Motion Estimation and Figure-Ground Segmentation Using Log-polar Images*

V. Javier Traver, Filiberto Pla
Computer Vision Group, Universitat Jaume I, E12080-Castellon (Spain)
{vtraver|pla}@uji.es

Abstract

A motion estimation algorithm for log-polar images, based on a previous general framework [2], is presented. The advantages and disadvantages brought by the use of this kind of foveal imaging are discussed. Within this context, a simple, but quite effective approach for figure-ground segmentation is also proposed. Performance results concerning motion estimation and target segmentation are reported.

1. Introduction

Motivation and related work. The use of space-variant images has proved advantageous over uniformly sampled images for a variety of problems in pattern recognition and computer vision problems. In particular, it turns to be especially suitable in active vision contexts. The log-polar transform, which mimics its biological retino-cortical mapping counterpart, has successfully been used in a number of problems, such as binocular tracking [1], time-to-impact computation [3], etc. Among the most direct benefits of log-polar images are their much smaller sizes —while preserving a wide field of view—, and their rotation and scale invariance properties.

For the problem of actively tracking a moving target, log-polar mapping is also helpful. But, besides a good image representation, for a tracking system to be successful, a fast and effective motion estimation algorithm is needed, in order to compensate for the relative motion between the observed target and the active observer.

A great deal of work has been done for solving motion estimation problems. However, even the same practitioners feel that the problem is by no means solved in all its generality. Comparatively, motion estimation for log-polar images is still in its infancy, partly because not every technique proposed for cartesian images is directly applicable to space-variant images. Thus, important research challenges remains to be addressed in this area. This is one of the motivations of the present work.

*Research supported in part by projects GV97-TI-05-27 from the *Conselleria d'Educaci, Cultura i Cincia, Generalitat Valenciana*, and CICYT TIC98-0677-C02-01 from the Spanish *Ministerio de Educacin y Cultura*.

(a) (b) (c) (d)

Figure 1. Log-polar mapping: (a) grid layout example (10×16), (b) original cartesian image (256×256), with grid (a) overlapped, (c) cortical image (64×128), (d) retinal image (256×256) obtained from (c) by the inverse mapping.

In this paper, a motion estimation and tracking algorithm in the log-polar domain is presented. It is based on a general framework introduced in [2] (H&B hereafter), which was originally proposed for cartesian images. Additionally, the figure-ground segmentation problem is addressed.

Log-polar mapping. The log-polar mapping used here defines the log-polar coordinates $(\xi, \eta) \triangleq \left(\log_a\left(\frac{\rho}{\rho_0}\right), \theta\right)$, with (ρ, θ) being the usual polar coordinates. The parameters of the transform are $q = \frac{S}{2\pi}$, ρ_0, and $a = \exp(\ln(\frac{\rho_{\max}}{\rho_0})/R)$, with R and S being the number of rings and sectors, respectively, of the log-polar image.

From their biological motivation, *retinal* images are those in the usual format, while *cortical* images are those resulting from the log-polar mapping (i.e., the log-polar images themselves). An example is shown in Fig. 1.

2. Theoretical background

Original formulation [2]. Let $I(\mathbf{p}, t)$ denote the gray-level value at a given image location $\mathbf{p}$ of an image I ac-

quired at time t. A general parametric *motion model* is defined by $\mathbf{f}(\mathbf{p};\boldsymbol{\mu})$, with $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)^T$ the motion parameter vector. We have that $\mathbf{f}(\mathbf{p};\mathbf{0}) = \mathbf{p}$. The image at an initial time t_0, I_0, will be denoted by the *reference image*, where a set of $N > n$ locations $\mathcal{R} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}$ define a *target region*. Let $\boldsymbol{\mu}^*(t)$ be the ground truth values of $\boldsymbol{\mu}$ at time t, and $\boldsymbol{\mu}(t)$ the corresponding estimate. If changes in subsequent images are only due to $\mathbf{f}$, then for any $t > t_0$, there is a $\boldsymbol{\mu}^*(t)$ such that $I(\mathbf{p}, t_0) = I(\mathbf{f}(\mathbf{p};\boldsymbol{\mu}^*(t)), t), \forall\, \mathbf{p} \in \mathcal{R}$.

The image of the target region, transformed as of $\boldsymbol{\mu}$, can be written in vector notation as:

$$\mathbf{I}(\boldsymbol{\mu}, t) = \begin{bmatrix} I(\mathbf{f}(\mathbf{p}_1;\boldsymbol{\mu}), t) \\ I(\mathbf{f}(\mathbf{p}_2;\boldsymbol{\mu}), t) \\ \vdots \\ I(\mathbf{f}(\mathbf{p}_N;\boldsymbol{\mu}), t) \end{bmatrix},$$

which will be referred to as the *rectified image*, I_R. The estimation of the motion parameter vector $\boldsymbol{\mu}$ can be found by minimizing a *least squares* objective function which, in vector notation, can be written as:

$$O(\boldsymbol{\mu}) = ||\mathbf{I}(\boldsymbol{\mu}, t) - \mathbf{I}(\mathbf{0}, t_0)||^2. \quad (1)$$

In the absence of a good initial guess of $\boldsymbol{\mu}$, a costly global optimization procedure would be needed to optimize (1). However, in a visual tracking scenario, the continuity of motion provides this starting point. Thus, the problem can be reformulated to that of determining a vector of offsets $\delta\boldsymbol{\mu}$, such that $\boldsymbol{\mu}(t+\tau) = \boldsymbol{\mu}(t) + \delta(\boldsymbol{\mu})$. If the components of $\delta\boldsymbol{\mu}$ has a small magnitude, continuous optimization can be applied to a linearized version of the problem. Taking this into account, and with the additional approximation $\tau\mathbf{I}_t \approx \mathbf{I}(\boldsymbol{\mu}, t+\tau) - \mathbf{I}(\boldsymbol{\mu}, t)$, the solution is [2]:

$$\delta\boldsymbol{\mu} = -(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T[\mathbf{I}(\boldsymbol{\mu}, t+\tau) - \mathbf{I}(\mathbf{0}, t_0)], \quad (2)$$

where $\mathbf{M}$ is the $N \times n$ Jacobian matrix of $\mathbf{I}$ with respect to $\boldsymbol{\mu}$. When $\mathbf{f}$ is a linear function of the image coordinate vector $\mathbf{p}$, $\mathbf{M}$ can be factorized into a product of two matrices, one of them being constant and, therefore, needing only to be computed off-line, once and for all. This renders the tracking algorithm more efficient (see [2] for details).

Adaptation to log-polar geometry. As we are dealing with log-polar images rather than cartesian ones, I denotes a log-polar image, and its coordinates are $\mathbf{p} = (\xi, \eta)$. Although other motion models would be possible, in this paper we focus on a similarity motion model (i.e., translation, rotation and scaling), as it is reasonable simple and yet useful:

$$\mathbf{f}(\mathbf{p};\boldsymbol{\mu}) = \mathbf{p} + \mathbf{t}_l(\boldsymbol{\mu}) + \mathbf{J}(\mathbf{p}) \cdot \mathbf{t}_c(\boldsymbol{\mu}), \quad (3)$$

where $\mathbf{t}_l = (r, s)^T$ is a translation in the log-polar domain, and $\mathbf{t}_c = (b, c)^T$ is the common cartesian translation. Therefore, our 4-parameter motion vector is $\boldsymbol{\mu} = (r, s, b, c)$. Note how the use of the log-polar geometry simplifies the expression of the motion model regarding rotation and scaling (they are just a translation; the actual rotation angle is $\phi = r/q$, and the actual scaling factor is found as $\alpha = a^s$). However, the usual translation is modified by the log-polar Jacobian matrix

$$\mathbf{J} = \begin{bmatrix} \partial\xi/\partial x & \partial\xi/\partial y \\ \partial\eta/\partial x & \partial\eta/\partial y \end{bmatrix} = 1/\rho \begin{bmatrix} \frac{\cos\theta}{\ln a} & \frac{\sin\theta}{\ln a} \\ -q\sin\theta & q\cos\theta \end{bmatrix}.$$

An interesting advantage of using log-polar images, instead of cartesian ones, is that we do not explicitely select a target region, as it is done in [2]. Our $\mathcal{R}$ will therefore be the whole image (i.e., $N = R \cdot S$), The *implicit focus-of-attention* of log-polar images [1] will effectively deal with images with a foveated target, even when it only occupies a small part of the visual field, without the background becoming too distracting.

3. Figure-ground segmentation

As mentioned above, we make no *a priori* selection of what the target region will be. On the contrary, if a moving object (the target) is kept foveated (which is the case with an active tracking mechanism), it is possible to automatically discover the target and segment it from the background. To that end, we propose the following probabilistic approach, in which pixels are classified as either *target* or *background* pixels.

Let $P_c(\mathbf{p})$ be the *current* probability of the pixel at $\mathbf{p}$ being a target pixel, which can be estimated on the basis of the most recently estimated motion. Let $P_h(\mathbf{p}, t)$ be the *historic* probability of the same pixel being a target pixel. This *history* of the target is updated at each time step t as $P_h(\mathbf{p}, t) = \lambda \cdot P_h(\mathbf{p}, t-1) + (1-\lambda) \cdot P_c(\mathbf{p})$, where $\lambda \in [0, 1]$ can be regarded as a *forget/memory* factor, which weights the historic probability against the more recent confidence. Initially, $P_h(\mathbf{p}, t_0) = 0$ (i.e., no target identified yet).

As for $P_c(\mathbf{p})$, we choose to compare the reference and rectified images (I_0 and I_R) on a pixel-by-pixel basis, by using the squared frame difference function $D(\mathbf{p}) = (I_0(\mathbf{p}) - I_R(\mathbf{p}))^2$. The rational behind this is that, if motion estimates are accurate enough, the rectified image will look similar to the initial, reference image, at those pixels which belong to the target (whose motion is being estimated). Then, to get a probability from D, we use $P_c(\mathbf{p}) = \exp\left(-\frac{1}{2}\frac{D(\mathbf{p})}{\sigma^2}\right)$, with σ being a noise estimate.

(a) (b) (c) (d)

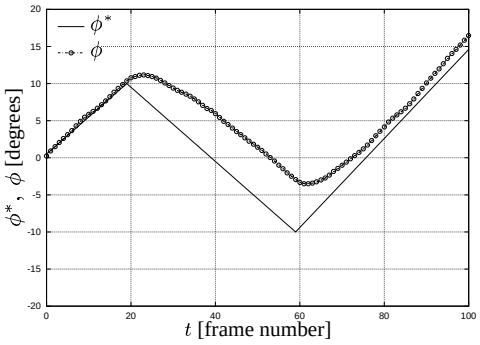
Figure 2. Target motion simulation: (a) target mask, (b) target image, (c) background image, (d) target-over-background image.

4. Experimental work

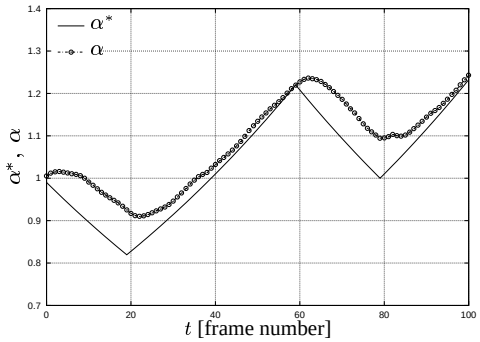
Set-up. Fig. 2 illustrates the elements of our experiments. The desired shape of a target is defined by a binary target mask (Fig. 2a) whose contents are taken from a target image (Fig. 2b). Another image serves as a background (Fig. 2c). Pasting the masked target image on this background, results in a target-over-background image (Fig. 2d). To simulate motion, we apply known transformations, according to ground truth $\boldsymbol{\mu}^*$, to the target (both the mask and image), while the background is left unchanged. The log-polar transformations of the resulting cartesian images are used as input to the approach described in Sect. 2, and the estimated motion parameter vector $\boldsymbol{\mu}$ is computed.

Results. All the four motion parameters are made to change in a range and at a rate as shown in Fig. 3, over a 100-frame sequence. Motion parameters are not changed too much at each time step, so that not to violate the assumption of "small" $\delta\boldsymbol{\mu}$. This, in turn, requires the target to move slowly, and/or a very fast processing system, which are a strong limitation in many situations. Estimation results, as can be seen, are quite good on average, taking into account how small log-polar images are (32×64), and that the target is occupying only a part of the visual field. In an active vision context, which is where the use of space-variant images makes full sense, the active camera's parameters can be readjusted when motion parameters (mainly $\mathbf{t}_c$) are big, which would also be the right moment to reset the reference image and the global $\boldsymbol{\mu}$.

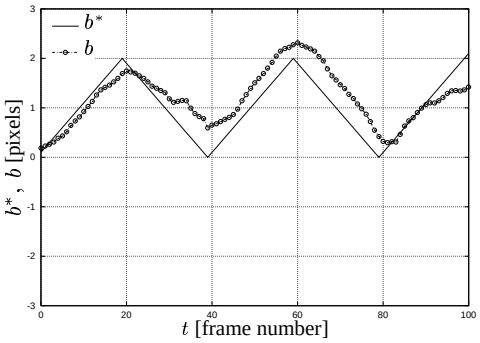
In Fig. 4, different cortical images, at selected steps in the sequence, are remapped to cartesian space for visualization purposes. The first row shows the target transformed as of the corresponding (ground truth) motion parameters. Rectified images are shown at the second row. It can be appreciated that even though the target has moved (shifted, rotated and scaled), the rectified target looks very similar to the target at the first step (reference image). It is worth noticing that, despite the target pixels are correctly rectified, the background pixels are not. This is due to the fact that we consider all the whole image as the target region, and, therefore, all the pixels are rectified according to the



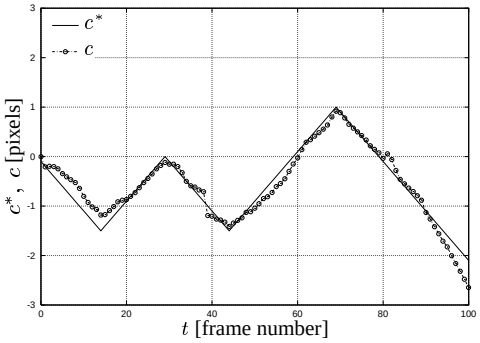
(a) True and estimated rotation angle



(b) True and estimated scaling factor



(c) True and estimated horizontal shift



(d) True and estimated vertical shift

Figure 3. Comparison of true and estimated motion parameters along the sequence.

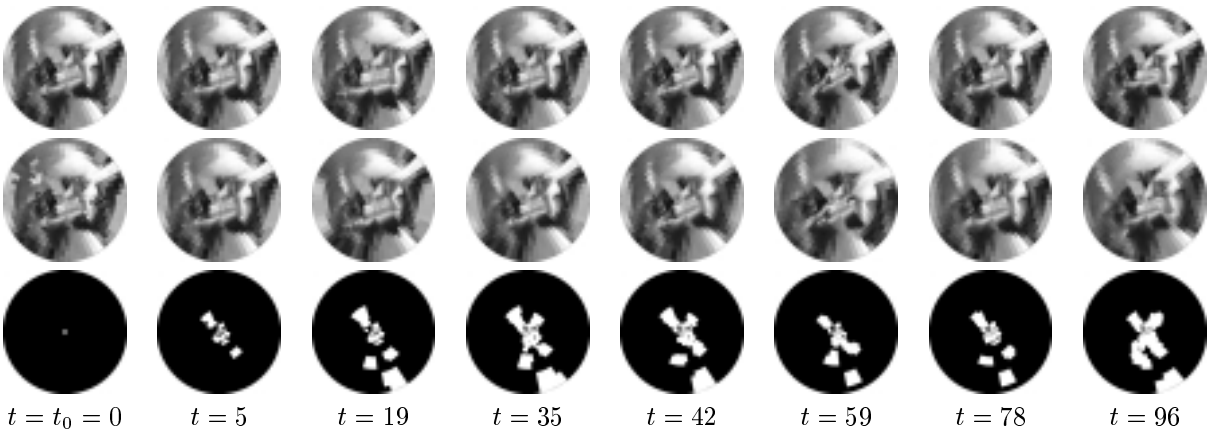


Figure 4. Results at some frames: current image, rectified image, and estimated target mask.

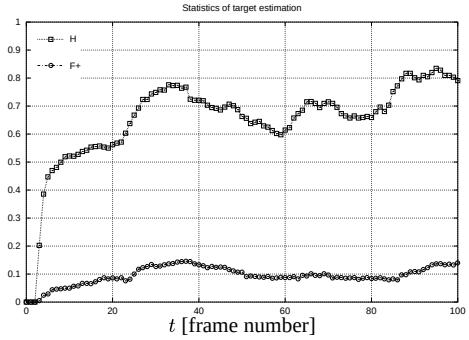


Figure 5. Figure-ground segmentation: evolution of hits and false positives over time.

estimated *target* motion parameters. The third row corresponds to the estimated target mask, i.e., a binarized version of $P_h(\mathbf{p}, t)$. It can be seen how the shape of the target is being discovered by integrating information over time. It is important to consider that even for a human observer may be difficult to segment the x-shaped target from the background because, in this example, they have similar patterns in many parts of the image. The actual target size, position, and orientation would be found by applying the estimated motion parameters to the estimated target pixels. As can be observed, some false positives can arise, because the proposed figure-ground segmentation approach relies, not only on accurate motion parameters, but also on having a non-uniform changing background. In a real active tracking scenario, the background pixels change not only due to the rectification process, but also due to the motion of the active observer, thus helping the segmentation process.

In Fig. 5 the percentage of *hits* H (number of correctly classified target pixels over the total number of target pixels), and the percentage of *false positives* F^+ (pixels wrongly classified as being target) are plotted. It can be observed how H grows as time progresses, as it should be expected, while F^+ keeps reasonably small.

5. Conclusions

A motion estimation and region tracking algorithm for log-polar images has been presented. The method is based on the general framework introduced in [2]. Results with a similarity motion model show that H&B's framework can successfully be applied with space-variant images. Besides the big information reduction, an advantage of using log-polar images is that, as opposed to [2] and other works based on it, no prior selection of any target region is needed. In this sense, the proposed figure-ground segmentation process blends particularly well in this formulation. Further work can explore other motion models, deal with factorization issues, and improve the segmentation process. Active vision can also be of help to stabilize the target image. Predicting the target motion could compensate for the limitation of requiring small changes from frame to frame.

References

[1] A. Bernardino and J. Santos-Victor. Binocular tracking: Integrating perception and control. *IEEE Trans. on Robotics and Automation*, 15(6):1080–1094, Dec. 1999.

[2] G. D. Hager and P. N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 20(10):1025–1039, Oct. 1998.

[3] M. Tistarelli and G. Sandini. On the advantages of polar and log-polar mapping for direct estimation of time-to-impact from optical flow. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 15:401–410, 1993.