MVA '98 IAPR Workshop on Machine Vision Applications, Nov. 17-19, 1998, Makuhari, Chiba, Japan

2—5 Autonomous Robot Navigation by Active Visual Motion Analysis and Understanding

Sung Shic Park* and Arcot Sowmya†
Department of Artificial Intelligence, School of Computer Science and Engineering
University of New South Wales
Kensington NSW 2052 Australia

Abstract

Over the past decade, there have been numerous attempts to achieve intelligent levels of autonomy for mobile robots. Many such systems however, have intensive programmer-encoded knowledge and fail to operate in unknown environments. This leads to intelligence ought to be based on behavioural capabilities inspired by biological systems. Using this approach, a real-time active visual motion understanding system has been developed to detect obstacles in a mobile robot's environment, without requiring *a priori* knowledge about the world. Robot navigation incorporating it has shown encouraging results.

1 Introduction

In intelligent robot development, tasks that are trivial for humans, such as obstacle avoidance, become non-trivial. This is because a conflict exists between real-world tasks and the usual numerical problems that computers were originally designed to solve. This also explains why classical vision methods such as that of Marr (1982), which suggest an explicit internal modelling of a world, have failed in dynamically changing real-world environments. Interestingly, biological studies suggest a natural way of looking at robot competence as a complex behaviour constructed out of many trivial behaviours, each of which is simple to model and implement. Such biologically inspired models often provide more robust robot control than traditional models.

Our goal is to develop a biologically inspired autonomous robot control system in which the internal programmer-encoded knowledge of a world can be omitted. Further this system will contain an active vision module, because active vision not only conforms to biological systems but also promises to simplify and facilitate the development of higher level competences such as navigation. As will be discussed later, the combination of vision and behaviour (ie active vision) becomes necessary in achieving robust real-time perception for a robot system which interacts with a dynamically changing environment.

In this research, a low-level visual reaction system, whereby an independent agent can detect colliding threats without *a priori* knowledge, is achieved by developing a novel *real-time optical flow* algorithm, then *clustering* retinal pixels *to segment objects* in 2-D space, and finally computing the *looming aspect* of objects in the scene to detect possible collision. A behavioural control systems uses the looming aspect to navigate in the world.

In Section 2, a modified optical flow algorithm is presented, followed by description of the clustering step in Section 3. The looming aspect computation is described in Section 4, and the actual navigation control algorithm in Section 5. The last section presents concluding remarks.

2 Real-time Optical Flow Estimation

Optical flow is a popular technique in dynamic scene analysis. However, establishing correspondence in a discrete sequence of consecutive images involving displaced objects is a cause of computational inefficiency in optical flow calculation. Spatial neighbourhood correlation (Aggarwal, Davis & Martin 1981), which measures the degree of correspondence between two points, is computationally more efficient than other methods, including gradient-based approaches (Horn 1986). However, the algorithm is still far from providing real-time performance, which is necessary for developing mobile robot systems.

The problem with conventional correlation is that it tries to correlate totally irrelevant pairs of points and, in fact, spends more time looking at irrelevant pairs than relevant ones. Our new approach to establishing correspondence relies upon an *a priori* relevance check between a candidate pair of points, and spatial correlation is computed only for relevant pairs in the two frames. Relevance is measured by comparing *photo-signatures* of the two points; photo-signature of a pixel is a spatially invariant identification based on its spatio-temporal intensity gradients.

Let $E(x, y, t)$ be the brightness at time t at an image point (x, y), and let the corresponding point at time $t + dt$ be $(x + dx, y + dy)$; the x and y components of the optical flow vector at point (x, y) are $u(x, y)$, and $v(x, y)$ respec-

*E-mail: unixor@chollian.net
†Email: sowmya@cse.unsw.edu.au

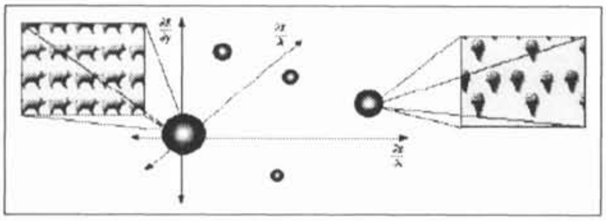


Figure 1: An artist's impression of clusters of relevant pixels in spatio-temporal gradient space.

tively, and $dx = u \cdot dt$ and $dy = v \cdot dt$;

In the optimal case, without considering any sensor noise or discretization, the two points will have identical brightness:

$$E(x + udt, y + vdt, t + dt) = E(x, y, t) \qquad (1)$$

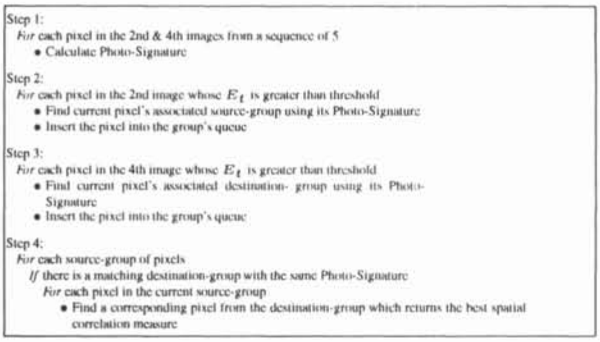
The optical flow constraint equation (Horn 1986) as in gradient based algorithms can be derived from Equation 1 and we get

$$E_x u + E_y v + E_t = 0 \qquad (2)$$

where $u = \frac{dx}{dt}$, $v = \frac{dy}{dt}$, $E_x = \frac{\partial E}{\partial x}$, $E_y = \frac{\partial E}{\partial y}$ and $E_t = \frac{\partial E}{\partial t}$. Equation 2 constrains the derivatives E_x, E_y and E_t to remain constant with respect to the displacement vector (u, v). The combination of these two spatial gradients (E_x and E_y) and a temporal gradient (E_t) can be used as a spatially invariant identification for the point (x, y) and is called the Photo-Signature of the pixel. Figure 1 illustrates clusters of closely related pixels in terms of their corresponding Photo-Signature values. The frame with ice-cream cones is a dramatised plot of pixels from the rightmost cluster, where each ice-cream cone can be considered as a pixel. All the ice-cream cones have similar spatial intensity gradients (ie E_x & E_y), and are moving at a similar temporal speed (E_t). The same is true for the frame with pigs with its associated cluster. This illustrates the rationale behind using Photo-Signature to measure relevance between pairs of points in two frames, whereby correlating pigs to ice-cream cones is pointless.

This new method overcomes the problems associated with the gradient-based and correlation-based methods. The former, which utilises the constraint equation (See Equation 2) suffers from the aperture problem (Nakayama and Silverman 1988, Davies 1997), which can be reduced to a certain extent by employing neighbourhood correlation. The latter lacks the means to predetermine the relevance of a given pair of points, which can be eliminated by using the Photo-Signature to select candidate pairs for spatial correlation. An example of an estimated optical flow map and its associated images is shown in Figure 2. The pseudo-code for this algorithm is given in Figure 3. Given a sequence of 5 consecutive images, an optical flow image can be estimated for the 2nd least recent image from the sequence.

Figure 2: Optical flow map.



```
Step 1:
  For each pixel in the 2nd & 4th images from a sequence of 5
    • Calculate Photo-Signature
Step 2:
  For each pixel in the 2nd image whose $E_t$ is greater than threshold
    • Find current pixel's associated source-group using its Photo-Signature
    • Insert the pixel into the group's queue
Step 3:
  For each pixel in the 4th image whose $E_t$ is greater than threshold
    • Find current pixel's associated destination- group using its Photo-
      Signature
    • Insert the pixel into the group's queue
Step 4:
  For each source-group of pixels
    If there is a matching destination-group with the same Photo-Signature
      For each pixel in the current source-group
        • Find a corresponding pixel from the destination-group which returns the best spatial
          correlation measure
```

Figure 3: Pseudocode for estimating optical flow.

3 Unsupervised Image Segmentation

Optical flow vectors are based on pairs of points and do not convey information about object size. Clusters of spatially spread optical flow vectors may be used to segment the image and estimate object size. There are three features available from optical flow as input to this segmentation process: the magnitude of an optical flow vector, and the x and y coordinates where the vector is defined. Spatial clusters based on these features can help detect groups of pixels close to each other and moving at similar speed. In this work, Kohonen's (1982) Self-Organising Feature Map (SOM) was used to cluster optical flow vectors (See Figure 4).

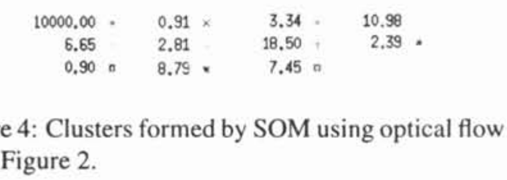


Figure 4: Clusters formed by SOM using optical flow data from Figure 2.

4 Collision Detection by Visual Looming

For obstacle detection, the focus of expansion (FOE) may be utilised for the construction of a scene depth map (Davies 1997). The FOE however, is usually not known *a priori*, its calculation is non-trivial, and subject to change with different camera gaze directions and errors in the optical flow field. Moreover, this method only allows the detection of static obstacles, assuming there is no self-moving object in the field of view, though this is likely in the real world.

The looming characteristics of perceived objects provide strong and natural clues for the detection of obstacles. An equation for computing time-to-contact (TTC) with visual looming was initially noted by the astronomer Hoyle (1957), who pointed out that

$$T \approx \frac{\theta}{d\theta/dt} \tag{3}$$

where T is the TTC derived from an angular diameter θ and $\mathrm{d}\theta/dt$ the temporal derivative or the rate-of-expansion (ROE) of a rigid spherical object moving at a constant speed along the line of sight. Various studies show that humans use a similar method to help guide goal-directed discriminative motor action, for example in sport, highway-driving and aviation (Beverley and Regan 1979, Schiff and Detwiler 1979, Todd 1981, Kruk and Regan 1983).

The problem is to find the size of an object (ie θ) without *a priori* knowledge about that object. Here, we utilize the spatial clusters based on optical flow features mentioned earlier. In segmenting an object, humans seem to exploit exploit primitive features which apply generically to arbitrary objects of any shape, and consequently our approach is justifiable. To estimate object size from clusters of pixels, an ellipse is fitted to each cluster, and either of the two axes lengths is used as an estimate of the corresponding object size θ.

Each cluster of optical flow vectors is associated with two object clusters, the source point cluster and destination point cluster (see Figure **??**). Thus two successive θ values returned over time may be used to calculate the ROE of the cluster, thereby enabling TTC estimation, which is used to obtain an obstacle image such as in Figure 6. Values listed with various marks at the bottom of Figure 4 represent uncalibrated TTCs for clusters having identical marks. The value for the cluster labelled with a darker X mark is reasonably small compared to others, and this is consistent with the actual distance to the related object (see the chair to the right in Figure 2).

Figure 6 shows an obstacle image where each pixel value is the thresholded TTC for its corresponding cluster. The brightness of each pixel in the image tells the robot how close it is to collision with the corresponding object; the brighter the intensity the more imminent it is to collision. As is seen in the figure, the obstacle detection is reasonably robust, despite any clustering errors.

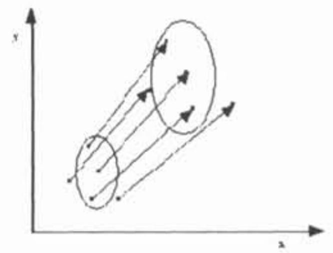


Figure 5: Two temporally different *eta* values determined by a set of clustered optical flow vectors.

Figure 6: Imminent obstacle image.

5 Active Visual Behaviours for Collision Avoidance

The active vision-based robot navigation system developed for this project accommodates a behavioural stimulus-reaction mechanism that combines Brooks' subsumption architecture (1986) with Nelson's behavioural vision system (1991), and is illustrated in Figure 7. Lower layers provide information to successively higher layers, and suppression and inhibition of control between layers take place, resulting in robust robot control.

The two layers for detecting motion and obstacles use the vision algorithms discussed earlier, based on Photo-Signature and visual looming. The Obstacle Detection layer, in conjunction with the Motion Detection layer, supplies critical visual information regarding imminent collision to other control layers. Because the optical flow estimation algorithm based on Photo-Signature requires a steady sequence of images, shafter encoder values, from which the robot can estimate the distance it has travelled, are used to synchronise the frame-grabbing frequency with a constant interval of robot movement; ie ego-motion estimation and image synchronization. The Gaze Control layer drives the alteration of camera parameters, and may be expanded to three active visual behaviours shown below:

Behaviour-I	Giving attention to peripheral threats
Behaviour-II	Tracking the fovea
Behaviour-III	Directing the gaze toward a free path

Behaviours I & III are extensions of gaze change, while behaviour II is extended from gaze stabilisation (Brown

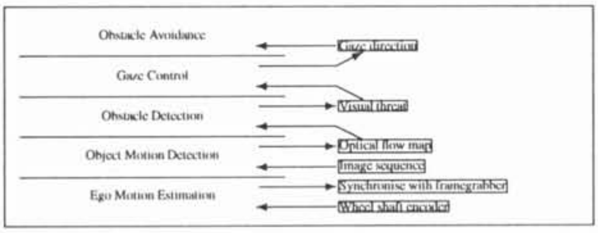


Figure 7: Hierarchical view of the system.

1990a, 1990b). The layers up to this level embody the active vision aspects of the system, and the core behaviour for navigation is left to the top layer. In Behaviour I, peripheral visual threats are used to choose the system's next attention, while Behaviour III can override Behaviour I whenever a higher level decision such as avoiding an imminent obstacle is made by the system to look for a freeway. Meanwhile, Behaviour II helps to stabilise the camera gaze so that the set of grabbed images is not affected by the inevitable rotational effects of robot movement.

The top layer simply accepts a signal triggered by Behaviour III, which specifies a change in gaze direction, and whenever the new gaze direction is mis-aligned with the forward body direction of the robot, the body is rotated to the side where the gaze is currently fixated (ie left/right with respect to the forward body direction).

6 Conclusion

This paper presented a behavioural approach to building robot intelligence, incorporating active visual image analysis. The novel motion understanding technique introduced, which uses Photo-Signature based optical flow estimation and visual looming, achieved real-time An active camera control mechanism facilitating this visual proximity sensing has been developed.

A robot with this control system responds to nearby obstacles by turning its head to view them more closely through the use of foveal vision, and if an imminent collision is detected by doing so, the robot will look for free space. As a consequence, the body of the robot chases the head in an attempt to align the forward body direction to the gaze direction. This aligning behaviour eventually makes the robot to stand facing free space.

The mobile robot in our robotics laboratory, running this control system, could navigate in an unconstrained lab environment for more than 10 minutes without any collision. Future extensions include goal-directed navigation and chasing behaviour.

7 Acknowledgments

The authors thank Mark Peters, Department of Artificial Intelligence, for showing the way to active vision and many insightful discussions on its transfer to robotics.

References

[1] J. K. Aggarwal, L. S. Davis, W. N. Martin (1981), "Correspondence processes in dynamic scene analysis", Proceedings of the IEEE, 69: 562-572.

[2] K. J. Beverley & D. Regan (1979), "Separable aftereffects of changing size and motion-in-depth: Different neural mechanisms?", Vision Research, 19, 727-732.

[3] R. A. Brooks (1986), "A Robust Layered Control System For A Mobile Robot", IEEE Journ. of Robotics and Automation, vol. RA-2, no. 1, March.

[4] C. M. Brown (1990a), "Gaze controls with interactions and delays", IEEE Transactions of System, Man and Cybernetics, 20:3, May.

[5] C. M. Brown (1990b), "Prediction and cooperation in gaze control", Biological Cybernetics, 63: 61-70, May.

[6] E. R, Davies (1997), "Machine Vision", Academic Press, San Diego, CA, pp. 431-453.

[7] B. K. P. Horn (1986), "Robot Vision", The MIT Press, Cambridge, MA.

[8] F. Hoyle (1957), "The black cloud", Penguin, London, pp. 26-27.

[9] K. Joarder & D. Raviv (1992), "Autonomous Obstacle Avoidance Using Visual Fixation and Looming", SPIE: Intelligent Robots and Computer Vision XI, vol. 1825, 733-744.

[10] T. Kohonen (1982), "Self-Organized Formation of Topologically Correct Feature Maps", Biological Cybern. 43, 59-69.

[11] R. Kruk & D. Regan (1983), "Visual test results compared with flying performance in telemetry-tracked aircraft", Aviation, Space and Environ. Medicine, 54, 906-911.

[12] D. Marr (1982), "Vision: a computational investigation into the human representation and processing of visual information", W H Freeman, New York.

[13] K. Nakayama & G. Silverman (1988), "The Aperture Problem-I", Vision Research, 28(6):739-746.

[14] R. C. Nelson (1991), "Introduction: Vision as Intelligent Behavior - An Introduction to Machine Vision at the University of Rochester", Intl. Journ. of Comp. Vision, 7:1, 5-9.

[15] D. Regan & A. Vincent (1995), "Visual Processing of Looming and Time to Contact Throughout the Visual Field", Vision Research, vol. 35, no. 13, 1845-1857.

[16] W. Schiff & M. L. Detwiler (1979), "Information judged in impending collision", Perception, 8, 647-658.

[17] M. J. Swain & M. A. Stricker *et al.* (1993), "Promising Directions in Active Vision", Interl. Journ. of Comp. Vision, 11:2, 109-126.

[18] J. T. Todd (1981), "Visual information about moving objects", Journ. of Exper. Psych., 7, 795-810.