Global Motion Estimation: Feature-Based, Featureless, or Both ?!

Rui F.C. Guerreiro[1] and Pedro M.Q. Aguiar[2]

[1] Philips Research, Eindhoven, Netherlands
rui.guerreiro@philips.com
[2] Institute for Systems and Robotics / Instituto Superior Técnico, Lisboa, Portugal
aguiar@isr.ist.utl.pt

Abstract. The approaches to global motion estimation have been naturally classified into one of two main classes: *feature-based methods* and *direct (or featureless) methods*. Feature-based methods compute a set of point correspondences between the images and, from these, estimate the parameters describing the global motion. Although the simplicity of the second step has made this approach rather appealing, the correspondence step is a quagmire and usually requires human supervision. In opposition, featureless methods attempt to estimate the global motion parameters directly from the image intensities, using complex nonlinear optimization algorithms. In this paper, we propose an iterative scheme that combines the *feature-based simplicity* with the *featureless robustness*. Our experiments illustrate the behavior of the proposed scheme and demonstrate its effectiveness by automatically building image mosaics.

1 Introduction

In this paper, we address the problem of estimating the global motion of the brightness pattern between a pair of images.

1.1 Motivation and State of the Art

Efficient methods to estimate global motion find applications in diverse fields. For example, in remote sensing and virtual reality, it is often necessary to build large images from partial views, *i.e.*, to register, or align, the input images. The key step for the success of these tasks is the estimation of the global motion between the images. In digital video, image alignment is also crucial, for stabilization and compression [1,2] and content-based representations [3].

Under common assumptions about the camera motion or the scene geometry, the motion of the image brightness pattern is described by a small set of parameters, see for example [4,5] for different parameterizations. In many situations of interest, the scene is well approximated by a plane, *e.g.*, when the relevant objects are far from the camera. In this case, the image motion is described by an 8-parameter homographic mapping. The homography also describes the image

A. Campilho and M. Kamel (Eds.): ICIAR 2006, LNCS 4141, pp. 721–730, 2006.

motion when the scene is general, *i.e.*, unrestricted, but the motion of the camera is restricted to a (three-dimensional) rotation (an approximation is when the camera is fixed to a tripod), see for example [6,7]. In this paper, we address the estimation of the homographic mapping from a pair of uncalibrated input images.

Two very distinct approaches to homography estimation are found in the literature: *feature-based* methods estimate the homography by first matching feature points across the images; and *featureless*, *direct*, or *image-based*, methods estimate the homography parameters directly from the image intensity values. Other techniques include the use of integral projections [8] and Fourier transforms [9].

Feature-based methods, *e.g.*, [6,7,10], became popular due to the simplicity of the geometric problem of estimating the homography from the feature point correspondences. In fact, the homography parameters are linearly related to simple functions of the feature coordinates. Thus, given a set of point correspondences, the Least-Squares (LS) estimate of the homography parameters is obtained in closed-form by simply using a pseudo-inverse. The bottleneck of these methods is the feature correspondence step, which is particularly hard in several practical situations, *e.g.*, when processing low textured images [11,12].

Featureless methods, *e.g.*, [4,13,14], avoid pre-processing steps by attempting to estimate the homography directly from the images. Naturally, these methods lead to robust estimates. However, since the homography parameters are related to the image intensities in an highly nonlinear way, featureless methods use complex and time-consuming optimization algorithms, *e.g.*, Gauss-Newton, gradient descent. See [15,16] for a discussion on the feature-based/featureless dichotomy.

1.2 Proposed Approach

Our method splits the first image into four blocks (quadrants) and deals with each one as if it was a pointwise feature, *i.e.*, it determines the displacement of each quadrant by using standard correlation techniques. Then, it computes the homography described by these four displacements and registers the first image according to this homography. The registered image is naturally closer to the second image. The procedure is repeated (and the computed homographies are successively composed) until the displacement of each quadrant is zero, *i.e.*, until the registered image coincides with the second image.

Our approach combines the simplicity of the feature-based methods with the robustness of the featureless ones. It has the robustness of the image-based approaches because it matches the intensities in the entire image (rather than using only a subset of pointwise features). Our method is however much simpler than current featureless approaches because the nonlinear optimization is taken care of by using an iterative scheme where each iteration is as trivial as estimating the homography from feature point correspondences.

1.3 Paper Organization

In section 2, we introduce the notation needed to parameterize the image global motion in terms of an homographic mapping. Section 3 details the method we

propose in this paper to estimate the parameters of the homography describing the global motion between a pair of images. In section 4, we present experimental results that illustrate our approach and section 5 concludes the paper.

2 Global Motion Parameterization

An homography describing the global motion of the brightness pattern is characterized by an 8-parameter vector

$$\mathbf{h} = \left[a\ b\ c\ d\ e\ f\ g\ h \right]^T . \tag{1}$$

The homography parameter vector $\mathbf{h}$ relates the coordinates (x_1, y_1) of a point in image $\mathbf{I}_1$, with the coordinates (x_2, y_2) of the corresponding point in image $\mathbf{I}_2$, through

$$\begin{cases} x_2\left(\mathbf{h}; x_1, y_1\right) = \left(a x_1 + b y_1 + c\right) / \left(g x_1 + h y_1 + 1\right) \\ y_2\left(\mathbf{h}; x_1, y_1\right) = \left(d x_1 + e y_1 + f\right) / \left(g x_1 + h y_1 + 1\right) . \end{cases} \tag{2}$$

For simplicity, in projective geometry, the vector $\mathbf{h}$ is often re-arranged into a 3×3 homography matrix $\mathbf{H}$ and the equalities in (2) are written in homogeneous coordinates $\mathbf{p} = [x, y, 1]^T$:

$$\begin{bmatrix} x_2 \\ y_2 \\ 1 \end{bmatrix} \propto \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \\ 1 \end{bmatrix} \qquad \Leftrightarrow \qquad \mathbf{p}_2 \propto \mathbf{H}\mathbf{p}_1 , \tag{3}$$

where $\propto$ represents the projective space equality, *i.e.*, it denotes equal up to a scale factor [6,7].

Re-arranging (2,3), we see that the homography parameters in $\mathbf{h}$, or in $\mathbf{H}$, are linearly related to simple functions of the image coordinates. In fact, (2,3) can be written as

$$\boldsymbol{\Psi}\mathbf{h} = \boldsymbol{\psi}, \tag{4}$$

where

$$\boldsymbol{\psi} = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} , \tag{5}$$

and

$$\boldsymbol{\Psi} = \begin{bmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -x_2 x_1 & -x_2 y_1 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -y_2 x_1 & -y_2 y_1 \end{bmatrix} . \tag{6}$$

Given the correspondences of a set of N feature points, *i.e.*, given the set of pairs of coordinates $\left\{(x_1, y_1)^i, (x_2, y_2)^i, i = 1, \ldots, N\right\}$, the LS estimate of the homography parameter vector $\mathbf{h}$ is easily obtained by using the pseudo-inverse to invert the set of $2N$ linear equations like (4). Since vector $\mathbf{h}$ is 8-dimensional (1), at least 4 feature points are required to unambiguously determine the homography between the images. In this case, which is the relevant one for the algorithm

we propose in this paper, the homography describing the global motion of the 4 feature points, is simply obtained as

$$\mathbf{h} = \left[\begin{array}{c} \boldsymbol{\Psi}_1 \\ \hline \boldsymbol{\Psi}_2 \\ \hline \boldsymbol{\Psi}_3 \\ \hline \boldsymbol{\Psi}_4 \end{array}\right]^{-1}_{8\times 8} \left[\begin{array}{c} \boldsymbol{\psi}_1 \\ \hline \boldsymbol{\psi}_2 \\ \hline \boldsymbol{\psi}_3 \\ \hline \boldsymbol{\psi}_4 \end{array}\right]_{8\times 1} , \tag{7}$$

where $\boldsymbol{\Psi}_i$ and ψ_i denote matrices and vectors defined as in (4), now computed with the coordinates of each feature point i.

A reader familiar with feature-based methods may note that usually the number of feature points is much larger than the minimum 4 required and the estimate of the homography is usually computed from the singular value decomposition of a matrix that collects the observations, rather than from the generalization of (7). In opposition, we use the simple closed-form solution in (7) for 4 features, because, as it will become clear in the sequel, the robustness of our method comes from using the intensity values in the entire images, rather than from using an huge number of pointwise features.

3 Global Motion Estimation

Although the homography describing the global motion is easily estimated from point correspondences using (7), pointwise features are difficult to match from image to image in an automatic way. This inspired us to develop an algorithm that overcomes the difficulty without resorting to time-consuming nonlinear optimization, unlike current image-based, or featureless, methods. This section describes our algorithm.

3.1 Feature-Based, Featureless, or Both ?!

Rather than attempting to match hundreds of pointwise features, we use the minimum possible number, *i.e.*, four, but with a much larger dimension—each feature occupies one quadrant of the first image $\mathbf{I}_1$. This way we take into account all the intensity levels of the entire images. Our method proceeds by matching each of these four feature blocks to the second image $\mathbf{I}_2$, using standard correlation techniques. Naturally, this results in a very rough matching, unless the global motion between $\mathbf{I}_1$ and $\mathbf{I}_2$ is a pure translation. Then, using expression (7) with the coordinates of the centers of the feature blocks[1], we obtain a first estimate $\widehat{\mathbf{h}}_1$ of the homography describing the global motion between images $\mathbf{I}_1$ and $\mathbf{I}_2$.

Now, apply the homographic mapping characterized by the estimate $\widehat{\mathbf{h}}_1$, to the first image $\mathbf{I}_1$. If $\widehat{\mathbf{h}}_1$ was an accurate estimate, this would align $\mathbf{I}_1$ with $\mathbf{I}_2$.

[1] Due to the chosen location for the feature blocks, the matrix inversion in (7) is well conditioned. It is singular only in degenerate situations, *e.g.*, when the centers of three blocks become colinear or the centers of two blocks collapse into a point.

Since in general $\widehat{\mathbf{h}}_1$ will be a rough estimate, applying this homography to $\mathbf{I}_1$ will generate an image $\mathbf{I}'$ that is just "closer" to being aligned with $\mathbf{I}_2$. This is the fact exploited by our algorithm—we now use $\mathbf{I}'$ as the first input image and proceed again as just described, now to estimate $\mathbf{h}'$, the homography describing the motion between $\mathbf{I}'$ and $\mathbf{I}_2$. The estimate $\widehat{\mathbf{h}}'$ will be more accurate than $\widehat{\mathbf{h}}_1$ because the corresponding input images are closer to being aligned, thus the block features (the image quadrants) will be better matched.

To obtain the second estimate $\widehat{\mathbf{h}}_2$ of the homography between the original image $\mathbf{I}_1$ and $\mathbf{I}_2$, we just need to combine the first estimate $\widehat{\mathbf{h}}_1$ with the update $\widehat{\mathbf{h}}'$. From (3), we see that this operation is easily expressed using the homogeneous representation of the homographies in terms of the corresponding matrices $\widehat{\mathbf{H}}_2$, $\widehat{\mathbf{H}}_1$, and $\widehat{\mathbf{H}}'$:

$$\widehat{\mathbf{H}}_{i+1} \propto \widehat{\mathbf{H}}_i \widehat{\mathbf{H}}' \,, \tag{8}$$

where, at this point, $i=1$. The process is repeated until the displacements of the block features is zero (in practice, until they are below a small threshold). The update matrix $\widehat{\mathbf{H}}'$ converges to the identity $\mathbf{I}_{3\times 3}$ and the sequence of estimates $\widehat{\mathbf{H}}_i$ converges to the homography describing the global motion between images $\mathbf{I}_1$ and $\mathbf{I}_2$.

This paragraph summarizes our claims relative to the approach just described. Like the feature-based methods, our approach exploits the fact that the homography is easily estimated from a set of spacial correspondences but, unlike those, it avoids matching hundreds of pointwise features. Like the image-based methods, we match the intensity levels in the entire images, but, unlike those, we avoid complex and time-consuming optimization algorithms.

3.2 Multiresolution Processing

Naturally, attempting to match large regions, such as the image quadrants, with a simple translational motion model may lead to a very poor match when the global motion is far from a pure translation. As a consequence, in such situations, the algorithm described above may exhibit slow convergence or even get stuck at a point that does not correspond to the true solution. To speedup the convergence and better cope with global motions far from pure translations, we use a coarse-to-fine strategy.

We build a multiresolution pyramid [17] and start running the algorithm at its coarsest level. The estimate of the homography at each level is used to initialize the algorithm at the following (finer) level. The final estimate of the homography is then obtained at the finest level of the pyramid, *i.e.*, at the full resolution of the input images. The loss of detail at the coarser levels of the pyramid is what enables a fast convergence to the true solution, even with a model as simple as the pure translation for the motion of each quadrant.

3.3 Summary of the Algorithm

In short, our method computes the homography describing the global motion between images $\mathbf{I}_1$ and $\mathbf{I}_2$ through the following steps (image coordinates are normalized such that $x, y \in [0, 1]$):

1- Build multiresolution pyramids for $\mathbf{I}_1$ and $\mathbf{I}_2$. Initialize the resolution to its coarsest level $l = 0$, the iteration number $i = 0$, and the estimate of the homography $\widehat{\mathbf{H}}_0 = \mathbf{I}_{3\times 3}$.

2- Apply the current homography estimate $\widehat{\mathbf{H}}_i$ to image $\mathbf{I}_1$ at resolution l, obtaining $\mathbf{I}'$.

3- Split $\mathbf{I}'$ into its four quadrants and compute their displacements that best match image $\mathbf{I}_2$ at resolution l, using standard correlation techniques.

4- Compute the vector $\mathbf{h}'$ that describes the motion of the centers of the image quadrants, using (7).

5- Update the homography estimate $\widehat{\mathbf{H}}_{i+1}$ by composing the previous estimate $\widehat{\mathbf{H}}_i$ with $\widehat{\mathbf{H}}'$ through (8).

6- If the displacements of the quadrants are below a small threshold, increase the resolution level, $l = l + 1$ (if it was already the full resolution of the input images, then stop).

7- Increase the iteration number, $i = i + 1$, and go to **2-**.

4 Experiments

In this section, we describe experiments that illustrate the efficiency of the proposed approach.

4.1 Chess Table Images

The first experiment illustrates the fast convergence of the algorithm by comparing the images before and after the first iteration. We applied an homographic mapping to an image of a chess table, obtaining an highly distorted version of it, see the two images on the top of Fig. 1. These images were the input to our algorithm. Superimposed with them, we represent the initial (very rough) correspondences of the 4 quadrants of the image. Using these 4 correspondences, our algorithm computes the corresponding homography and registers the first input image according to it. This leads to the bottom left image. Note how closer to the second input image is the bottom left one, when compared to the first input image (top left one). The displacements of the new blocks are now much smaller—see the rectangles superimposed to the bottom images of Fig. 1. Our experience has shown that they converge to zero in a very fast way, which, in turn, leads to a fast convergence of the estimated homography between the two images.

In order to demonstrate the performance of the proposed algorithm with real video, we built several mosaics from video sequences, in a fully automatic way. Our system estimates the homographies between successive video frames and composes them to align all the images. Then, the intensity of each pixel of the mosaic is computed by averaging the intensities of the frame pixels that overlap at the corresponding position. We now illustrate with two of these mosaics.

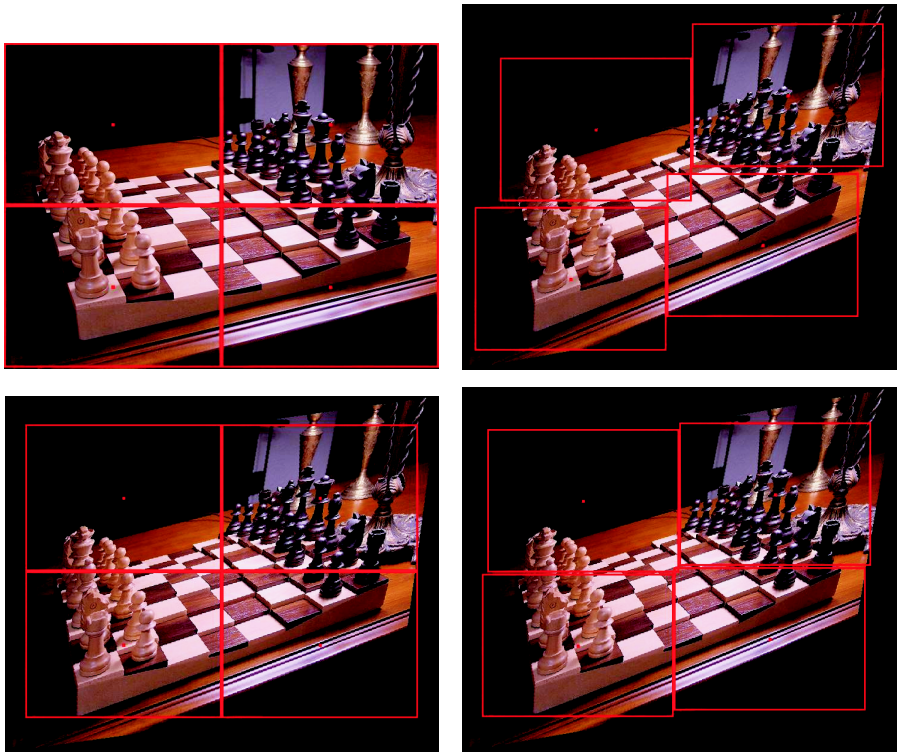

Fig. 1. Behavior of the proposed algorithm. Top line: images to register, with correspondences of the four quadrants superimposed. Bottom line: the same, after the first iteration. Notice how the left image approximates the right one.

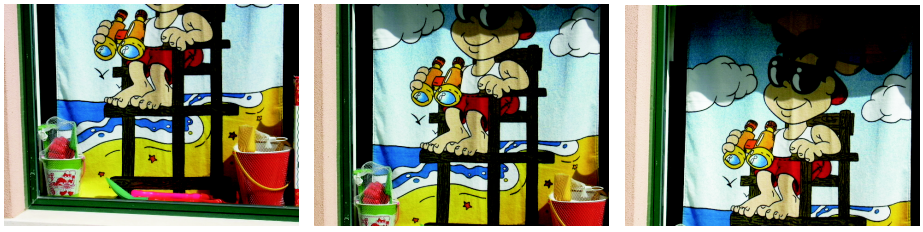

Fig. 2. Towel video sequence: three sample frames

4.2 Beach Towel Video Sequence

We used several 640×480 images showing partial views of a beach towel, see the three sample images in Fig. 2. From our experience with images of this size, it suffices to use a multiresolution pyramid with four resolution levels. The number of iterations needed to estimate each homography was very small, typically less

Fig. 3. Mosaic recovered from the towel video sequence in Fig. 2

Fig. 4. Carpet video sequence: eight sample frames

than 5. The recovered mosaic, shown in Fig. 3, illustrates that the homographies estimated by our algorithm are accurate and appropriate for this kind of application.

4.3 Carpet Video Sequence

To demonstrate the performance of our method when dealing with long video sequences, we used a 512×512 video stream with 26 frames of a carpet. Sample images are shown in Fig. 4. Although these images, obtained with an ordinary

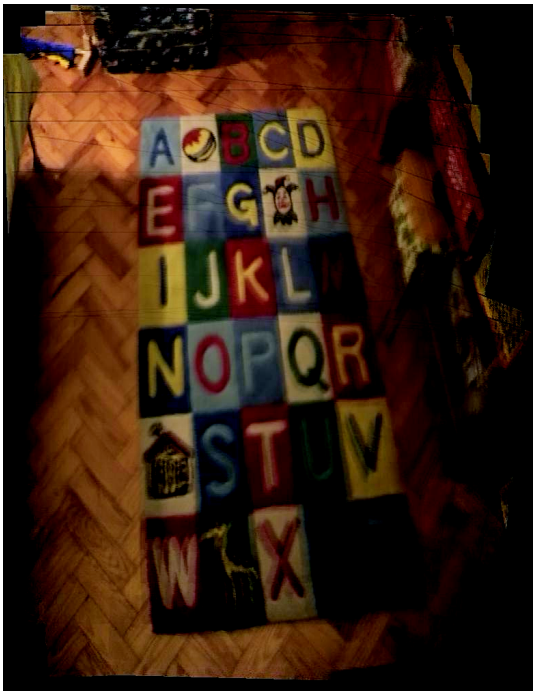

Fig. 5. Mosaic recovered from the carpet video sequence in Fig. 4

camera, are rather noisy due to video compression and fast camera movement, the recovered mosaic, shown in Fig. 5, confirms the good performance of our method. Note that feature-based methods usually fail to automatically align low contrast images, as the ones we use in this experiment, because it is very hard to compute the correspondences of pointwise features in this scenario.

5 Conclusion

We proposed a new method to estimate the homography describing the global motion between a pair of images. Our method combines the simplicity of the feature-based approaches with the robustness of the featureless ones. We illustrate the efficiency of the proposed algorithm when building, in a fully automatic way, image mosaics from uncalibrated video streams.

Acknowledgment

This work was partially supported by the (Portuguese) Foundation for Science and Technology, under grant # POSI/SRI/41561/2001.

References

1. Dufaux, F., Konrad, J.: Efficient, robust, and fast global motion estimation for video coding. IEEE Trans. on Image Processing **9**(3) (2000) 497–501
2. Petrovic, N., Jojic, N., Huang, T.: Hierarchical video clustering. In: Proc. of IEEE Multimedia Signal Processing Workshop, Siena, Italy (2004)
3. Aguiar, P., Jasinschi, R., Moura, J., Pluempitiwiriyawej, C.: Content-based image sequence representation. In Reed, T., ed.: Digital Video Processing. CRC Press (2004) 7–72 Chapter 2.
4. Mann, S., Piccard, R.: Video orbits of the projective group: a simple approach to featureless estimation of parameters. IEEE Trans. on Image Processing **6**(9) (1997) 1281–1295
5. Kim, D., Hong, K.: Fast global registration for image mosaicing. In: Proc. of IEEE Int. Conf. Image Processing, Barcelona, Spain (2003)
6. Faugeras, O.: Three-Dimensional Computer Vision. MIT Press, Cambridge, MA, USA (1993)
7. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)
8. Lee, J., Ra, J.: Block motion estimation based on selective integral projections. In: Proc. of IEEE Int. Conf. Image Processing, Rochester NY, USA (2002)
9. Reddy, B., Chattery, B.: An FFT-based technique for translation, rotation, and scale-invariant image registration. IEEE Trans. on Image Processing **5**(8) (1996) 1266–1271
10. Perez, P., Garcia, N.: Robust and accurate registration of images with unknown relative orientation and exposure. In: Proc. of IEEE Int. Conf. Image Processing, Genova, Italy (2005)
11. Shi, J., Tomasi, C.: Good features to track. In: IEEE Int. Conf. on Computer Vision and Pattern Recognition. (1994)
12. Aguiar, P., Moura, J.: Image motion estimation – convergence and error analysis. In: Proc. of IEEE Int. Conf. on Image Processing, Thessaloniki, Greece (2001)
13. Altunbasak, Y., Merserau, R., Patti, A.: A fast parametric motion estimation algorithm with illumination and lens distortion correction. IEEE Trans. on Image Processing **12**(4) (2003)
14. Pires, B., Aguiar, P.: Featureless global alignment of multiple images. In: Proc. of IEEE Int. Conf. Image Processing, Genova, Italy (2005)
15. Irani, M., Anandan, P.: About direct methods. In: Vsion Algorithms: Theory and Practice. Volume 1883 of Springer Lecture Notes in Computer Science. (1999)
16. Torr, P.: Feature based methods for structure and motion estimation. In: Vsion Algorithms: Theory and Practice. Volume 1883 of Springer LNCS. (1999)
17. Rosenfeld, A., ed.: Multiresolution Image Processing and Analysis. Volume 12 of Springer Series in Information Sciences. Springer-Verlag (1984)