
john.perrone@waikato.ac.nz
[2] School of Engineering, University of Waikato, Hamilton 3240, New Zealand
michael.cree@waikato.ac.nz, hedi.hedayati@waikato.ac.nz

Abstract. Humans and most animals can run/fly and navigate efficiently through cluttered environments while avoiding obstacles in their way. Replicating this advanced skill in autonomous robotic vehicles currently requires a vast array of sensors coupled with computers that are bulky, heavy and power hungry. The human eye and brain have had millions of years to develop an efficient solution to the problem of visual navigation and we believe that it is the best system to reverse engineer. Our brain and visual system appear to use a very different solution to the visual odometry problem compared to most computer vision approaches. We show how a neural-based architecture is able to extract self-motion information and depth from monocular 2-D video sequences and highlight how this approach differs from standard CV techniques. We previously demonstrated how our system works during pure translation of a camera. Here, we extend this approach to the case of combined translation and rotation.

Keywords: biologically-based motion sensor · visual odometry · monocular visual sensor · optical flow · image motion · depth-from-motion

1 Introduction

It has long been known that the motion occurring on the back of our eyes (optic flow) provides a rich source of information regarding our own movement through the world ('self-motion') as well as the 3-D layout of the scene in front of us [10, 18, 19]. We are able to extract this information from just the motion projected on the retina of a single eye. Hence, we often perceive depth while viewing 2-D movies and can navigate and avoid obstacles while playing 2-D video games. This 3-D from 2-D motion extraction process occurs within fractions of a second which grants us the ability to run/drive through complex environments at speed while avoiding obstacles in our way.

Many animals share this amazing skill with us and there is a large industry dedicated to emulating this biological navigational ability in software/hardware

in order to solve the visual odometry problem [9, 35]. However, the disparity between what animals and machines can achieve is currently still wide and attempts to map out of the environment in front of a moving robot or vehicle tend to rely on active sensors (e.g., LIDAR) or binocular systems [9, 35]. The former tend to be bulky and power hungry and the latter require the simultaneous processing of dual video streams and often require careful camera calibration to be able to exploit epipolar geometry to simplify feature matching. A passive sensor that worked with standard monocular video inputs would have many advantages.

The primate visual system seems to have taken a very different approach to most computer vision techniques for extracting odometry and depth information. The majority of computer vision methods either find corresponding pixels (indirect) [22, 17, 36, 33] or minimise photometric error (direct) [15, 6, 5] between two frames to estimate the camera displacement and depth information. These models can be further categorised as sparse, which only consider a set of independent points to solve the correspondence problems [22, 17], and as dense which use all pixels in the frame to estimate camera motion [36, 33, 6]. In contrast, the primate visual system has a series of stages ('the visual motion pathway' [23, 3]) where the image motion is first registered using banks of spatiotemporal filters, and after a series of integration stages [21, 3], produces a signal proportional to the image speed [14]. The brain then recovers self-motion (odometry) information at a relatively late stage of processing [3]. What are the advantages to be gained from this long chain of computational steps? Some of the motion processing steps seem to provide an economical use of neural hardware [25] but other potential advantages are currently unknown.

Discovering the benefits of neural-based approaches to odometry and depth estimation would be very useful in the design of smart sensors for robots and autonomous vehicles. Unfortunately, we do not yet have a full understanding of how the human or non-human primate visual systems are able to recover depth from 2-D video sequences. Some motion processing areas of the brain have been well studied [2] but many aspects of how visual motion is analyzed remain a mystery.

We have recently implemented a scheme based on the properties of neurons in the primate visual system that is able to solve the depth estimation problem [27]. This technique is able to derive 3-D scene depth estimates from the 2-D image motion generated during pure forward translation of the camera. Here we extend this system to work with a combined translation and rotation of the camera. We present a 'proof of principle' test that demonstrates that a system based on primate neuron properties is able to extract camera motion information as well as 3-D depth information from a brief 2-D video sequence. We see this as a first step in the process of comparing biologically-based methods with conventional computer vision approaches to odometry and depth recovery. We eventually hope to discover and highlight the advantages that the brain has developed through millions of years of evolution.

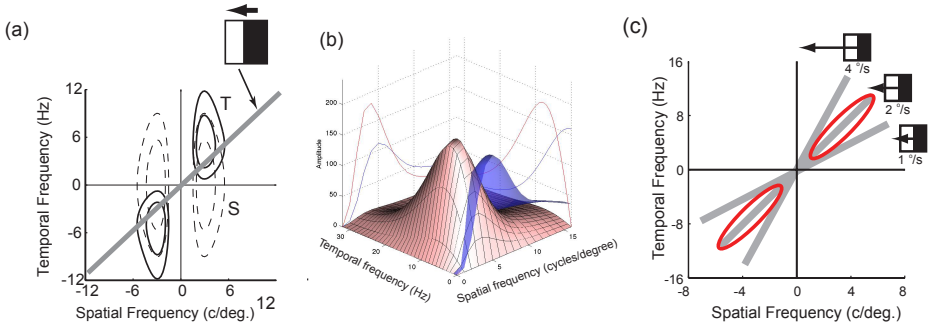

Fig. 1. Frequency space representations of the model early stage motion sensor filters and the representation of edge motion in spatial frequency (f_s)–temporal frequency (f_t) space. (a) Plan view of two types of model filters. A moving edge has a spectrum that falls on an oriented line with the slope proportional to the speed [38, 39]. (b) 3-D surface plot of the upper right quadrant of $f_t - f_s$ space with the amplitude spectrum of the sustained type (S) of spatiotemporal filter shown in blue and the transient type (T) in red. (c) Representation of multiple edge speeds in the frequency domain. The red ovals represent a speed tuned filter tightly tuned to a particular edge speed.

2 Extracting velocity vector flow fields (optic flow) from image sequences.

Our approach to the visual odometry/obstacle avoidance problem makes use of measurements of the image motion occurring in a video input and these measurements are based on intensity changes in the images rather than on feature matching [9, 35, 22, 17]. Fig. 1 shows the early stage filtering used in our system for deriving image velocity estimates from an 8-frame video sequence.

2.1 Early stage motion sensors based on neurons in primate primary visual cortex (V1)

In a plot of spatial frequency versus temporal frequency, a moving edge has a spectrum that falls along a line (grey line in Fig. 1a) with the slope of the line proportional to the edge speed [38, 39]. The first stage in determining the slope of the spectrum (speed of the edge) in our system is to use two spatiotemporal frequency tuned filters, one with low pass temporal frequency tuning (dashed curve in Fig. 1 marked S for 'sustained') and one with bandpass tuning (solid lines in Fig. 1a marked T for 'transient').

The S filters are separable and made up of separate spatial and temporal frequency functions as given below. The T filters are one-quadrant separable and are constructed from individual non-directional spatiotemporal filters using the combination rules specified by Watson & Ahumada [39].

The temporal frequency amplitude response function of the S-type filter is given by:

$$\tilde{f}_{\text{sust}}(f_t) = e^{-f_t^2\sigma^2/2}\, e^{-i2\pi f_t \theta}. \tag{1}$$

where f_t is the temporal frequency measured in Hz and $i = \sqrt{-1}$. The θ term (phase) controls the temporal delay (lag) of the response and σ controls the spread of the Gaussian.

The T-type filter temporal frequency tuning function is band-pass in shape and is given by:

$$\tilde{f}_{\text{trans}} = 0.25\tilde{f}_{\text{sust}}(f_t)f_t i. \tag{2}$$

The magnitude part of both of these functions are good matches [25] to the temporal frequency tuning functions often observed in primate V1 neurons [8, 12]. The two functions (S = blue, T = red) can be seen on the right wall of Fig. 1b and the parameter values can be found in [25].

The spatial frequency tuning functions of the front-end spatiotemporal filters (Fig. 1a, b) are based on the difference of difference of Gaussians with separation (dDOGs) function used by Hawken and Parker [11] to fit their primate V1 spatial frequency tuning data. (see [24]). The two S & T spatial frequency functions can be seen on the back wall of Fig. 1b. Only one size of filter is shown in the figure. The full model uses four different sizes in $\log_2$ steps [26].

2.2 Gradient optical flow models versus the introduction of tight speed tuning

In the space domain (inverse Fourier transformed versions), the S and T spatial filters look similar to standard Gabor or Difference of Gaussian (DoG) functions often used in computer vision. They come in both even and odd (quadrature) versions and the latter can act as a 1^{st} derivative spatial operator over space (x). The T-type temporal bandpass filter is biphasic in time and can also be considered to be a 1^{st} derivative operator over time (t). Therefore the ratio of the T and S filter outputs (T/S) is equal to $\Delta x/\Delta t = V$ (the image speed). The common gradient-based optical flow methods used in computer vision [13, 40, 7] use this operation inside the optical flow energy function to find V. Therefore an indication of the speed of a moving image feature is available at the very earliest stage of filtering in the primate visual system yet, for some reason, the estimation of the actual speed is delayed. The electrophysiological evidence points to speed-tuned sensors prior to the direct estimation of speed. Subsequent to the S&T filter stage (Fig. 1a), there are filters (Middle Temporal or MT neurons) that are precisely tuned to a particular image speed and their output is not linear with input speed [20, 29]. It is only at a post-MT stage (see below) that an output proportional to input speed is found in the primate visual system [14].

We follow this primate motion processing pathway design and introduce tight speed tuning after the spatiotemporal filtering stage. The spatial filters in our model look superficially like standard Gabor functions but they differ in important ways [24, 25]. The spatial and temporal frequency tuning functions used in the model motion sensor have been especially designed to create a filter that is very selective to a particular spectrum orientation (edge speed) and to be precisely tuned to a narrow range of image speeds (red oval outline in Fig. 1c). This is done via a type of AND operation whereby the speed tuned filter gives a

large response whenever both the sustained and transient filter outputs are high and equal. For the S and T spectra profiles shown in Fig. 1b, this occurs along a locus that forms a straight line and is oriented relative to the two axes.

It has been shown that the slope of the line of intersection can be altered (and hence the speed tuning of the filter) by simply changing the weight of either the S or T energy output [25]. The speed tuned mechanism is therefore referred to as a 'Weighted Intersection Mechanism' (WIM) and is given by:

$$\text{WIM} = \frac{S + T}{|S - T| + \delta} \tag{3}$$

where δ is a constant that controls the bandwidth of the speed tuning (set to 12 in the model tests reported here).

This design is based on theoretical work using frequency representations of visual motion [30, 25, 38] and the model filters mimic the oriented spectral receptive fields of neurons found in the Middle Temporal (MT) region of the primate brain [29]. This speed tuning stage is an important and unique feature of our flow field estimation scheme.

2.3 Image velocity estimation from three separate MT unit outputs

The MT unit sensors in our optic flow estimation system do not output a signal proportional to the stimulus speed of the feature passing over them; they are speed and direction tuned only. In the primate visual system evidence for cells that respond in a linear fashion to the input speed only appear at a stage after MT, namely the dorsal Medial Superior Temporal area (MSTd) [14]. Based on a theory that outlines a possible velocity code used by the primate visual system [26], our velocity estimation stage replicates this MT to MSTd transition and derives a velocity signal from the outputs of a 'triad' of MT units that come in two spatial scales.

This system uses competition (via a 2nd derivative stage) between two different sized MT units tuned to speed MTv and MT2v as well as a unit tuned to $V/2$. The latter input has the same spatial size as the MTV unit. In frequency space (see Fig. 1c) the $2V$and $V/2$ units sit on either side of the ridge occupied by the main MTv unit (see red oval in Fig. 1c). They therefore isolate the correct location of the oriented line spectrum generated by an edge moving at speed V to a single velocity 'channel'. A precise speed signal is calculated using a centroid operation on the three triad unit outputs, which interpolates between the discrete MT unit speed tuning values $1, 2, 4, \ldots$ pixels/frame. The direction is similarly found with greater precision by using vector addition to estimate the direction from the output of multiple velocity sensors tuned to different directions [26].

3 Heading estimation and depth extraction

The direct approach to deriving camera motion (odometry) and depth from the optical flow field is to directly input the velocity vectors (or image brightness

changes) into the equations for observer motion [19] and solving for the camera parameters and depth. Again, the primate visual system seems to have taken a slightly different approach and has neural processing units designed to extract information about the observer's heading direction separately from information regarding depth [3]. The mechanism used seems to be based around a population of heading tuned units rather than a single neuron coding for all of the observer self-motion parameters and scene depth.

We follow this design and our approach makes use of a well-known property of flow fields that occurs during pure forward translation of an observer/camera; the image motion radiates out from a single point in the image (the focus of expansion or FOE) and this coincides with the heading direction [10]. The location of the FOE can be found using special 'heading detectors' or radial templates based on the properties of primate MSTd neurons [31, 32]. In our model an array of detectors tuned to a range of heading azimuth and elevation values (-50° to 50° in 2.5° steps for azimuth and elevation) is used to search for the FOE location in the image sequence.

A major problem encountered while attempting to estimate image motion is the aperture problem [41]; when just an edge is located in a motion sensor aperture, only the motion orthogonal to the edge direction can be detected and the estimated motion direction and speed are perturbed away from the true optic flow values. Our heading estimation units are very tolerant of noise in the flow field vector directions. As long as there are a sufficient number of vectors distributed across the field and the edge orientations causing the aperture problem are randomly distributed around the radial direction out from the putative FOE locations, the heading can still be estimated accurately [4, 27]. Once the FOE has been determined, the true direction of the image motion is constrained to lie along the radial direction (α) of a line joining the derived FOE location to the vector location. The corrected magnitude of the vector can be found from $V_c = V/\cos(\alpha - \beta)$ where β is the vector direction and V is the magnitude. This correction is only applied to vector locations where $\alpha - \beta < 70\,\mathrm{deg}$.

3.1 Heading estimation in the presence of camera rotation

Most camera motion scenarios with a moving vehicle or aerial platform include rotation of the camera, which adds a rotation component (R) to the vector flow field created via pure translation of the camera (T). The resultant flow vectors (T + R) produce a flow field that no longer has a focus of expansion that coincides with the heading direction [34, 37]. The flow is no longer purely radial and so the depth cannot be easily recovered from the image motion. This is known as the 'rotation problem' and somehow the T vectors need to be recovered from the T + R flow in order to determine the heading direction and depth. A biologically feasible method for this has been proposed [28]. Rotation activity is removed from the heading detector map distribution that is equivalent to vector subtraction at the local level (T + R − R = T). We use this same mechanism in our model.

3.2 Depth extraction from heading and radial flow

If heading direction (α, β) is known, and the position of two points w_1, w_2 are fixed in the world, the ratio of the distances to the two points D_1/D_2 can be found from s_1 and s_2, the image velocity vector magnitudes in the radial flow pattern. If the camera/observer's forward speed (V_O) is known, it is possible to obtain absolute values of D_1 and D_2. Therefore, our system first determines the heading direction (with rotation removed), derives the radial optic flow field and then estimates the depth of the points. An overall plan of the system can be found in [27] (see Fig. 2). This system is very different from the majority of computer vision approaches to the structure from motion problem [9, 35, 15]. After introducing a number of radically different approaches to the standard structure-from-motion estimation problem one may well ask if our system works? Are we able to estimate 3-D depth from a 2-D video sequence using these biologically-based motion filters? We next present a proof of principle that our system can solve the depth from 2-D motion problem.

4 Testing Methodology

We used a computer-controlled camera (Basler acA1920-150um) mounted on a Pan-Tilt unit attached to an X-Y translation table (Newmark CS Series XY Gantry-1500-1500-1). The camera (field of view = 42° horizontal and 26.3° vertical) moved towards a laboratory scene containing identifiable target objects (Fig. 2). The camera forward speed was 0.25 m/s while rotating to the left (from straight ahead) about a vertical axis at 2.5°/s. This scene contained a range of object sizes, contrasts and intensity distributions similar to what is commonly found in both indoor and outside environments. A series of eight frames (1984×1264 pixels) was extracted from the video stream at a 100 Hz sample rate. The output of the velocity code model develops over the eight-frame sequence and we use the output from the fourth frame as an estimate of the vector flow field. The current Matlab implementation is not capable of 'real-time' analysis but many of the model stages could be run in parallel.

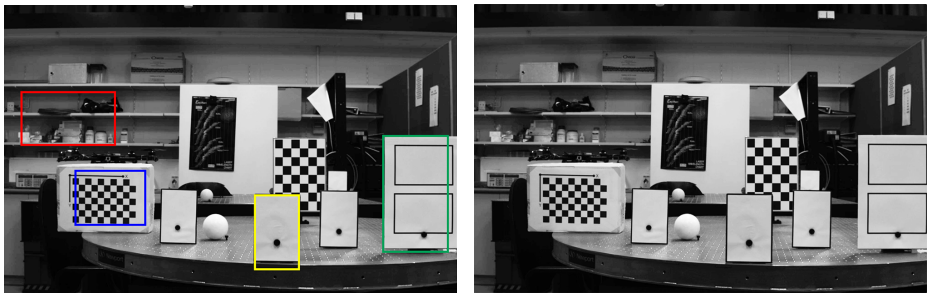

Fig. 2. Test input images. Frames 1 and 8 from the video sequence.

Fig. 2 shows the first and last frames of the eight-frame movie sequence with some of the objects that will be used to assess the depth extraction stage of the model. We used the blue zone (grid on left side of image) as a reference object and compared the depth estimates of other objects (each at a different distance) against the reference to see if the model could distinguish the depth location of the different objects. The far wall (red zone) was close to 4 m further than the reference object (true distances were 6 m and 2.1 m respectively). The yellow zone (middle card) was 0.6 m closer than the reference object (true yellow zone distance = 1.5 m) and the green zone object (card on extreme right) was 0.4 m in front of the standard (true green zone distance = 1.7 m). The distance between the green and yellow zone objects was 0.2 m.

5 Results

The raw vector flow field output from the velocity estimation stage is shown in Fig. 3a. This vector field was passed through the heading detector array and the activity distribution from the array is plotted in Fig. 3b. Without the rotation being removed the heading estimate was $(-35.3°, -.75°)$. The rotational flow vectors produced by the known camera rotation (2.5°/s to left) was removed from this raw flow field. The rotation-free heading detector distribution is shown in Fig. 3c. After rotation removal the estimated heading was $(2.51°, -0.75°)$ which is very close to the true heading $(2.5°, 0°)$.

Given the estimated heading direction (and associated expansion point in the image), the actual radial direction of each vector was determined and the vector magnitude was corrected. The resulting radial flow field is shown in Fig. 3d. The radial flow was used to estimate the distance to each point occupied by a vector in the output field. The estimated point cloud from each zone is noisy because slight variations in the vector magnitudes can result in large depth variations given the small projection angles involved. In order to quantify the depth estimation performance of the model we binned the estimates along the Z dimension (using the histogram function in Matlab) and the resulting frequency histograms are shown in Fig. 4.

The means (and standard deviations) for the blue and red depth distributions were 2.7 m (0.5) and 5.5 m (1.7). A t-test indicated that these two distributions are significantly different, $t(1, 164) = 15.0$, $p < .001$. Therefore, the model was able to extract depth from the monocular video sequence and successfully identified that the two zones were at different distances from the camera. For the yellow zone object (middle card) the actual separation from the reference object was −.6 m. The mean of the depth estimates was 2.1 m (0.9) and this was significantly different from the estimated standard distance (2.7 m), $t(1, 209) = 6.5$, $p < .001$. For the green zone object (Fig. 2 right card) the mean depth estimate was also 2.1 m (.71) and this was also significantly different from the standard, $t(1, 196) = 6.4$, $p < .001$. The separation distance from the standard was −0.4 m.

A test to see if the model was able to distinguish the average depth of the yellow and green zone objects (separated by 0.2 m) was non-significant,

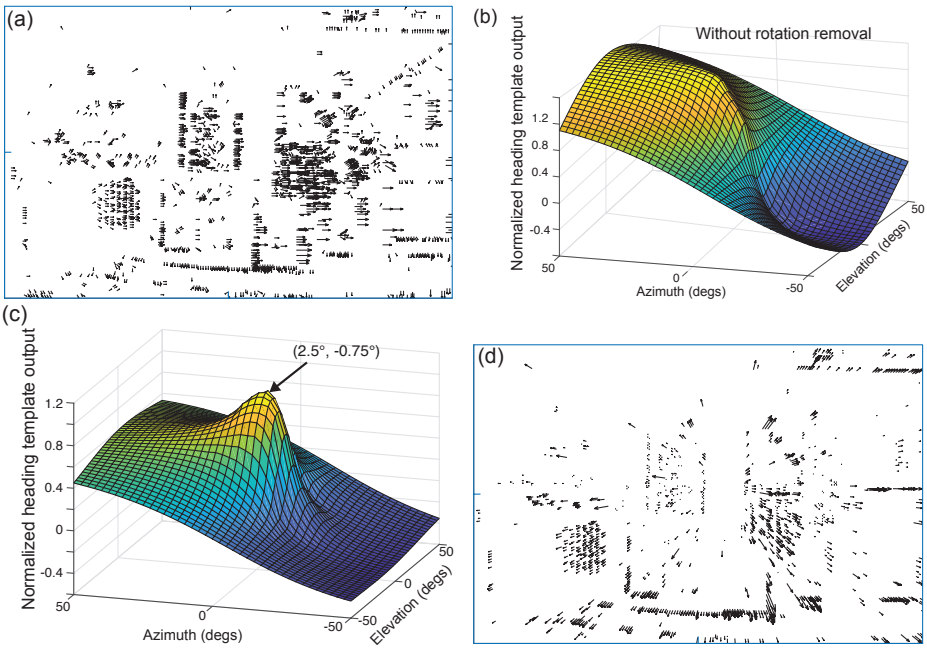

Fig. 3. (a) Initial vector flow field output from the velocity estimation stage of the model in response to the eight-frame test sequence. (b) Output of heading estimation stage of the model when the rotation is not removed from the flow field. The 3-D graph shows the activity of each heading detector (tuned to a particular azimuth and elevation value) in response to the vector flow field shown in a. (c) Heading template distribution after rotation removal. (d) Radial vector flow field output from the velocity estimation stage of the model in response to the eight-frame test sequence.

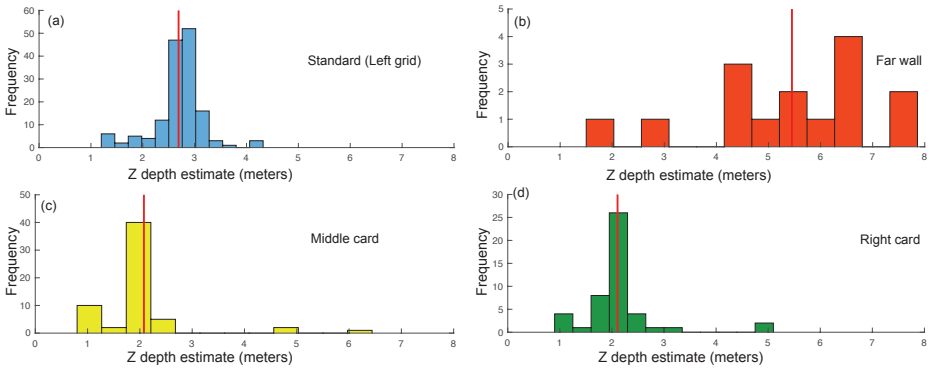

Fig. 4. Frequency histograms of the estimated distances (x-axis) found in each of the four image zones with the reference object at the top. The red vertical lines are the means of the distributions.

$t(1, 105) = .2$, $p = .8$. The means were different in the right direction (yellow, 2.08 m closer than green, 2.11 m) but the spreads of the two distributions were too high to detect the difference. It should be noted that human observers also cannot distinguish the depth of these two objects while viewing the eight-frame movie sequence.

6 Discussion

Our test demonstrates that camera motion (heading direction) and 3-D depth information can be recovered from an eight frame, monocular video sequence using a model based on the properties of neurons in the primate visual system. In order to be effective for obstacle avoidance the detection of objects along the path of travel needs to occur very quickly if evasive action is to be executed in time. The temporal filters in the early stage of our velocity detection algorithm have an epoch of around 200 ms and the timeline for the extraction of depth is not much longer than this because the later stages mainly involve integration of the first stage motion signals. We use a feedforward pipeline only and we argue that this gives it an advantage over schemes that rely on iterative searches for solving the velocity or odometry stages [1]. However, we recognize that the output from the later stages of the model could be used to refine the depth estimates over time. The depth distributions could be used to refine future estimates of the extracted depth signals and implement some form of Kalman filtering [16].

7 Conclusion

The primate visual system uses a different series of steps and stages from standard computer vision approaches in its attempt to derive information about self motion (odometry) and depth. The computation of the visual flow field is delayed in the biological system relative to computer vision methods with the inclusion of a speed-tuned sensor stage (see 2.2 above). This speed tuning stage could be part of a system for determining the overall direction of an object from its separate edge components [2, 24] but this is not known for certain. Once the flow field is determined, the estimation of depth and 3-D information from the visual motion occurring in a monocular video stream is still very difficult because the image motion is hard to measure accurately. The aperture problem perturbs the velocity vectors away from the true direction. We have taken a novel approach (based on knowledge of primate neuron properties) whereby we determine the heading direction using detectors tuned to radial motion that make use of redundant information distributed across the full visual field and which are therefore tolerant of the image motion noise introduced by the aperture problem. Once the heading direction is established it becomes relatively easy to derive the depth from the radial flow field. This is a different method to how depth from motion is usually estimated using computer vision approaches. We plan next to examine what these additional stages add to the depth and odometry recovery process and what advantages they provide.

References

1. Anandan, P.: A computational framework and an algorithm for the measurement of visual motion. Int. J. Comput. Vis. **2**(3), 283–310 (1989)
2. Born, R.T., Bradley, D.: Structure and function of visual area MT. Annual Review of Neuroscience **28**, 157–189 (2005)
3. Britten, K.H.: Mechanisms of self-motion perception. Annu Rev Neurosci **31**, 389–410 (2008)
4. Cree, M.J., Perrone, J.A., Anthonys, G., Garnett, A.C., Gouk, H.: Estimating heading direction from monocular video sequences using biologically-based sensors. Proceedings of the 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ) pp. 116–121 (2016)
5. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(3), 611–625 (2018)
6. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular SLAM. In: European Conference on Computer Vision (ECCV). Lecture Notes in Computer Science, vol. 8690, pp. 834–849 (2014)
7. Fortun, D., Bouthemy, P., Kervrann, C.: Optical flow modeling and computation: A survey. Journal of Computer Vision and Image Understanding **134**, 1–21 (2015)
8. Foster, K.H., Gaska, J.P., Nagler, M., Pollen, D.A.: Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the Macaque monkey. Journal of Physiology **365**, 331–363 (1985)
9. Fraundorfer, F., Scaramuzza, D.: Visual odometry part II: matching, robustness, optimization, and applications. IEEE Robotics & Automation Magazine **19**(2), 78–90 (2012)
10. Gibson, J.: The perception of the visual world. Houghton Mifflin, Boston (1950)
11. Hawken, M., Parker, A.: Spatial properties of neurons in the monkey striate cortex. Proceedings of the Royal Society of London B. **231**, 251–288 (1987)
12. Hawken, M., Shapley, R., Grosof, D.: Temporal frequency selectivity in monkey visual cortex. Journal of Neuroscience **13**, 477–492 (1996)
13. Horn, B.K.P., Schunk, B.G.: Determining optic flow. Artificial Intelligence **17**, 185–203 (1981)
14. Inaba, N., Shinomoto, S., Yamane, S., Takemura, A., Kawano, K.: MST neurons code for visual motion in space independent of pursuit eye movements. Journal of Neurophysiology **97**(5), 3473–3483 (2007)
15. Irani, M., Anandan, P.: About direct methods. In: International Workshop on Vision Algorithms. Lecture Notes in Computer Science, vol. 1883, pp. 267–277. Corfu, Greece (1999)
16. Kalman, R.E.: A new approach to linear filtering and prediction problems. Journal of Basic Engineering **82**(1), 35–45 (1960)
17. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: 6th IEEE and ACM International Symposium on Mixed and Augmented Reality. pp. 1–10. Nara, Japan (2007)
18. Koenderink, J.J., van Doorn, A.: Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer. Optica Acta **22**(9), 773–791 (1975)
19. Longuet-Higgins, H.C., Prazdny, K.: The interpretation of moving retinal images. Proceedings of the Royal Society of London B. **B 208**, 385–387 (1980)
20. Maunsell, J., Van Essen, D.: Functional properties of neurons in the middle temporal visual area of the Macaque monkey. I. selectivity for stimulus direction, speed, orientation. Journal of Neurophysiology **49**, 1127–1147 (1983)

21. Movshon, J.A., Adelson, E., Gizzi, M.S., Newsome, W.T.: The analysis of visual moving patterns. In: Chagas, C., Gross, C. (eds.) Pattern recognition mechanisms, pp. 117–151. Springer, New York (1985)
22. Mur-Artal, R., Montiel, J.M.M., Tardos, J.D.: ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015)
23. Nakayama, K.: Biological image motion processing: A review. Vision Res. **25**(5), 625–660 (1984)
24. Perrone, J.A.: A visual motion sensor based on the properties of V1 and MT neurons. Vision Res **44**(15), 1733–55 (2004)
25. Perrone, J.A.: Economy of scale: A motion sensor with variable speed tuning. Journal Of Vision **5**(1), 28–33 (2005)
26. Perrone, J.A.: A neural-based code for computing image velocity from small sets of middle temporal (MT/V5) neuron inputs. Journal of Vision **12**(8) (2012)
27. Perrone, J.A., Cree, M.J., Hedayati, M., Corlett, D.: Testing a biologically-based system for extracting depth from brief monocular 2-D video sequences. In: International Conference on Image and Vision Computing New Zealand. Auckland, New Zealand (Nov 2018)
28. Perrone, J.A., Krauzlis, R.: Vector subtraction using visual and extraretinal motion signals: A new look at efference copy and corollary discharge theories. Journal of Vision **8**(14), 1–14 (2008)
29. Perrone, J.A., Thiele, A.: Speed skills: measuring the visual speed analyzing properties of primate MT neurons. Nat Neurosci **4**(5), 526–32 (2001)
30. Perrone, J.A., Thiele, A.: A model of speed tuning in MT neurons. Vision Res **42**(8), 1035–51 (2002)
31. Perrone, J.: Model for the computation of self-motion in biological systems. Journal of the Optical Society of America **9**, 177–194 (1992)
32. Perrone, J., Stone, L.: A model of self-motion estimation within primate extrastriate visual cortex. Vision Research **34**, 2917–2938 (1994)
33. Ranftl, R., Vineet, V., Chen, Q., Koltun, V.: Dense monocular depth estimation in complex dynamic scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4058–4066 (2016)
34. Regan, D., Beverley, K.I.: How do we avoid confounding the direction we are looking and the direction we are moving? Science **215**(8), 194–196 (1982)
35. Scaramuzza, D., Fraundorfer, F.: Visual odometry part I: the first 30 years and fundamentals. IEEE Robotics & Automation Magazine **18**(4), 80–92 (2011)
36. Valgaerts, L., Bruhn, A., Mainberger, M., Weickert, J.: Dense versus sparse approaches for estimating the fundamental matrix. International Journal of Computer Vision **96**(2), 212–234 (2012)
37. Warren, W.: Optic flow. In: Chalupa, L., Werner, J. (eds.) The Visual Neurosciences, vol. 2, pp. 1247–1259. Bradford, Cambridge, Massachusetts (2003)
38. Watson, A., Ahumada, A.: A look at motion in the frequency domain. In: Tsotsos, J. (ed.) Motion:Perception and representation, pp. 1–10. Association for Computing Machinery, New York (1983)
39. Watson, A., Ahumada, A.: Model of human visual-motion sensing. Journal of the Optical Society of America A **2**, 322–342 (1985)
40. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Deepflow: Large displacement optical flow with deep matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1385–1392 (2013)
41. Wuerger, S., Shapley, R., Rubin, N.: 'On the visually perceived direction of motion' by Hans Wallach: 60 years later. Perception **25**(11), 1317–1367 (1996)