Focus of Expansion Estimation for Motion Segmentation from a Single Camera

José Rosa Kuiaski, André Eugênio Lazzaretti and Hugo Vieira Neto
Graduate School of Electrical Engineering and Applied Computer Science
Federal University of Technology - Paraná, Brazil
jose.kuiaski@yahoo.com.br, andrelzt@yahoo.com.br, hvieir@utfpr.edu.br

Abstract

This paper proposes a new approach to motion segmentation from video sequences acquired using a single camera, whose aim is to identify which components are due to pure egomotion *and which components are due to* independent *moving objects within the observed motion field. The model has three main steps, namely computation of the optical flow field, estimation of the focus of expansion and classification under constraints. Preliminary results show that estimation of the focus of expansion followed by simple clustering techniques is promising for the achievement of motion segmentation when using a single camera with no* a priori *information about the environment.*

1. Introduction

Animal survival relies on the ability to perceive movements, distinguishing the static environment from moving agents, so that the animal can act appropriately. As the majority of objects in real environments is usually stationary, biological visual systems became sensitive to movement events and the environment itself can be considered stationary [14]. This ability is also very useful for artificial agents aiming to operate in dynamic environments.

Reproducing the ability of motion segmentation computationally is a complex task, which involves defining decision rules and handling noise. The computational problem is even more complicated when there is no *a priori* information about the *egomotion* (i.e. camera motion) or the behaviour of moving objects in the environment, which is almost always the case.

According to Gibson's Ecological Theory of Perception [4], the human visual system is able to extract all the necessary environmental information from the optical flow field at the viewer's retina. In his theory, Gibson states that the environmental optical information converges to a single point in space, forming the *Dynamic Ambient Optical Array*. Although this may be a controversial statement, it will be considered here when dealing computationally with movement analysis.

Let the focus of expansion be the representation of the *Dynamic Ambient Optical Array* on the 2D image projection of the environment, namely a single point in space where all optical flow vectors due to camera movement (image background optical flow vectors) intersect. So, estimating the focus of expansion may provide, according to Gibson's theory, all the *egomotion* information necessary to motion segmentation.

In a computational vision system, the information about movement is obtained from computing the optical flow field, which can be obtained in several ways, such as in Horn and Schunk [6], Lucas and Kanade [8], Shi and Tomasi [16], Nordberg and Farneback [13] or Fleet and Weiss [3]. Therefore, it may be possible to extract all – or almost all – information about the environment from the calculated optical flow field.

Video sequences acquired with stereo cameras simulate similar conditions as the ones present in the human visual system, but the availability of more than a single camera may not always be possible. With a single camera, depth information without *a priori* information is not available, but a possible solution for this problem is to model movement from image motion, rather than from object motion.

In this paper we compute image motion in the form of optical flow vectors obtained from a single camera and make some assumptions due to the unavailability of *a priori* information about *egomotion*:

1. The majority of optical flow vectors within the optical flow field are relative to *egomotion*, so the majority of pixels within a frame of a video sequence are related to the static environment (background).
2. The environment optical flow vectors theoretically converge to a single point in space – the *Dynamic Ambient Optical Array* – from now on called the focus of expansion (FoE) point.

3. The presence of independent moving objects and also noise in the optical flow computation will generate multiple additional FoE candidate locations, corresponding to different moving groups of pixels (objects).

The first step to consider is computing the optical flow field. Our current model is based on the algorithm by Nordberg and Farneback [13] for dense optical flow computation – in section 2, considerations about the use of dense or sparse algorithms will be made. Once the optical flow field is available, we estimate FoE locations for the optical flow vectors – this is a key step to our motion analysis approach and will be detailed in section 3. The proposed idea for the estimation of FoE points is to find the intersections of every pair of optical flow vectors within a pair of frames from a video sequence. Clustering of the resulting intersections and a decision rule are then necessary to determine the FoE location for the environment, which corresponds to *egomotion*, and for any moving objects that may be present in the scene.

2. Optical Flow Estimation

The image registration technique proposed by Lucas and Kanade [8] shows that it is possible to use the spatial intensity gradient to optimise the search for the position that yields the best match between a pair of images. This is particularly important when dealing with optical flow. In motion analysis, the optical flow field is the distribution of motion vectors, whose magnitudes mean the apparent velocity and whose directions indicate the relative trajectory between moving objects and the viewer.

Let us first consider image registration techniques as the mathematical foundation for optical flow understanding when dealing with two consecutive frames of a video sequence. Each pixel at the first frame possibly corresponds to a specific pixel at the second frame, as shown in figures 1 and 2. Figure 1 shows two consecutive frames of the optical_flow_input.avi video sequence, in which a moving car is followed by the camera [17], while figure 2 shows the corresponding optical flow field computed by the algorithm by Lucas and Kanade (the first frame is the reference for the optical flow vectors). Each optical flow vector shows the estimated movement for its corresponding underlying pixel.

In figure 2 one can clearly notice that the vast majority of optical flow vectors are due to the camera translation to the right, which results in apparent background image translation to the left.

One of the assumptions in computation of optical flow is that the brightness of an image feature does not change between frames, i.e. if $E(x, y, t)$ is the brightness of a given pixel at time t, its first partial derivative in time should be zero, as follows:

Figure 1. Two consecutive frames of the optical_flow_input.avi video sequence in which a moving car is followed by the camera [17].

$$\frac{\partial E}{\partial t} = 0. \tag{1}$$

Equation 1 can be rewritten as a function of the brightness gradient, leading to equation 2:

$$\frac{\partial E}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial E}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial E}{\partial t} = \nabla E^T \cdot \vec{v} = 0, \tag{2}$$

where $\vec{v} = (v_x, v_y, 1) = \left(\frac{\partial x}{\partial t}, \frac{\partial y}{\partial t}, 1\right)$ is the motion vector for a single point or region Ω within an image and ∇E^T is the spatio-temporal brightness gradient of the same region.

According to Nordberg and Farneback [13], there is no unique solution for $\vec{v}$ in equation 2. A possible solution for this equation would involve a constraint, say $\nabla E = 0$, and considering a region Ω where $\vec{v}$ could be assumed constant. Therefore, the local mean value over Ω carries all the spatio-temporal information of the object's orientation, as follows:

$$\left[\int_{\Omega} p(x)\left(\nabla E_x\right)\left(\nabla E_x\right)^T dx\right] \cdot \vec{v} = 0, \tag{3}$$

where p is a Gaussian function and the first term in the multiplication is a second-moment matrix derived from ∇E, the so-called *structure tensor*.

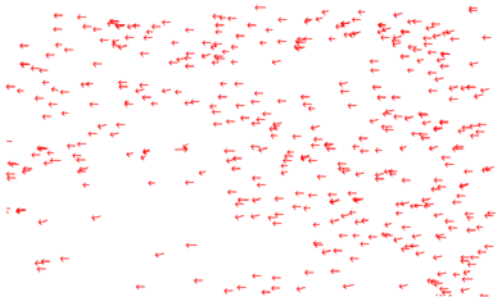

Figure 2. Optical flow field from matching points of the frames in figure 1 computed with the algorithm by Lucas and Kanade.

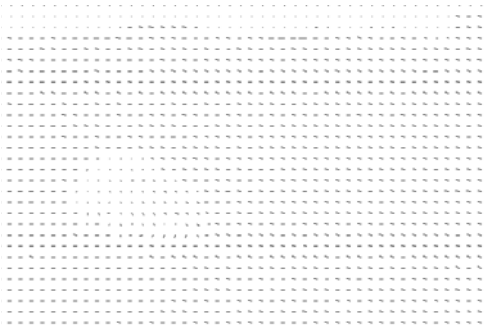

Figure 3. Dense optical flow field for the frames in figure 1 computed with the algorithm by Nordberg and Farneback.

Horn and Schunk [6] also suggest an additional smoothness constraint in which $\nabla^2\left(\frac{\partial x}{\partial t}\right) = \nabla^2 v_x$ and $\nabla^2\left(\frac{\partial y}{\partial t}\right) = \nabla^2 v_y$ (the Laplacians of velocity components) must be minimised.

2.1. Dense Optical Flow

Horn and Schunk's algorithm and also Nordberg and Farneback's algorithm consider an $m \times n$ window within the image to estimate the optical flow at its central point. Therefore, a uniformly sampled grid of points with their respective optical flow vectors is obtained. Figure 3 shows such a dense optical flow field, in which each vector connects the grid point to the calculated position for that same point in the consecutive frame. The use of dense optical flow in figure 3 clearly shows that there are noticeable differences in motion between object and background.

Dense optical flow may be more sensitive to the aperture problem, i.e. more prone to select edges from straight line segments in the input video sequence, which will result in the computation of ambiguous motion. Also, rigid moving objects that are larger than the considered analysis window may reflect in noisy components in the optical flow field, requiring a multiscale approach do be dealt with. However, the use of dense optical flow favours the analysis of optical flow vectors as stochastic processes. The static locations of flow vectors obtained in the dense approach do not depend on any tracking algorithm and ensure their behaviour as random variables for the stochastic process at time t.

The experiments reported in this paper were based on dense optical flow in order to model motion as a stochastic process, making it possible to use Independent Component Analysis [7] for motion classification and segmentation in future work, when we plan to work with optical flow vectors from two different observation points – right and left frames acquired from a stereo vision system, for example. The technique that involves estimation of the focus of expansion reported here is a preliminary approach to the problem.

2.2. Sparse Optical Flow

The most noticeable difference when dealing with sparse optical flow is the use of interest point detectors before computing the optical flow field. Figure 2 already showed an example of an sparse optical flow field in which the Harris interest point detector [5] was used to select which pixels within the image are those whose information is more relevant, and the optical flow field was computed using a multiscale version of Lucas and Kanade's algorithm [2]. This approach ensures greater robustness to the aperture problem, reducing noisy components. The sparse optical flow computation approach reduces wrong estimations of flow vectors when frames from the video sequence under analysis have a relevant proportion of homogeneous regions.

3. Focus of Expansion Estimation

The focus of expansion (FoE) corresponds to Gibson's *Dynamic Ambient Optical Array* [4], which is a single point in space where all the flow vectors should converge. Its main uses in vision applications are the estimation of the time-to-impact (TTI) in visual navigation and the 3D reconstruction of the environment. There are three usual approaches to FoE estimation in the literature [12]: discrete, differential and least-squares, all of them yielding good results for pure translational motion.

This papers adopts the differential approach, since our aim is to estimate the focus of expansion from the optical flow field. The differential approach is more robust, but

computationally heavy, as a consequence of not performing any least-squares minimisation [15].

The focus of expansion estimation plays an important role in motion segmentation. Our hypothesis is that it is possible to classify all optical flow vectors within the optical flow field as belonging to the background (environment) or to independent moving objects, from the estimated FoE, using a simple rule:

> "If a flow vector belongs to the environment, its backwards line segment should cross the FoE."

In this way, it seems to be possible to estimate all *egomotion* components for pure translation. Considering the general assumption that the environment is static and also assuming that the majority of pixels within an image frame represents the environment, it is possible to argue that the majority of flow vectors from a dense optical flow field are *egomotion* components due to camera translation. From these assumptions, it is possible to state the following:

1. For every pair of optical flow vectors, there may be an associated FoE. If there is not, one of the optical flow vectors is related to an independent moving object within the image frame.
2. FoE candidates originated from two environmental optical flow vectors will be concentrated within an small area in space, whose variational width will be a function of noise on the optical flow field computation due to numerical approximations.
3. FoE candidates originated from mixed optical flow vectors (one from the environment and one from an independent moving object), or from two optical flow vectors from independent moving objects, will result in sparsely and randomly distributed locations in space.

Therefore, it is assumed that the great majority of calculated FoE candidates will lie within a small area in space. A simple histogram analysis could inform precisely where the largest concentration of candidates lies. However, FoE candidates can be distributed across the infinite 2D space and computing such a histogram would be computationally infeasible. Therefore, clustering the resulting FoE candidates is necessary. Using a clustering approach, one can infer that the centroid of the cluster with the largest number of samples is the most likely FoE for the environment.

3.1. Naïve Estimation

A first approach to the focus of expansion estimation is to naïvely consider some *a priori* information about the optical flow vectors. If at least three of them are known to be from the environment, they can be simply triangulated into a small region and their average location will be considered the FoE. This approach works fine for simple video sequences with pure translational motion, but relies on *a priori* information.

3.2. Clustering by K-means

If no *a priori* information should be used, a possible solution is to perform clustering on the estimated FoE array. The K-means technique [9] can be used to split an array with n samples $(X_0, X_1, \ldots, X_n)$ into m clusters $(C_0, C_1, \ldots, C_m)$ whose total squared energy is minimal, as follows:

$$\text{Energy} = \arg\min \sum_{i=1}^{m} \sum_{\mathbf{X}_j \in \mathbf{C}_i} \parallel \mathbf{X}_j - \mu_i \parallel^2, \qquad (4)$$

where $\mathbf{X}_j$ is the j^{th} sample and $\mathbf{C}_i$ is the i^{th} cluster with centroid μ_i.

The most noticeable disadvantage in using K-means to cluster the FoE candidates is the need of *a priori* information about the number of clusters m. In most general cases, this information is not available.

4. Experimental Setup

After dense optical flow field computation, the first step is to count the total number of optical flow vectors that are parallel to the x or y axes. If the majority of optical flow vectors is parallel to one of these axes, it is inferred that the FoE lies in the positive or negative infinity for x or y, depending on the case. The second step is to count the number of parallel vectors lying in the same direction. This step informs whether the FoE lies at the positive or negative infinity for x or y. These two steps can be merged into a single step if a directional angle histogram is computed and the majority of directional angles lies within a small region of the histogram. However, this procedure is computationally more expensive than the previous two separate steps.

In both cases, we considered that two vectors are still considered parallel if they differ by a small value ϵ from each other. This approximation compensates for numerical errors in the optical flow field computation and yields good approximations for FoE candidates that lie at infinity. If all cases above do not apply, this implies that the FoE is at a finite location in 2D space and this is the case in which FoE estimation must take place. The model was implemented using OpenCV and the experiments were conducted using the optical_flow_input.avi video sequence [17]. Ground-truth data was obtained by counting correlated and uncorrelated optical flow vectors frame by frame.

5. Results

The naïve estimation used locations known to be background *a priori*, namely the coordinates (160, 160), (320, 560) and (320, 160). For the first 50 frames of the optical_flow_input.avi video sequence, the average error rate for the background was 13.8% and the average error rate for objects was as high as 90%. This large average error for objects led us to notice that if there are homogeneous regions in the frame (such as the sky in the optical_flow_input.avi video sequence), there will be no optical flow within these regions.

Since the approach being used was naïve, it was possible to deal with the sky region by labelling it *a priori* as background. By doing this, the average error rate for the background went down to 8.7% and the average error rate for objects went down to about 74%. A more careful observation showed that the optical_flow_input.avi video sequence yielded significantly small optical flow vectors in the majority of its frames – a consequence of having a camera following the car at low speed – which seemed to be the main reason for such a large average error rate for the object. Using another video sequence, Office_left.avi [11], with less camera movement and much more significant object translation – therefore resulting in larger optical flow vectors – yielded an average error rate for the background of 12% and an average error rate for the object of 22.4%, confirming our hypothesis.

Finally, the K-means clustering was applied to the optical_flow_input.avi video sequence. Since there is no *a priori* information about the spatial location of the FoE for every pair of optical flow vectors, the number of clusters had to be determined empirically and was set to seven. Figure 4 shows the spatial distribution of the estimated FoE candidates after clustering, where the x and y axis represent spatial coordinates in pixels.

In figure 4, the location most likely to be the FoE is the centroid of the cluster with the largest sample density. In this case, the FoE was considered to lie in coordinates (0, 250000), which is the centroid of the yellow cluster. In our implementation, this result corresponds to an FoE at positive infinity in x, which is coherent with the motion observed in the corresponding video sequence. The K-means clustering approach yielded an error rate of 5.9% for the background and 80% for objects, not dealing with the homogeneous sky region problem. After filtering the sky region using *a priori* information, the error rate for the background dropped to 5% and the error rate for objects dropped to 30%.

No K-means clustering was necessary in the experiments with the Office_left.avi video, due to the *a priori* information that no background movement, i.e. camera movement, was present.

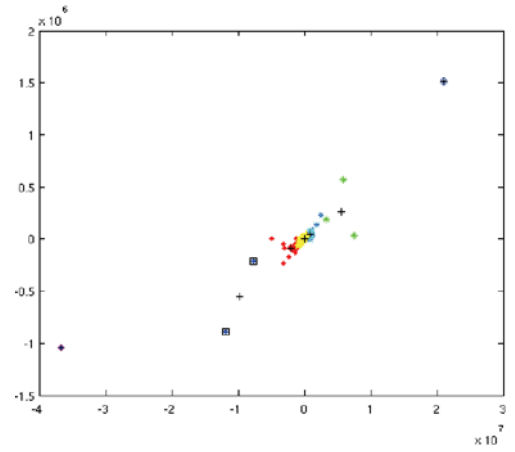

Figure 4. Spatial distribution of FoE candidates. The cluster with highest sample density lies in coordinates (0, 250000), which was considered the FoE.

6. Conclusions

The preliminary results shown in this paper, despite being beyond a reasonable error tolerance, show that it is possible to classify image motion through focus of expansion estimation. The large average error rates are attributed to the simplistic approaches for motion segmentation that were used (naïve estimation and K-means clustering). There seems to be great potential of use of the method in situations where a second camera or information about the camera's trajectory and speed is not available *a priori*.

Other issues, such as the interference of homogeneous regions, lack of optical flow vectors of significant magnitude and presence of noise in the optical flow computation suggest the use of adaptive clustering algorithms. Possible adaptive clustering algorithms to be used in future investigations include the GWR Neural Network [10] and Support Vector Machines [1], avoiding the need of *a priori* definition of the number of clusters.

The use of the approach presented here is intended to deal with pure camera translation and zooming when there are much more optical flow vectors due to *egomotion* than to object motion. Dealing with rotations would require the estimation of a focus of rotation in addition to the estimation of the focus of expansion. It should be noticed that the current approach is not well suited for real-time applications, as the amount of calculations slows down considerably the input video sequences. We expect that the use of neural network approaches for clustering will result in more robust classification for the determination of the focus of expansion.

References

[1] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik. Support vector clustering. *Journal of Machine Learning Research*, 2:125–137, December 2001.

[2] J.-Y. Bouguet. *Pyramidal Implementation of the Lucas Kanade Feature Tracker: Description of the Algorithm*. Intel Corporation, Microprocessor Research Labs, 2002.

[3] D. J. Fleet and Y. Weiss. *The Handbook of Mathematical Models in Computer Vision*, chapter Optical Flow Estimation, pages 239–258. Springer, 2005.

[4] J. J. Gibson. *The Ecological Approach to Visual Perception*. Lawrence Erlbaum Associates, Hillsdale, 1986.

[5] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conference*, pages 147–151, Manchester, 1988.

[6] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1–3):185–203, August 1981.

[7] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley Series on Adaptive and Learning Systems for Signal Processing, Communications, and Control. Wiley, New York, 2001.

[8] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference in Artificial Intelligence*, pages 674–679, Vancouver, Canada, August 1981.

[9] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, Berkeley, 1967. University of California Press.

[10] S. Marsland, J. Shapiro, and U. Nehmzow. A self-organising network that grows when required. *Neural Networks*, 15(8–9):1041–1058, October–November 2002.

[11] B. McCane, K. Novins, D. Crannitch, and B. Galvin. On benchmarking optical flow. *Computer Vision and Image Understanding*, 84(1):126–143, October 2001.

[12] S. Negahdaripour and B. K. P. Horn. A direct method for locating the focus of expansion. *Computer Vision, Graphics, and Image Processing*, 46(3):303–326, June 1989.

[13] K. Nordberg and G. Farneback. A framework for estimation of orientation and velocity. In *Proceedings of the 2003 IEEE International Conference on Image Processing*, volume 3, pages 57–60, 2003.

[14] S. E. Palmer. *Vision Science: Photons to Phenomenology*. MIT Press, Cambridge, 1999.

[15] D. Sazbon, H. Rotstein, and E. Rivlin. Finding the focus of expansion and estimating range using optical flow images and a matched filter. *Machine Vision and Applications*, 15(4):229–236, 2004.

[16] J. Shi and C. Tomasi. Good features to track. In *Proceedings of the 1994 IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, Seattle, June 1994.

[17] D. Stavens. Lecture notes on optical flow and OpenCV. URL: http://robotics.stanford.edu/~dstavens/cs223b/, 2007. CS223-B Guest Lecture, Stanford University.