All About Direct Methods

M. Irani[1] and P. Anandan[2]

[1] Dept. of Computer Science and Applied Mathematics,
The Weizmann Inst. of Science, Rehovot, Israel.
irani@wisdom.weizmann.ac.il
[2] Microsoft Research, One Microsoft Way,
Redmond, WA 98052, USA.
anandan@microsoft.com

1 Introduction

This report provides a brief summary of the review of "Direct Methods", which was presented by Michal Irani and P. Anandan.

In the present context, we define "Direct Methods" as methods for motion and/or shape estimation, which recover the unknown parameters directly from *measurable image quantities* at *each pixel* in the image. This is contrast to the "feature-based methods", which first extract a sparse set of distinct features from each image separately, and then recover and analyze their correspondences in order to determine the motion and shape. Feature-based methods minimize an error measure that is based on *distances* between a few corresponding features, while direct methods minimize an error measure that is based on direct image information collected from *all pixels* in the image (such as image brightness, or brightness-based cross-correlation, etc).

2 The Brightness Constraint

The starting point for most direct methods is the "brightness constancy constraint", namely, given two images $J(x, y)$ and $I(x, y)$,

$$J(x, y) = I(x + u(x, y), y + v(x, y)),$$

where (x, y) are pixel coordinates, and (u, v) denotes the displacement of pixel (x, y) between the two images. Assuming small (u, v), and linearizing I around (x, y), we can obtain the following well-established constraint [7]:

$$I_x u + I_y v + I_t = 0, \tag{1}$$

where (I_x, I_y) are the spatial derivatives of the image brightness, and $I_t = I - J$. All the quantities in these equations are functions of image position (x, y),

hence every pixel provides one such equation that constrains the displacement of that pixel. However, since the displacement of each pixel is defined by two quantities, u and v, the brightness constraint alone is insufficient to determine the displacement of a pixel. A second constraint is provided by a "global motion model", namely a model that describes the variation of the image motion across the entire image. These models can be broadly divided into two classes: Two-dimensional (2D) motion models and three-dimensional (3D) motion models. Below we describe how direct methods have been used in connection with these two classes of models. A more complete description of a hierarchy of different motion models can be found in [1].

3 2D Global Motion Models

The 2D motion models use a single global 2D parametric transformation to define the displacement of every pixel contained in their region of support. A frequently used model is the *affine* motion model[1], which is described by the equations:

$$\begin{aligned} u(x,y) &= a_1 + a_2 x + a_3 y \\ v(x,y) &= a_4 + a_5 x + a_6 y \end{aligned} \tag{2}$$

The affine motion model is a very good approximation for the induced image motion when the camera is imaging *distant* scenes, such as in airborne video or in remote surveillance applications. Other 2D models which have been used by direct methods include the *Quadratic* motion model [1, 14], which describes the motion of a planar surface under small camera rotation, and the 2D *projective* transformation (a homography) [19], which describes the exact image motion of an arbitrary planar surface between two discrete *uncalibrated* perspective views.

The method of employing the global motion constraint is similar, regardless of the selected 2D global motion model. As an example, we briefly describe here how this is done for the affine transformation.

We can substitute the affine motion of Equation 2 into the brightness constraint in Equation 1 to obtain,

$$I_x(a_1 + a_2 x + a_3 y) + I_y(a_4 + a_5 x + a_6 y) + I_t = 0. \tag{3}$$

Thus each pixel provides one constraint on the six unknown global parameters $(a_1, \ldots, a_6)$. Since these parameters are *global* (i.e., the same parameters are shared by all the pixels), therefore, theoretically, six independent constraints

[1] The affine transformation accurately describes the motion of a an arbitrary planar surface for a fully rectified pair of cameras - i.e., when the optical axes are parallel and the baseline is strictly sideways.

from six different pixels are adequate to recover these parameters. In practice, however, the constraints from *all* the pixels within the region of analysis (could be the entire image) are combined to minimize the error:

$$E(a_1, \ldots, a_6) = \sum (I_x(a_1 + a_2 x + a_3 y) + I_y(a_4 + a_5 x + a_6 y) + I_t)^2 \quad (4)$$

Note that different pixels contribute differently to this error measure. For example, a pixel along a horizontal edge in the image will have significant I_y, but zero I_x, and hence will only constrain the estimation of the parameters (a_4, a_5, a_6) and not the others. Likewise a pixel along a vertical edge will only constrain the estimation of the parameters (a_1, a_2, a_3). On the other hand, at a corner-like pixel and within a highly textured region, both the components of the gradient will be large, and hence the pixel will constrain all the parameters of the global affine transformation. Finally, a pixel in a homogeneous area will contribute little to the error since the gradient will be very small.

In other words, the direct methods use information from all the pixels, weighting the contribution of each pixel according to the underlying image structure around that pixel. This eliminates the need for explicitly recovering distinct features. In fact, even images which contains no distinct feature points can be analyzed, as long as there is sufficient image gradient along different directions in different parts of the image.

4 Coarse-to-Fine Iterative Estimation

The basic process described above relies on *linearizing* the image brightness function (Equation 1). This linearization is a good approximation when (u, v) are small (e.g., less than one pixel). However, this is rarely satisfied in real video sequences. The scope of the direct methods has therefore been extended to handle a significantly larger range of motions via *coarse-to-fine* processing, using *iterative refinement* within a multi-resolution pyramid.

The basic observation behind coarse-to-fine estimation is that given proper filtering and subsampling, the induced image motion decreases as we go from full resolution images (fine pyramid levels) to small resolution images (coarse pyramid levels). The analysis starts at the coarsest resolution level, where the image motion is very small. The estimated global motion parameters are used to warp one image toward the other, bringing the two images closer to each other. The estimation process is then repeated between the *warped* images. Several iterations (typically 4 or 5) of warping and refinement are used to further increase the search range. After a few iterations, the parameters are propagated to the next (finer) pyramid level, and the process is repeated there. This iterative-refine estimation process is repeated and propagated all the way up to the finest resolution level, to yield the final motion parameters. A more complete description of the coarse-to-fine approach can be found in [1].

With the use of coarse-to-fine refinement, direct methods have been extended to handle image motions typically upto 10-15 percent of the image size. This range is more than adequate for handling the type of motions found in real video sequences. Direct methods are also used for aligning images taken by different cameras, whose degree of misalignment does not exceed the abovementioned range. For larger misalignments, an initial estimate is required.

5 Properties of Direct Methods

In addition to the use of constraints from all the pixels, weighted according to the information available at each pixel, direct methods have a number of properties that have made them attractive in practice. Here we note three of these: (i) high sub-pixel accuracy, and (ii) the "locking property", and (iii) dense recovery of shape in the case of 3D estimation. Properties (i) and (ii) are briefly explained in this section, while property (iii) is referred to in Section 6.

5.1 Sub-Pixel Accuracy

Since direct methods use "confidence-weighted" local constraints from every pixel in the image to estimate a few global motion parameters (typically 6 or 8), these parameters are usually estimated to very high precision. As a result, the displacement vector induced at each pixel by the global motion model is precise upto a fraction of a pixel (misalignment error is usually less than 0.1 pixel). This has led to its use in a number of practical situations including mosaicing [11, 9, 18, 17], video enhancement [11, 9], and super resolution [12], all of which require sub pixel alignment of images. Figure 1 shows an example of a mosaic constructed by aligning a long sequence of video frames using a direct method with a frame-to-frame affine motion model. Note that the alignment is seamless. Figure 2 shows an example of video enhancement. Note the improvement in the fine details in the image, such as the windows of the building. For examples of Super-Resolution using direct image alignment see [12].

5.2 Locking Property and Outlier Rejection

Direct methods can successfully estimate global motion even in the presence of multiple motions and/or outliers. Burt, et. al. [3] used a frequency-domain analysis to show that the coarse-to-fine refinement process allows direct methods to "lock-on" to a single dominant motion even when multiple motions are present. While their analysis focuses on the case of global translation, in practice, direct methods have been successful of handling outliers even for affine and other parametric motions. Irani et. al. [14] achieved further robustness by using an

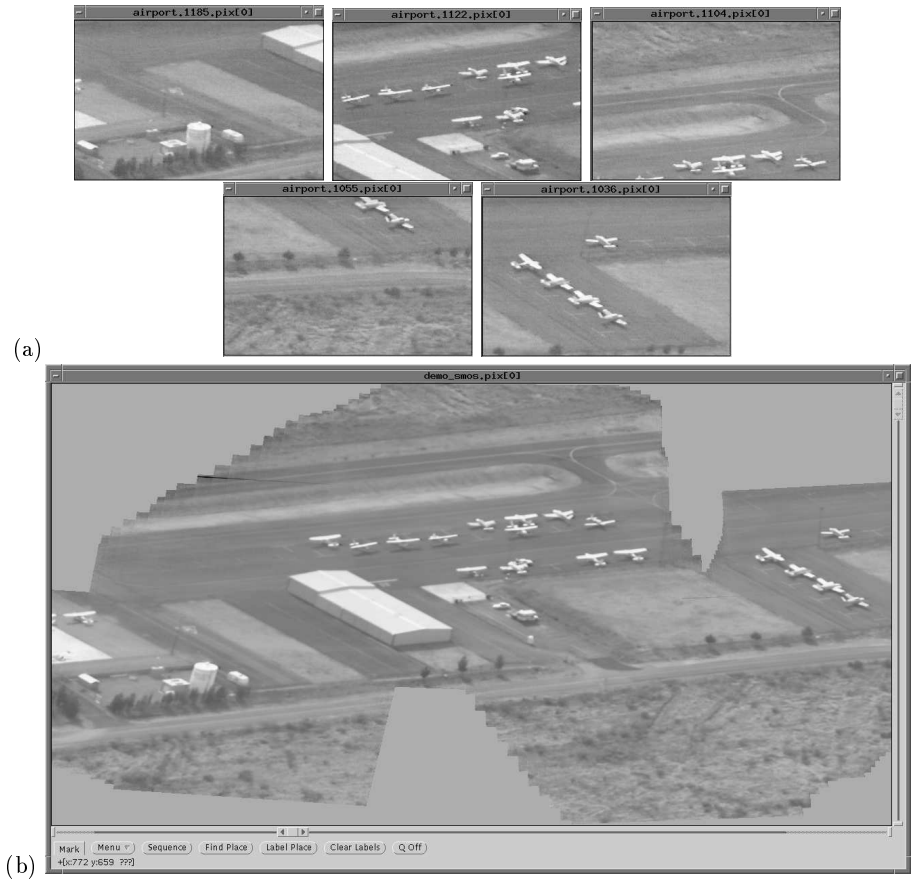

Fig. 1. Panoramic mosaic of an airport video clip. *(a) A few representative frames from a one-minute-long video clip. The video shows an airport being imaged from the air with a moving camera. (b) The mosaic image built from all frames of the input video clip. Note that the alignment is seamless.*

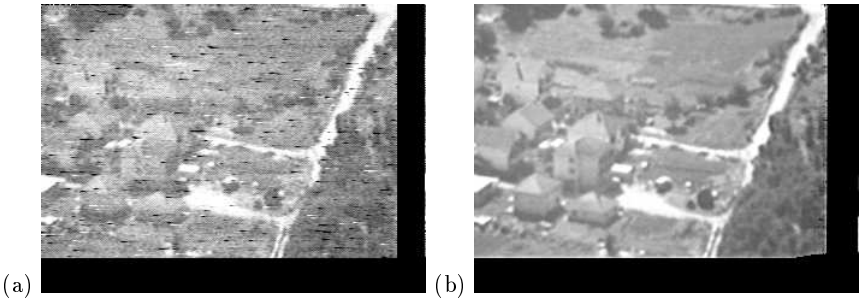

Fig. 2. Video enhancement. *(a) One out of* 20 *noisy frames (all frames are of similar quality). (b) The corresponding enhanced frame in the enhanced video sequence (all the frames in the enhanced video are of the same quality).*

iterative reweighting approach with an outlier measure that is easy to compute from image measurements. Black and Anandan [2] used M-estimators to recover the dominant global motion in the presence of outliers. Figure 3 (from [13]) shows an example of dominant motion selection, in which the second motion (a person walking across the room) occupies a significant area of the image. Other examples of dominant motion selection can be found in a number of papers in the literature (e.g., see [2, 14, 13]).

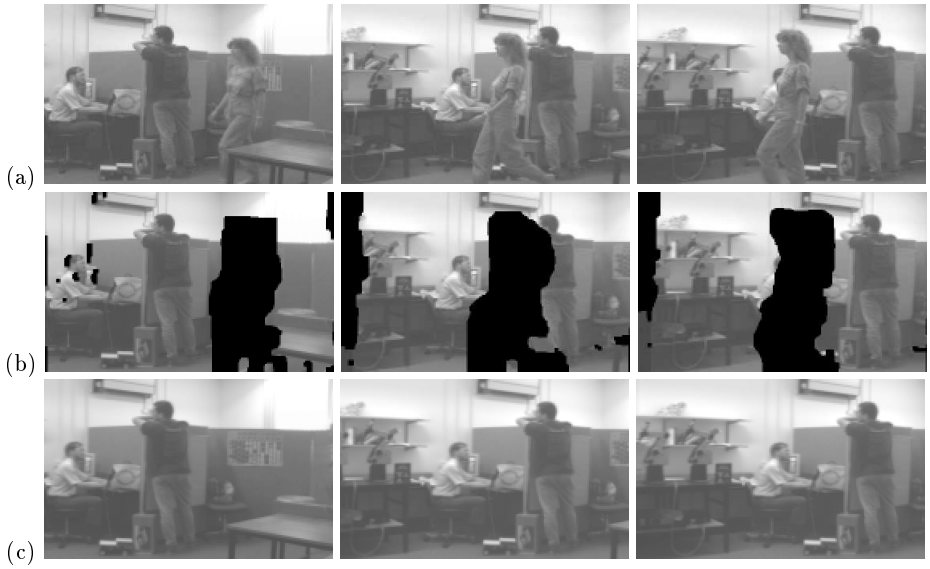

Fig. 3. Dominant motion selection and outlier rejection. *(a) 3 representative frames from the sequence. There are two motions present – that induced by the panning camera, and that induced by the walking woman. (b) Outlier pixels detected in those frame are marked in blacks. Those are pixels found to be moving inconsistently with the detected dominant motion. Those pixels correspond to the walking woman, to her reflection in the desk, to the boundaries of the image frames, and to some noisy pixels. (c) Full reconstructions of the dominant layer (the background) in all frames. The girl, her reflection, and the noise are removed from the video sequence by filling in the black regions with gray-level information from other frames according to the computed dominant background motion.*

6 3D Motion Models

So far, we have focused on using direct methods for estimating global 2D parametric motions. In these cases, a small number (typically 6 or 8) parameters

can describe the motion of every pixel in the region consistent with the global motion. However, these 2D motion models cannot model frame-to-frame motion when significant camera translation and non-planar depth variations are present. These scenarios require 3D motion models. The 3D motion models consist of two sets of parameters: a set of global parameters, which represent the effects of camera motion, and a set of local parameters (one per pixel), which represents the 3D structure or the "shape"[2]. Examples of 3D motion models include:
(i) The instantaneous velocity field model:

$$u = -xy\Omega_X + (1+x^2)\Omega_Y - y\Omega_Z + (T_X - T_Z x)/Z$$
$$v = -(1+y^2)\Omega_X + xy\Omega_Y + x\Omega_Z + (T_Y - T_Z y)/Z,$$

where $(\Omega_X, \Omega_Y, \Omega_Z)$ and (T_X, T_Y, T_Z) denote the camera rotation and translation parameters, and Z the depth value represents the local shape.
(ii) The discrete 3D motion model, parameterized in terms of a homography and the epipole:

$$u = \frac{h_1 x + h_2 y + h_3 + \gamma t_1}{h_7 x + h_8 y + h_9 + \gamma t_3} - x$$
$$v = \frac{h_4 x + h_5 y + h_6 + \gamma t_2}{h_7 x + h_8 y + h_9 + \gamma t_3} - y$$

where $(h_1, \ldots, h_9)$ denote the parameters of the homography, (t_1, t_2, t_3) represents the epipole in homogeneous coordinates, and γ represents the local shape.
(iii) The plane+parallax model:

$$u = x^w - x = \frac{\gamma}{1+\gamma t_3}(t_3 x - t_1)$$
$$v = y^w - y = \frac{\gamma}{1+\gamma t_3}(t_3 y - t_2)$$

where (x^w, y^w) correspond the image locations obtained after *warping* the image according to the induced homography (2D projective transformation) of a dominant planar surface (See [15, 10] for more details). Direct methods have been applied in conjunction with 3D motion models to simultaneously recover the global camera motion parameters and the local shape parameters from image measurements. For example, [4, 5] have used the instantaneous velocity equations to recover the camera motion and shape from two and multiple images. Szeliski and Kang [20] directly recovered the homography, the epipole, and the local shape from image intensity variations, and Kumar et. al. [15] and Irani et. al. [10] have applied direct methods using the plane+parallax model with two and multiple frames, respectively.

All of these examples of using direct methods with 3D motion models use multi-resolution coarse-to-fine estimation to handle large search ranges. The

[2] These types of 3D models are referred to as "quasi-parametric" models in [1].

computational methods are roughly similar to each other and are based on the approach described in [1] for quasi-parametric model.

Figure 4 shows an example of applying the plane+parallax model to the "block sequence" [15]. These results were obtained using the multiframe technique described in [10]. A natural outcome of using the direct approach with a 3D motion model is the recovery of a *dense* shape map of the scene, as is illustrated in Figure 4. Dense recovery is made possible because at *every* pixel the Brightness Constancy Equation 1 provides *one line constraint*, while the epipolar-constraint provides *another line constraint*. The intersection of these two line constraints uniquely defines the displacement of the pixel. Other examples of using direct methods for dense 3D shape and motion recovery can be found in the various papers cited above.

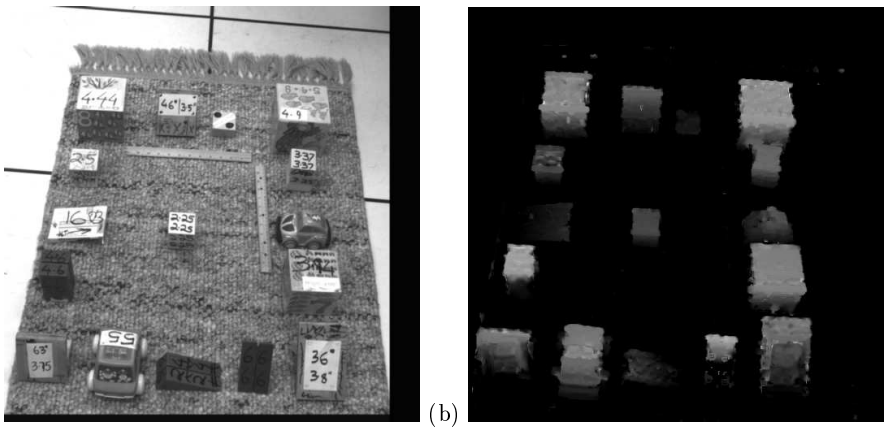

Fig. 4. Shape recovery using the Plane+Parallax model. *(a) One frame from the sequence. (b) The recovered shape (relative to the carpet plane). Brighter values correspond to taller points.*

7 Handling Changes in Brightness

Since the brightness constancy constraint is central to the direct methods, a natural question arises concerning the applicability of these techniques when the brightness of a pixel is *not* constant over multiple images. There are two ways of handling such changes. The first approach is to renormalize the image intensities to reduce the effects of such changes in brightness over time. For example, normalizing the images to remove global changes in mean and contrast often handles effects of overall lighting changes. More local variations can be handled

by using Laplacian pyramid representations and by applying local contrast normalizations to the Laplacian filtered images (see [6] for a *real-time* direct affine estimation algorithm which uses Laplacian pyramid images together with some local contrast normalization).

A second (and more recent) approach to handling brightness variation is to generalize the entire approach to use other local match measures besides the brightness error. This approach is discussed in more detail in Section 8.

8 Other Local Match Measures

Irani and Anandan [8] describe a general approach for extending direct methods to handle any user defined local match measure. In particular, instead of applying the linearization and the iterative refinement to *brightness surfaces*, the regression in [8] is applied directly to *normalized-correlation surfaces*, which are measured at *every pixel* in the image. A *global* affine transformation is sought, which *simultaneously* maximizes as many *local* correlation values as possible. This is done without prior commitment to particular local matches. The choice of local displacements is constrained on one hand by the global motion model (could be a 2D affine transformation or a 3D epipolar constraint), and on the other hand by the local correlation variations.

Irani and Anandan show that with some image pre-filtering, the *direct* correlation based approach can be applied to even extreme cases of image matching, such as *multi-sensor* image alignment. Figure 5 shows the results of applying their approach to recovering a global 2D affine transformation needed to align an infra-red (IR) image with an electro-optic (video) image. More recently, Mandelbaum, et. al. [16] have extended this approach to simultaneously recover the 3D global camera motion and the dense local shape.

9 Summary

In this paper we have briefly described the class of methods for motion estimation called direct methods. Direct methods use measurable image information, such as brightness variations or image cross-correlation measures, which is integrated from all the pixels to recover 2D or 3D information. This is in contrast to feature-based methods that rely on the correspondence of a sparse set of highly reliable image features.

Direct methods have been used to recover 2D global parametric motion models (e.g., affine transforms, quadratic transforms, or homographies), as well as 3D motion models. In the 3D case, the direct methods recover the *dense* 3D structure of the scene simultaneously with the camera motion parameters (or epipolar geometry). Direct methods have been shown to recover pixel motion

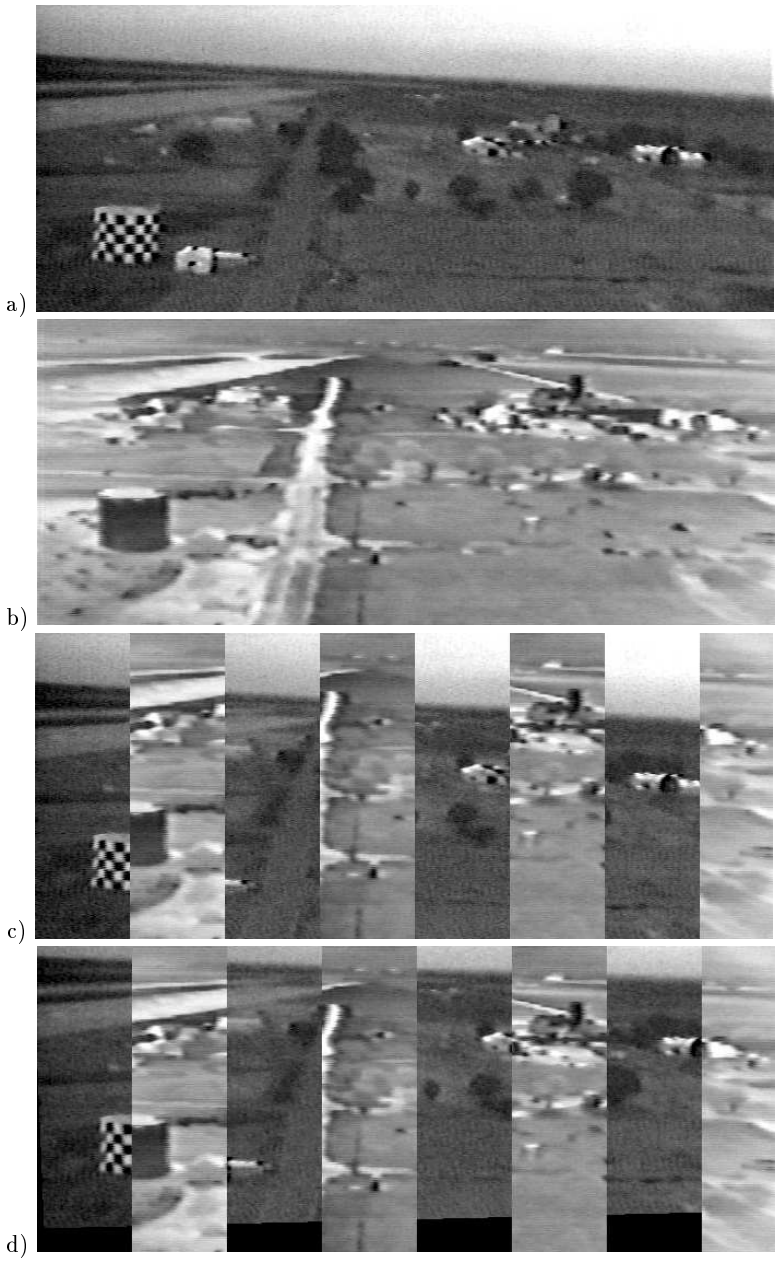

Fig. 5. Multi-sensor Alignment. *(a) EO (video) image. (b) IR (Infra-Red) image. (c) Composite (spliced) display* before *alignment. (d) Composite (spliced) display* after *alignment. Note in particular the perfect alignment of the water-tank at the bottom left of the images, the building with the arched-doorway at the right, and the roads at the top left of the images.*

upto high subpixel precision. They have also been applied to real-image sequences containing multiple motions and outliers, especially in the case of 2D motion models. The recent use of cross-correlation measures within direct methods have extended their applicability to image sequences containing significant brightness variations over time, as well as to alignment of images obtained by sensors of different sensing modalities (such as IR and video). Direct methods are capable of recovering misalignments of up to 10-15 % of the image size. For larger misalignments, an initial estimate is required.

References

1. J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *European Conference on Computer Vision*, pages 237–252, Santa Margarita Ligure, May 1992.
2. M.J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63:75–104, 1996.
3. P.J. Burt, R. Hingorani, and R.J. Kolczynski. Mechanisms for isolating component patterns in the sequential analysis of multiple motion. In *IEEE Workshop on Visual Motion*, pages 187–193, Princeton, New Jersey, October 1991.
4. K. Hanna. Direct multi-resolution estimation of ego-motion and structure from motion. In *IEEE Workshop on Visual Motion*, pages 156–162, Princeton, NJ, October 1991.
5. K. J. Hanna and N. E. Okamoto. Combining stereo and motion for direct estimation of scene structure. In *International Conference on Computer Vision*, pages 357–365, Berlin, May 1993.
6. M. Hansen, P. Anandan, K. Dana, G. van der Wal, and P. Burt. Real-time scene stabilization and mosaic construction. In *Proc. of the Workshop on Applications of Computer Vision II*, Sarasota, Fl., 1994.
7. B.K.P. Horn and B.G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.
8. M. Irani and P. Anandan. Robust multi-sensor image alignment. In *International Conference on Computer Vision*, Bombay, January 1998.
9. M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu. Efficient representations of video sequences and their application. *Signal Processing: Image Communication*, 8(4), 1996.
10. M. Irani, P. Anandan, and M. Cohen. Direct recovery of planar-parallax from multiple frames. In *Vision Algorithms 99*, Corfu, September 1999.
11. M. Irani, P. Anandan, and S. Hsu. Mosaic based representations of video sequences and their applications. In *International Conference on Computer Vision*, pages 605–611, Cambridge, MA, November 1995.
12. M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing*, 53:231–239, May 1991.
13. M. Irani and S. Peleg. Using motion analysis for image enhancement. *Journal of Visual Communication and Image Representation*, 4(4):324–335, December 1993.

14. M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12:5–16, February 1994.
15. R. Kumar, P. Anandan, and K. Hanna. Direct recovery of shape from multiple views: a parallax based approach. In *Proc 12th ICPR*, pages 685–688, 1994.
16. R.Mandelbaum, G. Salgian, and H. Sawhney. Correlation-based estimation of ego-motion and structure from motion and stereo. In *International Conference on Computer Vision*, pages 544–550, Corfu, September 1999.
17. H.S. Sawhney and S. Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18:814–830, 1996.
18. R. Szeliski. Image mosaicing for tele-reality applications. Technical Report CRL 94/2, DEC Cambridge Research Lab, May 1994.
19. R. Szeliski and J. Coughlan. Hierarchical spline-based image registration. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 194–201, June 1994.
20. R. Szeliski and S.B. Kang. Direct methods for visual scene reconstruction. In *Workshop on Representations of Visual Scenes*, 1995.