Estimating Heading Direction from Monocular Video Sequences Using Biologically-Based Sensors

Michael J. Cree*, John A. Perrone†, Gehan Anthonys*, Aden C. Garnett†, and Henry Gouk‡
*School of Engineering, University of Waikato, Hamilton 3240, New Zealand
†School of Psychology, University of Waikato, Hamilton 3240, New Zealand
‡Department of Computer Science, University of Waikato, Hamilton 3240, New Zealand
m.cree@ieee.org, john.perrone@waikato.ac.nz

***Abstract*—The determination of one's movement through the environment (visual odometry or self-motion estimation) from monocular sources such as video is an important research problem because of its relevance to robotics and autonomous vehicles. The traditional computer vision approach to this problem tracks visual features across frames in order to obtain 2-D image motion estimates from which the camera motion can be derived. We present an alternative scheme which uses the properties of motion sensitive cells in the primate brain to derive the image motion and the camera heading vector. We tested heading estimation using a camera mounted on a linear translation table with the line of sight of the camera set at a range of angles relative to straight ahead (0° to 50° in 10° steps). The camera velocity was also varied (0.2, 0.4, 0.8, 1.2, 1.6 and 2.0 $\mathrm{m/s}$). Our biologically-based method produced accurate heading estimates over a wide range of test angles and camera speeds. Our approach has the advantage of being a one-shot estimator and not requiring iterative search techniques for finding the heading.**

***Index Terms*—visual odometry, visual sensor, image motion**

I. Introduction

Humans are very adept at estimating their instantaneous heading direction from 2-D visual motion information [1], [2]. The determination of one's heading vector relative to the world is a critical step in obstacle avoidance and is a precursor for determining the relative depth of objects using 2-D image motion [3]–[5]. However the accurate measurement of 2-D image motion (optical flow) is difficult [6] and heading estimation is complicated by the presence of rotation during translation of the observer/camera [7]. Somehow humans and most biological species have overcome these problems and can safely navigate through complex environments using mainly 2-D image motion information.

Nistér coined the term *visual odometry* (VO) [8] for the process of recovering motion from visual input. There is enormous attention in the computer vision and robotics communities in developing systems for visual odometry because it is a challenging research problem and has potential to provide better data than is currently obtained from wheel odometry, GPS and inertial sensors [9], [10]. Most interest has been in stereo VO because it simplifies the problem and provides scale. Nevertheless much useful information can be gained from monocular vision and the scale problem can be overcome by the use of other sensors. Monocular VO can be divided into the feature based methods and the intensity based methods. In feature based methods salient features are extracted from each image and tracked from frame to frame, feeding into an estimation of the motion [8], [11], [12]. Each stage of the computation is subject to error and these errors accumulate through the pipeline limiting the achievable accuracy of the odometry. In contrast, intensity based methods exploit all information in the image or in subregions of the image [13], [14], but are not robust to occlusions and suffer from greater computational complexity.

Experiments carried out with humans using dense fields of identical moving dots and with very brief dot-lifetimes [2] suggest that humans do not use the feature tracking methods that form the core of most VO techniques. Instead, the primate visual system has evolved sophisticated neural mechanisms for registering 2-D image motion directly without the need to infer it from the changing positions of features such as corners [15], [16]. This motion analysis occurs very quickly ($\approx$ 200 ms) and no iterative searching for 'best matches' is required [9], [10]. Such search techniques are not easily implemented in biological systems and have only become viable in computer vision systems relatively recently because of the rapid advances in processor speeds.

Here we present an alternative method for carrying out VO that makes use of techniques based on the known properties of cells in the primate brain [5], [6], [17], [18]. The method registers image motion directly using motion sensors that emulate those found in the motion processing areas of the primate brain. Vector flow fields are generated from brief (8 frames, 233 ms) monocular video sequences and these flow fields are used to determine camera heading using specialised detectors based on those found in primates [19]. Because our method uses all image data it it could be considered an intensity based VO method but, unlike other intensity based methods, it is robust in the presence of occlusions. We believe our approach is a step towards replicating the powerful and effective navigational abilities of humans and many animals.

We tested the model over a range of heading directions using a video camera mounted on a precision linear translation rail. For comparison, we also tested standard computer-vision based methods for calculating optical flow from detected

978-1-5090-2748-4/16/$31.00

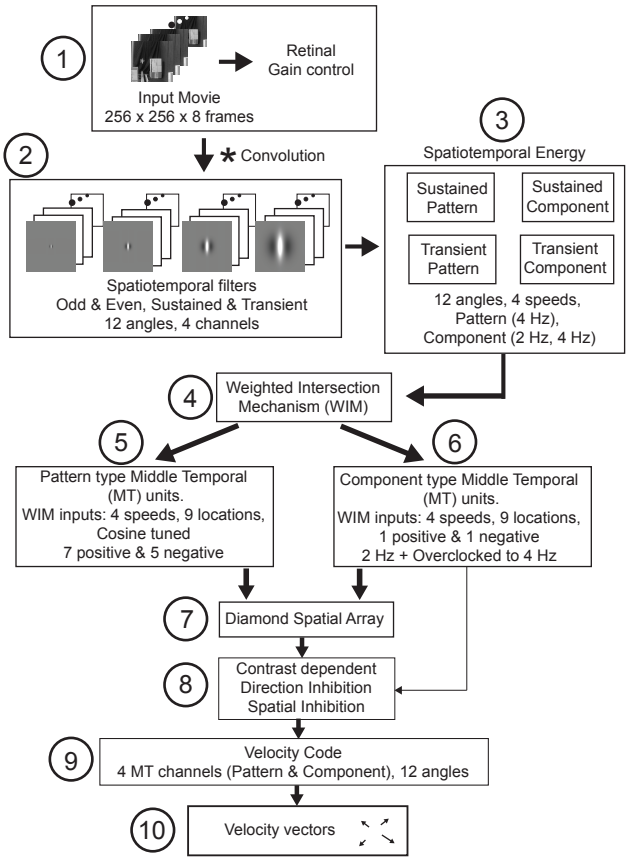


Fig. 1. Overview of velocity vector extraction stage of the model. Details can be found in Ref. [6].

and tracked features, with a least squares approximation for estimating the heading direction.

II. Model Description

A. Image velocity estimation

The image velocity estimation model has been described in detail previously [6]. The overall plan is shown in Fig. 1 which highlights the main stages of the code used to extract the optical flow vectors from the image sequences we tested. Each stage of the model is based on the known properties of motion sensitive neurones located along the 'visual motion pathway' of the primate brain [15], [16], [20]. As with many other image analysis tools, the image movie sequence is first convolved with a bank of spatiotemporal filters (step 1, in Fig. 1), tuned to a range of spatial scales and directions. However the model has some unique aspects which distinguish it from other approaches to motion detection using spatiotemporal filters [21], [22] and we briefly list those here. The spatiotemporal filters (step 2) come in two main classes (sustained and transient) which differ in their temporal frequency tuning. The sustained type of filters have low-pass temporal frequency tuning and respond best to static features although they still generate some output in response to moving features. The other class of filters (transient) have band-pass temporal frequency tuning and respond best to moving stimuli. These filters are convolved with the image sequence and the spatiotemporal 'energy' (step 3) is calculated [21]. It is the combination of the energy outputs from these two classes of filters that makes the Perrone [6] algorithm unique; this combination (step 4) is referred to as the Weighted Intersection Mechanism (WIM) and it generates tight temporal frequency tuning (speed tuning) in the motion sensors that follow [23].

Another unique aspect of the model is the combination of the output from a small number of WIM sensors tuned to a range of speeds and directions to produce sensors that are capable of determining the overall direction of a moving pattern rather than the direction of its constituent edges ('the aperture problem' [24], [25]). These pattern detectors (step 5) are based on the properties of Middle Temporal (MT) neurones in the primate brain which are known to respond best to the overall direction of a moving pattern rather than the edge components [26], [27]. The model also includes motion sensors that respond primarily to the direction of moving edges (component units, step 6) and these are used for generating contrast-dependent local spatial inhibition [6]. This stage (steps 7 and 8) carries out a form of redundancy reduction and thins out the velocity vector outputs along and on either side of moving edges in order to increase the signal to noise ratio in the heading estimation stage of the model (see below). The final stage of the velocity estimation code (steps 9 and 10) takes the signals from the pattern units (MT stage) across four spatial scales and, using a weighted vector average scheme (see Ref. [6]), finds the magnitude and direction of the image motion at a particular (x, y) location in the image. These are the vectors we use to determine the heading direction.

B. Heading estimation

The determination of the heading direction (α_H, β_H) from the velocity vectors is carried out using a 'heading template' model [5], [17], [18]. For cases of pure translation of the observer/camera (the case tested here), the location of the centre of expansion of the velocity vectors coincides with the heading direction [1] and the heading templates are designed to locate this point of expansion in the image plane. The principle is illustrated in Fig. 2. This technique is derived from the properties of neurones in the dorsal part of the Medial Superior Temporal area (MSTd) of the primate brain. These cells are known to respond best to the global, full-field patterns of radial expanding image motion that occur as we move through the world [19], [28].

The set of flow vectors estimated from the optical flow algorithm (Sect. II-A) is passed through a set of heading detectors tuned to a range of candidate heading directions (ranging from -60° to 60° in 5° steps) for both azimuth and elevation (α_H, β_H). The tuning of each heading detector can be represented as a 3-D vector in a coordinate space with x aligned with the direction of the camera, y left-right and z up-down. The final heading direction is found from the vector sum of the heading vectors that have activity that is greater than 95% of the most active unit's activity. This tends to produce more accurate estimates although a winner-takes-all

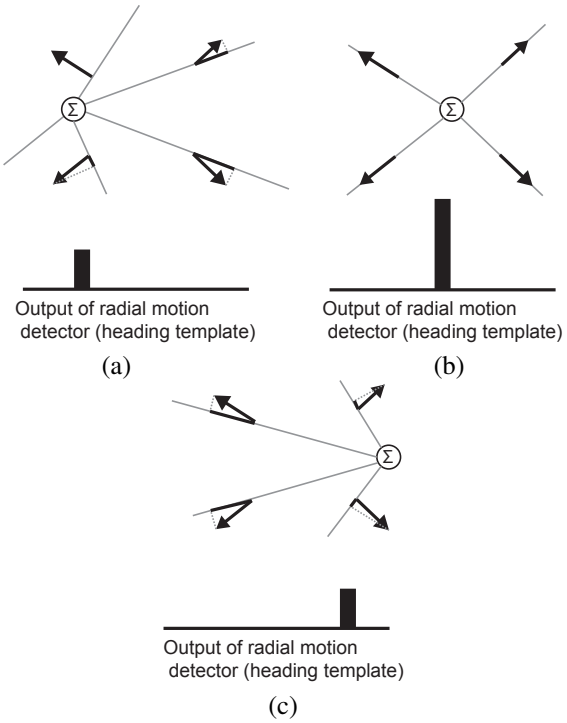


Fig. 2. Illustration of how model heading estimation detectors locate the correct heading direction. (a) Heading unit tuned to an eccentric location (to left) that does not coincide with the point of expansion of the image motion vectors. Summation of the projection (solid lines) of each vector (arrows) onto the radial direction (grey lines) out from the detector location produces a low output in this detector. The sum of the dot products is represented as the solid bar below the detector. (b) A heading detector that is tuned to the correct heading direction. The output is high. (c) Heading detector tuned to a direction eccentric to the true heading. Again the sum of the dot products is low compared to that shown in (b).

or max scheme can also be used [5]. The heading templates integrate motion information over the whole image and so the technique tends to be robust to noise in the optical flow vectors [5].

C. Implementation details

The model has been implemented as a collection of Matlab scripts that can be used to process a 1920×1080 pixel video sequence, with parallelism provided by taking advantage of the GNU parallel tool [29]. The model is currently designed to run on $256 \times 256 \times 8$ frame sequences and so the full movie was subdivided into 32 sub-movies with 16 pixel overlap. These were run individually and the vector outputs from the separate sub-movies were then stitched together to give velocity measurements over the full 1920×1080 image (excluding a 16 pixel region around the outside edges). The full frame vector flow field was then passed through the heading detector stage.

All computations involved in the experiments in this paper were carried out using the central processing unit (CPU) of a machine with 16 physical cores. Fast Fourier transforms were used to accelerate the convolution operations, which made up a significant fraction of the runtime.

Although the current implementation is several orders of magnitude from operating in real-time, recent advances in graphical processing unit (GPU) hardware and software indicates that the convolutions could be greatly accelerated through the use of CUDA libraries for training convolutional neural networks (CNN) [30]. Due to the very large overlap in the type of computation involved in performing a forward propagation in a CNN and processing a video sequence with our model, it is not unreasonable to expect a speedup of two orders of magnitude if the model discussed herein were to be executed on a GPU instead of a CPU.

D. Visual Odometry Feature Tracking

For a baseline comparison a standard VO feature tracking algorithm was also implemented with the Matlab Computer Vision Toolbox and run on the video sequences. The Harris-Stephens [31] corner detector was run on each of the eight frames in a video sequence, and Fast Retina Keypoint (FREAK) [32] descriptors were calculated at each detected corner. The minimisation of the sum of squared differences of the extracted FREAK descriptors was used to match a corner point from one frame to the next frame. The matched corner points, and their shift in pixels from one frame to the next, provided optical flow vectors at the locations of the detected corners.

These were used by the M-estimator sample consensus algorithm [33] (a variant of the random sample consensus algorithm [34]) to find the matrix describing the camera motion between consecutive frames. For the camera translating through the scene only the scaling and translation parameters of the determined similarity transform are relevant. From these the centre point of expansion of the camera was determined and used to calculate the azimuthal component of heading. This was repeated for each of the seven pairs of frames and the mean of the calculated azimuthal heading for the eight-frame sequence is reported in the results below.

III. Testing Methodology

A. Input sequences

A GoPro Hero3+ Black camera (operated at 1920×1080 pixel resolution, $30\,\mathrm{fps}$) with field of view $64.4°$ horizontal by $37.2°$ vertical using default image processing was mounted on a linear translation table (Macron Dynamics, Croydon, PA). The camera was positioned one foot above the table and translated at velocities of 0.2, 0.4, 0.8, 1.2, 1.6 and $2.0\,\mathrm{m/s}$, with the heading angle relative to the motion of the translation system varied from 0 to $50°$ in $10°$ increments for each speed. Each translation was to $2\,\mathrm{m}$ forward of the camera starting position and all sequences began at the same position. Recordings took place within a static environment under constant (fluorescent) lighting. For each video sequence, eight consecutive frames, the first taken when the camera was $1\,\mathrm{m}$ from the starting position, were extracted as uncompressed portable network graphic (PNG) images. These frames were then used as input for the model.

It was difficult to align the camera view direction with the heading direction during setup of the first trial. As such, true heading was manually obtained from visual analysis of each of the zero degree heading videos and used to compensate for the constant offset created by initial camera placement. The camera lens calibration was determined for the GoPro and all acquired images were rectified to the true perspective projection before analysis.

The scene can be seen in Fig. 3 (with extra annotations) and consists of regular structured objects (the grid patterns), box and line like objects (whose edges are subject to the aperture problem), a substantial region of dark draping exhibiting subtle texturing, and some background clutter.

IV. Results

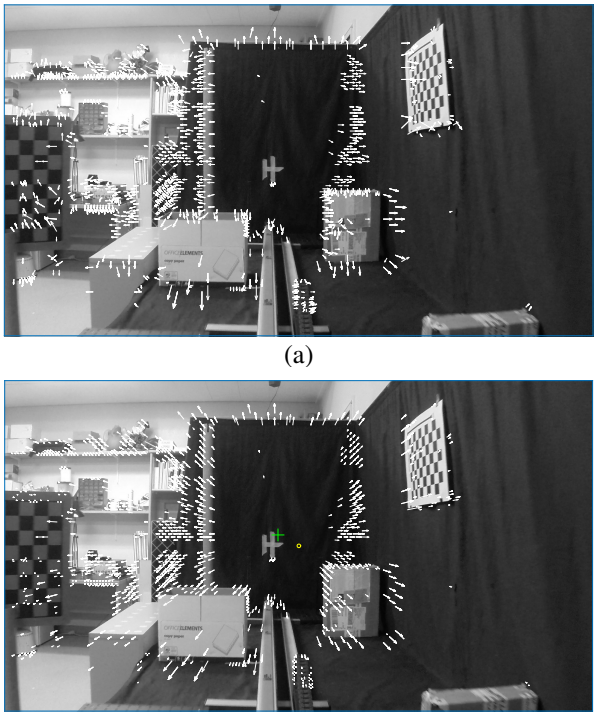
A. Model estimates and graphs

The flow vectors detected by the biological sensor are presented superimposed upon the an image of the scene for the case of the camera orientated to point forward along the linear translation table (the 0° case) in Fig. 3. The aperture problem is very evident with flow vectors pointing perpendicular to the edges (Fig. 3(a)). However, when the vectors are projected on to the radial direction from the estimated heading direction (Fig. 3(b)), the image motion is now much more consistent with what would be expected from forward translation. All determined headings in the results below are reported relative to the centre of the image which is marked by the yellow circle in Fig. 3(b), however the linear translation table is not quite orientated to point to the centre of the images, and the true centre of motion for the 0° case is 3.3° left of the centre of the image.

The output of the heading detectors in the model is shown in Fig. 4 as a surface plot (a) and as a contour plot (b). The vertical axis for the surface plot (out of page for contour plot) is the normalised output of each heading template with the azimuth and elevation tuning of the templates represented on the x and y-axes. This is for a $(-3.3°, 2.5°)$, $0.8\,\mathrm{m/s}$ test case when the output of the model was $(-3.5°, 2.5°)$.

Experiments were repeated for the camera rotated at various orientations to the translation table and for various speeds of translation along the table. The results for various speeds plotted against the azimuthal heading of the camera is plotted in Fig. 5. For angles of 30° and smaller the rotation angle of the camera relative to the motion is well estimated.

The results are also plotted for various headings against velocity in Fig. 6. Tests were performed with the camera rotated by $0\,\mathrm{deg}$ at increments of 10° up to 50° from the direction of travel (the azimuthal heading), however there was also a $-3.3°$ misalignment of the translation table to the centre of the images, thus a coloured dashed line with the misalignment incorporated is plotted for each heading on the graph to indicate the true heading value. The solid lines are the measured results for each heading. It can be seen that there is relatively good stability with respect to changing linear speed of the camera except for those with headings of 40° and above.



(a)

(b)

Fig. 3. The flow vectors determined from the motion tuned filters for the camera travelling at $0.8\,\mathrm{m/s}$ forward along the linear translation stage (visible in centre-bottom of image) and orientated at 0° to the translation stage, showing (a) the raw motion vectors and (b) the motion vectors projected on to the flow determined from the estimated heading. The yellow circle marks the centre of the image and the green cross the centre of the determined flow.

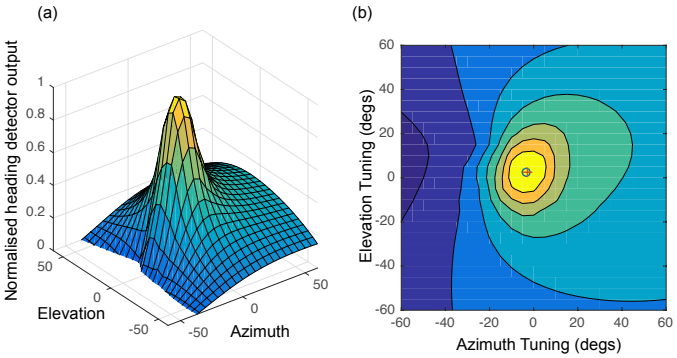


Fig. 4. Heading map shown in 3-D form (a) and as a contour map (b). The true heading is represented in (b) as a circle and the superimposed estimate of heading from the model is shown as a cross.

B. Comparison with computer vision methods

The 8-frame video sequences were also analysed with the more traditional feature tracking approach to estimate the heading (as described in Sect. II-D). The results are plotted for various headings against speed in Fig. 7. As can be seen the heading vectors particularly for heading angles below 20° are biased below the true heading value. As was the case for the biological sensor, the estimations for the large heading angles are poorer in quality and subject to significant errors.

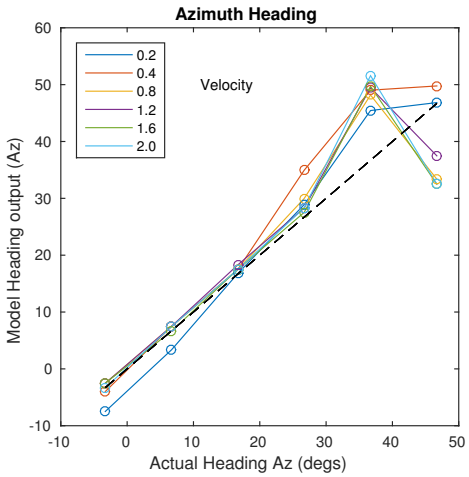


Fig. 5. The estimated azimuth heading estimation plotted against the actual azimuth heading. Each curve is for a certain velocity as indicated in the legend (velocity units are m/s).

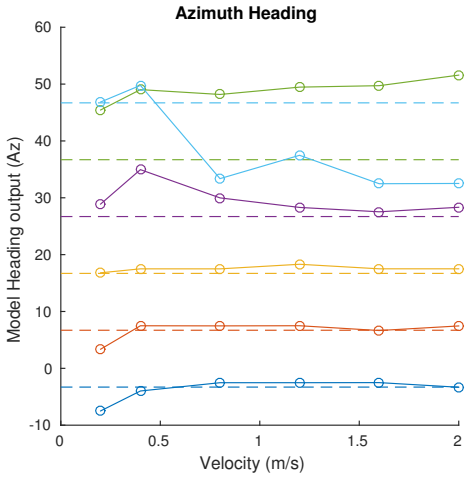


Fig. 6. The dependence of the estimated azimuth heading for various translational speeds. The dashed lines show the true azimuth heading for each coloured curve.

V. Discussion

We have shown that heading direction can be estimated from a brief (8 frame) monocular video sequence using motion sensors and heading detectors that are based on the properties of cells in the primate visual system [5], [6]. The model performed well over a wide range of heading angles out to approximately 30° eccentricity. Errors crept in at the higher test angles (40° and 50°) but these can be attributed to the fact that these sequences contained many regions with image velocities that exceeded the maximum tuning of the motion sensors. The motion sensors required to detect very high image speeds consist of filters which occupy most of the 256×256 image size we use for subdividing the full 1920×1080 video frames. These can suffer from edge effects that would not occur if we carried out our motion analysis over larger areas (e.g., 512×512). Therefore we consider the deterioration of the

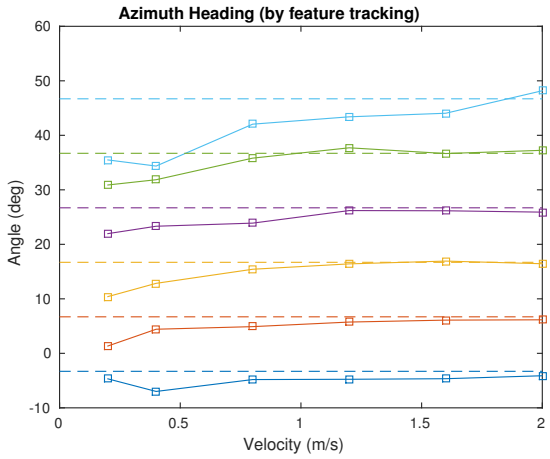


Fig. 7. Result of the VO feature tracking analysis. The line colours are the same as Fig. 6.

model's performance that occurred at higher eccentricities to be an artefact of our current implementation and not something inherently wrong with the way the model works. It should be pointed out that humans also make errors in heading estimation at large eccentricities [2] and tend to minimise the generation of high image velocities by tracking points in the scene with the eyes.

When the true heading direction was within the 0° to 30° range the model was accurate over a wide range of camera velocities (Fig. 6). Again, the errors increased for the case where the majority of the image motion was very low in magnitude and hence outside of the speed tuning range of the motion sensors. Such conditions tend to produce very sparse flow fields which affect the accuracy of the heading estimation stage. This is an intentional design feature and is designed to increase the signal-to-noise ratio of the heading detectors [6]. It should be noted that an advantage compared to feature based VO algorithms [8], [12] is that the motion sensors in our model produce flow in high-speed areas of the image, along extended edges and even in regions without distinctive features to detect.

For all heading eccentricities except 40° and 50°, the RMS error for the model compares favourably to the technique based on VO feature tracking (2.31° vs 2.58°). The biologically-based model also has the advantage of not requiring iterative search techniques to arrive at a solution. Given the appropriate hardware, a solution to the heading problem could in theory be generated within a 233 ms window (the time course of the temporal filters in the motion sensors) using our approach. This has obvious advantages in applications such as the control of fast moving micro-aerial vessels (UAVs).

The task of determining heading in the case of pure translation is relatively easy compared to the harder problem of determining heading in the presence of camera rotation. However the extraction of the flow field from video sequences is far from trivial and it is this aspect that we wanted to emphasis in this paper; it is a proof of principle that the flow can be derived from biologically-based motion sensors and that these velocity

estimates can be used to derive information regarding the camera's movement. We have already demonstrated that the impact of (known) camera rotation can be subtracted from the heading detector array activity to extract the pure translation heading signal [35] and this technique will be used in the future to test cases in which rotation of the camera is occurring during the translation.

Once heading has been determined, the derivation of the radial flow field (Fig. 3b) enables the relative depth of points in the scene to be estimated; there is a direct transformation from the vector magnitudes to the depth map (scaled by a factor dependent on the unknown camera speed). One of the main goals of this line of research is to obtain 3-D depth estimates from the monocular 2-D video input, something that humans can do with very brief stimulus exposures. Slight movements of the head reveal the depth structure of the world around us, even with one eye closed. We believe that the tests we have carried out in this paper are a first step in being able to emulate this amazing ability in software.

VI. Conclusion

We have demonstrated the extraction of heading information from monocular video sequences using a realistic biologically-based model. Over all heading angles and speeds tested our model had an RMS error of 6.97° compared to a conventional visual odometry feature tracking approach which achieved an RMS error of 3.87°. However, the model motion sensors are currently only tuned to deal with a smaller range of heading angles (30° and smaller) and when the analysis is restricted to this range the model performs favourably (RMS error of 2.31°) compared to the VO feature tracking (RMS error of 2.58°). In principle, the model implementation can be extended to analyse larger heading angles and we would expect it to then perform very favourably compared to VO feature tracking at all heading angles.

References

[1] J. Gibson, *The perception of the visual world*. Boston: Houghton Mifflin, 1950.

[2] W. Warren, *Optic flow*. Cambridge, Massachusetts: Bradford, 2003, vol. 2, pp. 1247–1259.

[3] J. J. Koenderink and A. van Doorn, "Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer," *Optica Acta*, vol. 22, no. 9, pp. 773–791, 1975.

[4] H. C. Longuet-Higgins and K. Prazdny, "The interpretation of moving retinal images," *Proceedings of the Royal Society of London B.*, vol. B 208, pp. 385–387, 1980.

[5] J. Perrone, "Model for the computation of self-motion in biological systems," *Journal of the Optical Society of America*, vol. 9, pp. 177–194, 1992.

[6] J. A. Perrone, "A neural-based code for computing image velocity from small sets of middle temporal (MT/V5) neuron inputs," *Journal of Vision*, vol. 12, no. 8, 2012.

[7] D. Regan and K. I. Beverley, "How do we avoid confounding the direction we are looking and the direction we are moving?" *Science*, vol. 215, no. 8, pp. 194–196, 1982.

[8] J. B. D. Nistér, O. Naaroditsky, "Visual odometry," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, vol. 1, Washington, DC, June 2014, pp. 652–659.

[9] D. Scaramuzza and F. Fraundorfer, "Visual odometry: Part I: The first 30 years and fundamentals," *IEEE Robotics and Automation Magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[10] F. Fraundorfer and D. Scaramuzza, "Visual odometry: Part II: Matching, robustness, optimization and applications," *IEEE Robotics and Automation Magazine*, vol. 19, no. 2, pp. 78–90, 2012.

[11] P. Corke, D. Strelow, and S. Singh, "Omnidirectional visual odometry for a planetary rover," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 4, 2004, pp. 4007–4012.

[12] A. Pretto, E. Menegatti, and E. Pagello, "Omnidirectional dense large-scale mapping and navigation based on meaningful triangulation," in *IEEE International Conference on Robotics and Automation*, Shanghai, China, 2011, pp. 3289–3296.

[13] R. Goecke, A. Asthana, N. Petterson, and L. Petersson, "Visual vehicle egomotion estimation using the fourier-mellin transform," in *IEEE Intelligent Vehicles Symposium*, Istanbul, Turkey, 2007, pp. 450–455.

[14] M. J. Milford and G. F. Wyeth, "Single camera vision-only SLAM on a suburban road network," in *IEEE International Conference on Robotics and Automation*, Pasadena, CA, 2008, pp. 3684–3689.

[15] K. Nakayama, "Biological image motion processing: A review," *Vision Research*, vol. 25, no. 5, pp. 625–660, 1985.

[16] C. W. G. Clifford and M. R. Ibbotson, "Fundamental mechanisms of visual motion detection: models, cells and functions," *Progress in Neurobiology*, vol. 68, no. 6, pp. 409–437, 2002.

[17] J. Perrone and L. Stone, "A model of self-motion estimation within primate extrastriate visual cortex," *Vision Research*, vol. 34, pp. 2917–2938, 1994.

[18] ——, "Emulating the visual receptive field properties of MST neurons with a template model of heading estimation," *Journal of Neuroscience*, vol. 18, pp. 5958–5975, 1998.

[19] K. Tanaka, K. Hikosaka, H. Saito, M. Yukie, Y. Fukada, and E. Iwai, "Analysis of local and wide-field movements in the superior temporal visual areas of the macaque monkey," *The Journal of Neuroscience*, vol. 6, no. 1, pp. 134–144, 1986.

[20] D. C. Bradley and M. S. Goyal, "Velocity computation in the primate visual system," *Nature Reviews Neuroscience*, vol. 9, no. 9, p. 686(10), 2008.

[21] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *J. Opt. Soc. Am. A*, vol. 2, no. 2, pp. 284–299, 1985.

[22] E. Simoncelli and D. Heeger, "A model of the neuronal responses in visual area MT," *Vision Research*, vol. 38, pp. 743–761, 1998.

[23] J. Perrone and A. Thiele, "A model of speed tuning in MT neurons," *Vision Research*, vol. 42, pp. 1035–1051, 2002.

[24] E. C. Hildreth, *The neural computation of the velocity field*. New York: Raven Press Ltd., 1990, pp. 139–164.

[25] S. Wuerger, R. Shapley, and N. Rubin, ""on the visually perceived direction of motion" by hans wallach: 60 years later," *Perception*, vol. 25, no. 11, pp. 1317–1367, 1996.

[26] J. A. Movshon, E. H. Adelson, M. S. Gizzi, and W. T. Newsome, *The analysis of moving visual patterns*. Ex Aedibvs Acaddemicis, Civitate Vaticana, 1983, vol. 54, pp. 117–151.

[27] R. T. Born and D. Bradley, "Structure and function of visual area MT," *Annual Review of Neuroscience*, vol. 28, pp. 157–189, 2005.

[28] K. H. Britten and R. J. van Wezel, "Electrical microstimulation of cortical area MST biases heading perception in monkeys," *Nat Neurosci*, vol. 1, no. 1, pp. 59–63, 1998.

[29] O. Tange, "GNU Parallel - The Command-Line Power Tool," *;login: The USENIX Magazine*, vol. 36, no. 1, pp. 42–47, Feb 2011. [Online]. Available: http://www.gnu.org/s/parallel

[30] S. Chetlur, C. Woolley, P. Vandermersch, J. Cohen, J. Tran, B. Catanzaro, and E. Shelhamer, "cuDNN: Efficient primitives for deep learning," *arXiv preprint arXiv:1410.0759*, 2014.

[31] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, August 1988, pp. 157–151.

[32] A. Alahi, R. Ortiz, and P. Candergheynst, "FREAK: Fast retina keypoint," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, June 2012, pp. 510–517.

[33] P. H. S. Torr and A. Zisserman, "MLESAC: a new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, pp. 138–156, 2000.

[34] M. A. Fischler and R. C. Bolles, "Random sample consensus: Paradigm for model fitting with applications to image analysis and automated cartography," *Comunnications of the ACM*, vol. 24, pp. 381–395, 1981.

[35] J. A. Perrone and R. Krauzlis, "Vector subtraction using visual and extraretinal motion signals: A new look at efference copy and corollary discharge theories," *Journal of Vision*, vol. 8, no. 14, pp. 1–14, 2008.