A Neural Network for Ego-motion Estimation from Optical Flow

A. Branca, G. Convertino, E.Stella, A. Distante
Istituto Elaborazione Segnali ed Immagini - C.N.R.
Via Amendola 166/5, 70126 Bari
ITALY
{branca,convertino,stella,distante}@iesi.ba.cnr.it

Abstract

In this work we consider the problem to determine qualitative information about the motion of a viewer moving in a stationary environment. First the optical flow (OF) is computed using a token based approach estimating the 2D velocity vectors only for some interesting points. Then our method estimates the motion of the viewer using only the available sparse OF.
A neural network extracts information about stable points useful for the computation of vehicle's heading and Time-to-Collision (TTC). A number of experiments showing the efficacy and robustness of the method have been performed both on synthetic image sequences and on real images acquired by a CCD camera mounted on a mobile platform.

1 Introduction

One of the most interesting applications of computer vision is the reconstruction of three-dimensional properties of a scene. from two-dimensional images. When a CCD camera is used as a sensor for autonomous vehicles, an important goal is to recover, from time-varying images, the relative motion between the viewer and the scene in order to avoid collisions or to perform on-line adjustments of the current navigational path. Psychophysical evidence exists [3] that two main ego-motion parameters, namely the *heading direction* and the *time-to-collision* (TTC), allow living organism to perform the above tasks.
A fundamental preliminary step for the extraction of 3D information from TV images is the computation of the 2D motion field from the variation of the 2D brightness patterns along the image sequence. It is called *Optical Flow* and it is qualitative similar to the 2D theoretic motion field, in the sense that it has the same topological structure.
In the present work we consider the application context of planar passive navigation in which the relative motion between the viewing camera and the scene is mainly a translation on a flat surface, with rotations occurring only around an axis orthogonal to both the surface and the translational component of motion. The resulting O.F. has a radial topology, with a singular point (the point where the flow vanishes) called *Focus of Expansion* (FOE), lied in the direction toward the

BMVC 1995 doi:10.5244/C.9.25

camera is moving. Moreover the TTC can be computed from the O.F. considering a small neighborhood of the FOE.
The FOE position is independent of the distances from the world surfaces and no assumptions about surface shape and smoothness are required.
Actually, the accurate computation of FOE seems to be a hard problem, mainly due to digitization errors that produce unreliable flow vectors. Moreover, it is impossible to produce a perfect radial flow due to small amounts of observer rotations or accidental camera vibrations that can change the radial shape of the flow field and the true position of the FOE. To manage this situation several methods have been proposed. They require a preliminary decomposition of the optical flow field into its translational and rotational components [6] or the computation of a 2D region of possible FOE locations (Fuzzy FOE) [2] instead of a single FOE.
In this work we propose a new contribution for the computation of autonomous vehicles motion. It is based on a neural module which detects the FOE associated with the O.F. field in order to recover vehicle's heading and TTC.
Since a dense flow field is generally required for motion segmentation tasks or to compute all motion parameters, a sparse flow field is sufficient in our context in which the visual control of locomotion requires only the heading direction.
We consider a sparse O.F. obtained by matching (with the Hopfield-type neural network proposed in [1]) image features extracted through Moravec's interest operator [5].
Based on the detected sparse optical flow, the proposed neural network locates the focus of expansion useful for heading direction detection and TTC estimate. Our neural network implements a gradient descent technique to compute optimal coefficient values, by means of which the three velocity fields (named basis flow fields), representing respectively the three elementary translational motions, can be combined to give a vector field with minimum distance from the analyzing OF. The FOE position is recovered from the so computed coefficients.
In the following sections the algorithm used to estimate *heading direction* and *time-to-collision* (section 2 and 3), with the architecture and the dynamic of the proposed neural network (section 4), are described. The most relevant experimental results will follow (section 5).

2 Three-Dimensional Interpretation of Visual Motion

The approach we propose, in order to recover the heading direction of a vehicle moving in a stationary world, attempts to compute 3D translational motion parameters from the 2D optical flow obtained by projecting the 3D velocities on the image plane. The OF is estimated only for sparse points with the method proposed in [1]. We assume a perspective projection model in which a world point $P = (X, Y, Z)$ projects on the image point $(x, y) = f\left(\frac{X}{Z}, \frac{Y}{Z}\right)$, where f is the focal length. Longuet-Higgins and Prazdny [4] derived the following equations to

describe the general rigid motion of an observer moving in a stationary world:

$$u = \frac{T_x - xT_z}{Z(x,y)} - xyR_x + (1+x^2)R_y - yR_z \tag{1}$$

$$v = \frac{T_y - yT_z}{Z(x,y)} - (1+y^2)R_x + xyR_y + xR_z \tag{2}$$

with (T_x, T_y, T_z) the 3D translational velocity components, (R_x, R_y, R_z) the 3D rotational velocity components, (u, v) the projected velocity of a point (X, Y, Z) on the image plane, $Z(x, y)$ the depth function.

Though various algebraic approaches have been proposed to resolve non-linear systems resulting from writing equations (1) and (2) in a suitable number of image points, the results are numerically instable due to large number of equations to be solved and the noise in the (u, v) velocity vector estimates.

We consider the application context in which the viewer translates on a flat ground and can rotate only around an axis orthogonal to the ground (*passive navigation*). The resulting 2D motion field has a radial topology: on the image plane all 2D velocity vectors radiate from a singular point (that is the point where the flow vanishes) named focus of expansion (FOE). The FOE is the projection on the image plane of the direction along which the observer moves. A rotation, occurring while the observer translates, will cause a FOE location shifting by preserving always the radial shape.

The FOE location can be correctly estimated as the point where the translational motion V_t vanishes:

$$V_t = \left(\frac{(T_x - xT_z)}{Z(x,y)}, \frac{(T_y - yT_z)}{Z(x,y)} \right) \tag{3}$$

$$(FOE_x, FOE_y) = f\left(\frac{T_x}{T_z}, \frac{T_y}{T_z} \right) \tag{4}$$

If $T_z = 0$ the FOE can be defined as a direction toward infinity defined by (T_x, T_y) and all the flow vectors point in that direction.

Moreover, it can be shown that the time to contact defined as $TTC = -\frac{Z}{T_z}$ can be easily computed from the OF in a small neighborhood of the FOE. From equations (3) and (4) we obtain the following equation for TTC estimation:

$$TTC = \frac{Z}{T_z} = \frac{FOE_x - x}{u} = \frac{FOE_y - y}{v} \tag{5}$$

The main problem in motion interpretation is due to unknown depth $Z(x, y)$ depending on different depth present in the scene. Though the FOE position is independent of the distances of world surfaces, the induced vector flow field depends on the unknown depths $Z(x, y)$ of the observed surfaces.

Generally this problem is overcome by making hypothesis of approximation to a planar surface. It is trivial to show that only the modulus of 2D velocity vectors depends on $Z(x, y)$, while the directions are independent. Since the FOE location is affected only by velocity direction we can make the OF, to be analyzed, independent of $Z(x, y)$ by normalizing the velocity vector modulus. In such manner we obtain a new OF $V(x, y)$ independent of world surfaces.

$$V(x,y) = (T_x - xT_z, T_y - yT_z) \tag{6}$$

3 A Least Square Error Technique

The 3D motion interpretation problem involves to solve the system of equations (6) for the three motion parameters: (T_x,T_y,T_z). For this purpose we need a minimum of three equations to constraint the three unknowns. This means that at least three flow vectors are required. If the flow vectors can be accurately measured we can apply the previous scheme to the few reliable vectors to find the motion parameters. In this case the 3D interpretation problem is very simple. However, in the real world, it is virtually impossible to get accurately measured flow vectors from imagery. Thus, it is clear that the principal difficulty in the 3D interpretation of visual motion comes from the unavoidable errors in visual motion measurement.

By (2) we derive that a normalized 2D motion field can be expressed as a linear combination of three basis vector fields $\{\psi_i(x,y)\}$ with coefficients corresponding to the parameters (T_x,T_y,T_z).

$$\psi_1(x,y) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \tag{7}$$

$$\psi_2(x,y) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{8}$$

$$\psi_3(x,y) = \begin{pmatrix} -x \\ -y \end{pmatrix} \tag{9}$$

We wish to represent $V(x,y)$ by projecting it onto this set of vectors $\{\psi_i(x,y)\}$. The solution is not simple because of the noise in (u,v) estimates and the non orthogonality of vectors to be combined $\{\psi_i\}$.

The correct projection coefficients $\{c_i\}$ $(c_1 = T_x,\ c_2 = T_y, c_3 = T_z)$ must minimize the following energy function:

$$E = \left\| V(x,y) - \sum_{i=1}^{3} (c_i \psi_i(x,y)) \right\| \tag{10}$$

The desired set of flow coefficients $\{c_i\}_{i=1,\ldots,3}$ are determined by an optimization criterion, minimizing the squared norm of the difference vector E. The norm E will be minimized only when its partial derivatives with respect to all of the three flow coefficients c_i equal zero.

Satisfying this condition for each of the c_i then a system of three equations in three unknowns is generated:

$$\sum_{x,y} \psi_i(x,y)(V(x,y) - \sum_{k=1}^{3} (c_k \psi_k(x,y)) = 0 \tag{11}$$

The three flow coefficients must be computed by solving equations (11).

It is completely impracticable to solve this huge system of simultaneous equations by algebraic methods such as matrix manipulation, because the complexity of such methods is $O(n!)$ (where n is the number of simultaneous equations). Methods based upon iterative improvement are much faster.

The difference-vector cost function E is quadratic in each member $\{c_i\}$, and so a unique global minimum exists.
We will demonstrate through some experimental results how correct FOE positions can be recovered by translational components (T_x,T_y,T_z) estimated minimizing the energy function E using the neural network proposed in the following section.

4 The Neural Network

We propose a neural network converging through iteration upon the desired coefficients by implementing a gradient descent along the $E(c_i)$ surface, which expresses the quadratic cost function's dependency on all of the $\{c_i\}$ coefficients. The network we propose consists of two layer units. In the first layer each i-th unit has the internal state representing the corresponding coefficient c_i. The first layer internal states are updated iteratively until a stable state is reached. The second layer has N units (corresponding to the N available sparse velocity vectors), and each is connected to all first layer units. The vector fields $\{\psi_i(x,y)\}$ are represented as fixed weights of the neural connections between the two layer units.

$$w_{ij} = \psi_i(x,y) \tag{12}$$

The normalized vector field $V(x,y)$, to be analyzed, is used as bias in the second layer

$$\forall j = 1,..,N \; b_j = V(x,y) \tag{13}$$

to compute the adaptive control signal Δ_i to adjust each of the internal states on the first layer S_{1i}

$$\Delta_i = \sum_j^N b_j - S_{2j} w_{ij} \tag{14}$$

where S_{2j} represent the second layer internal state

$$S_{2j} = \sum_i^3 S_{1i} w_{ij} \tag{15}$$

Thus, the iterative rule for adjusting the first layer internal states S_{1i} is:

$$S_{1i} = S_{1i} + \Delta_i \tag{16}$$

The equilibrium state of the network is reached when all $\Delta_i = 0$, that is the state in which the cost function E has reached its minimum; this is the point at which the partial derivatives of E with respect to all of the coefficients are null. Thus, in the stable state, the first layer of the network has internal states which represent the optimal coefficients $\{c_i\}_{i=1,...,3}$ for the projection of the optical flow $V(x,y)$ onto the set of elementary functions $\{\psi_i(x,y)\}_{i=1,...,3}$.

$$\forall i = 1..3 \; S_{1i} = c_i \tag{17}$$

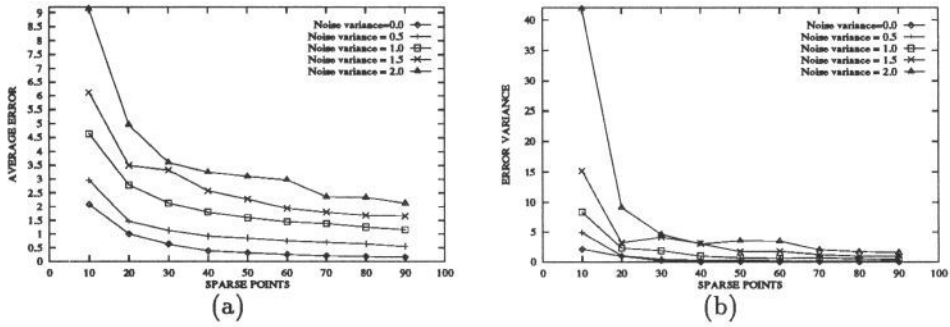


Figure 1: (a) Average error and (b) error variance of FOE location estimates for theoretical sparse flow fields of size 100×100, by varying the number of sparse points

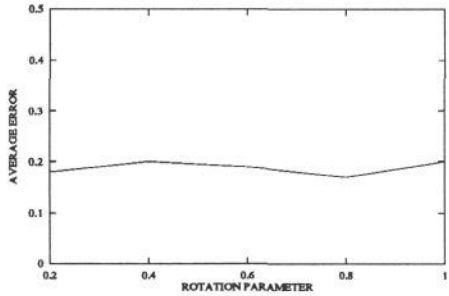


Figure 2: Average error of FOE location estimates for theoretical sparse flow fields of size 100×100, by varying the rotational velocity parameter. Uniform noise mean $= 0$. Uniform noise variance $= 1$. Sparse points $= 20$.

5 Experimental Results

Experiments on theoretical flow fields, computed by combining the basis flow fields and by adding uniform noise, and on real image flow fields, computed from sequences of natural images through the token-based approach proposed in [1], have been performed.

The TTC have been computed in number of frames before collision because the image sequences shown in this paper have not been acquired in real time.

Several experiments were performed on flow fields generated by setting randomly the FOE position and adding uniform noise with mean 0 and variance $0.5, 1, 1.5, 2$. Results obtained by varying the number of sparse points and the rotational velocity are plotted respectively in graphics in fig.1 and 2.

Some results obtained from synthetic and real image sequences are also reported. To evaluate the performances of the system the TTC has been computed through the equation (5) using the computed FOE.

In table 1 the estimated FOE coordinates and the computed against the actual TTC are reported, for the image sequences reported in fig.3, 4, 5, 6, 7. For each experiment a sequence image (a) with a OF estimated through the token based

approach [1] (b) are reported.

6 Conclusions

A vision-based "neural" system for the control of autonomous robot navigation in indoor environments has been proposed and simulated on a sequential machine. Experimental results on various synthetic and real images are extremely encouraging even for very complex scenes and poor OF maps.
It has been shown that also when very few features are present in the image good estimates of ego-motion parameters are provided. Extensive experimentation on sequences of images of different objects has revealed that in the special case of a highly textured plane, when the number of sparse points increases, a high accuracy of TTC and heading direction can be obtained.

References

[1] A.Branca, G.Convertino, A.Distante. *Hopfield Neural Network for Correspondence Problems in Dynamic Image Analysis.* Submitted to ICANN'95, International Conference on Artificial Neural Networks, October 1995, Paris.

[2] W. Burger, B. Bhanu. *Estimating 3D Ego-motion from Perspective Image Sequences.* IEEE Trans. on PAMI, vol. **12** no. 18, pp 1040–1058, November 1990.

[3] J.J. Gibson. *The Perception of the Visual World.* Houghton Mifflin, Boston 1950.

[4] H.C.Longuet-Higgins, K.Prazdny. *The Interpretation of a Moving Retinal Image.* Proc. Roy. Soc. Lond. Ser. B 208, pp 385–397, 1980.

[5] H.P.Moravec. *The Stanford Cart and the CMU Rover.* Proc. IEEE, vol. **71** no. 7, pp 872–878, 1983.

[6] K. Prazdny. *Determining the Instantaneous Direction of Motion from Optical Flow Generated by a Curvilinearly Moving Observer.* CGIP, vol. **17**, pp 238–248, 1981.

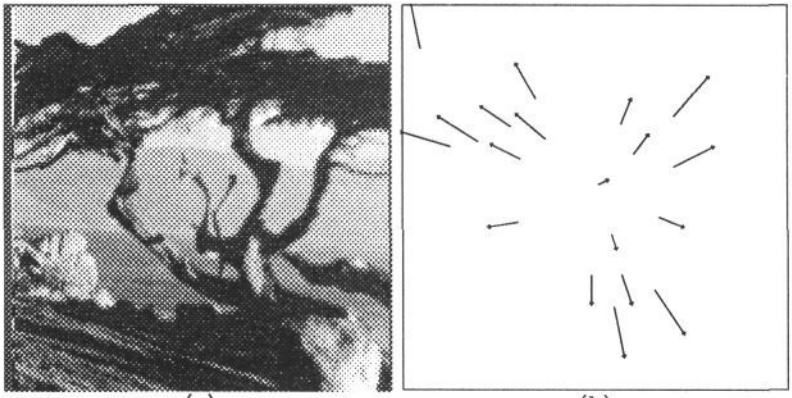


Figure 3: Diverging Tree Sequence created by Fleet. The camera translates along its line of sight and the FOE is at the center of the image plane whose size is 150×150 pixels. The FOE has been estimated at (74,76)

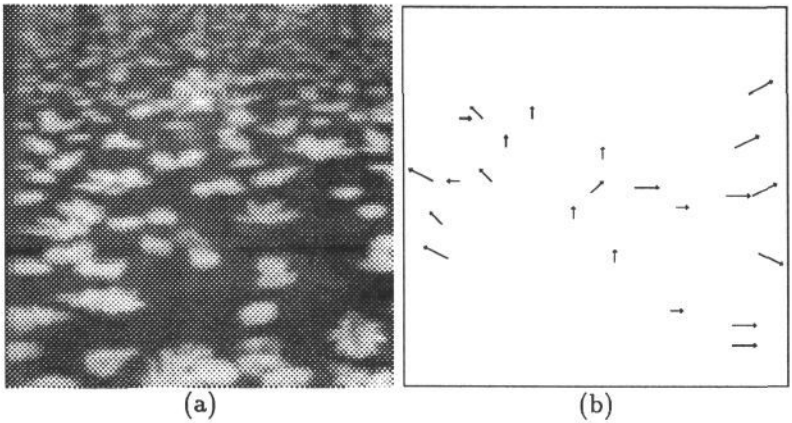
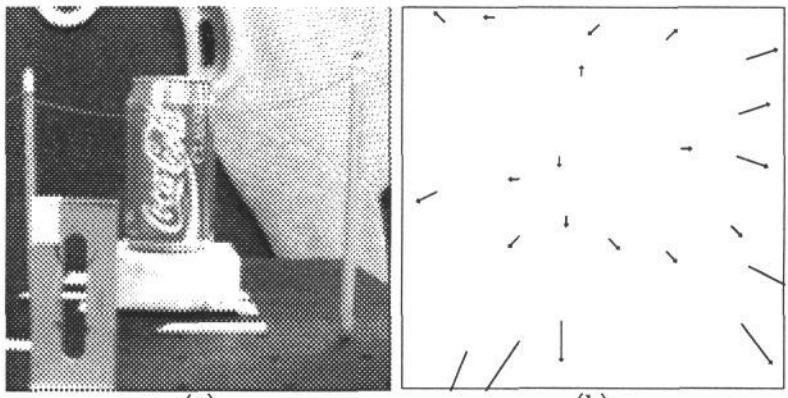
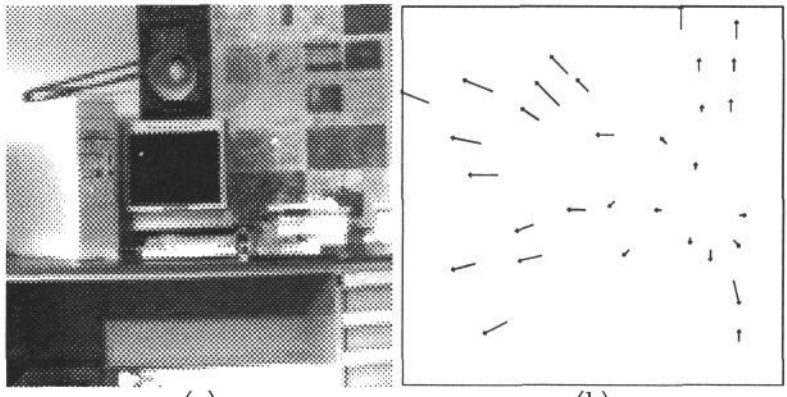


Figure 4: ESCHER: A poster by Escher. The experimental setup consists of a COHU camera mounted on a translating bench. Image size: 128×128 pixels. Initial distance camera-scene: 2350mm



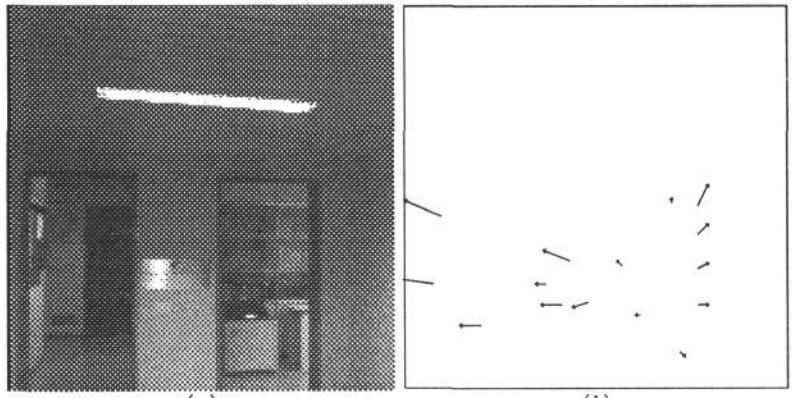
(a) (b)

Figure 5: NASA: Sequence collected at NASA Ames Research Center. Image size: 150 × 150 pixels. Initial distance camera-scene: 600mm



(a) (b)

Figure 6: LAB1: Image of a laboratory. The CCD camera is mounted on the mobile platform LAB-MATE by TRC. Image size: 128 × 128 pixels. Initial distance camera-scene: 3000mm



(a) (b)

Figure 7: LAB2: Image of a laboratory. The CCD camera is mounted on the mobile platform LABMATE by TRC. Image size: 143×143 pixels. Initial distance camera-scene: 6000mm

IMAGE	ACT. T_z (mm/frame)	FOE	OUT. TTC (frame)	ACT. TTC (frame)
ESCHER	40	14,102	50.39	57.75
ESCHER	60	44,86	38.85	38.166
ESCHER	80	44,89	29.65	28.375
NASA	15	69,34	35.11	39
NASA	20	71,33	28.97	29
NASA	25	67,35	23.70	23
LAB1	100	114,69	26.26	29
LAB1	150	115,70	21.57	19
LAB1	200	109,66	16.99	14
LAB2	400	93,112	12.03	13
LAB2	400	70,89	11.40	12
LAB2	400	74,89	10.51	11
LAB2	400	89,70	9.08	8
LAB2	400	71,90	7.38	7

Table 1: Estimate of FOE location and TTC (in number of frames) from some natural image sequences