Parallax-based view synthesis from uncalibrated images

Stefano Caldrer, Sara Ceglie, Andrea Fusiello, Vittorio Murino

Università di Verona
Dipartimento di Informatica
Strada Le grazie 15, 37134 Verona, Italy
Email: {fusiello|murino}@sci.univr.it

Abstract

In this paper we present an image-based system for novel view synthesis from multiple model views. Our method works by segmenting images of a static scene in background and foreground, basing on motion parallax. From this segmentation we are able to recover the *relative affine structure*. Finally, we synthesize novel views with an original method based on step-wise replication of the epipolar geometry acquired from few model or "seed" views. The method is *uncalibrated*, for it does not need the rigid displacements in the Euclidean frame (which is unknown), and it is *automatic*, for it does not require the user to manually specify viewing parameters.

1 Introduction

Nowadays we see an increasing interest in the convergence of Computer Vision and Computer Graphics [15]. One of the most promising and fruitful topic is *Image-Based Rendering* (IBR) [17, 11]. While the traditional geometry-based systems use a 3-D model, in IBR views are generated by resampling one or more example images, using appropriate warping functions. The advantage is that photographs of real scenes can be used as a basis to create very realistic images.

The warping functions are based on the observation that certain relationships exist between the positions of pixels representing the same points in the scene observed from different viewpoints [4].

In the case of calibrated camera, algorithms based on image interpolations yield satisfactory results [22, 21]. Uncalibrated techniques, that do not assume any knowledge on the imaging device, utilize image to image constraints such as the Fundamental matrix [14], trilinear tensors [1] or the "plane+parallax" [24, 8], to reproject pixels from a small number of reference images to a given view.

Although uncalibrated point transfer algorithms are well understood, a "natural" way of specifying novel views is missing. With an uncalibrated setting, one cannot specify the position and orientation of the virtual camera in the familiar Euclidean frame, because it is not accessible[1]. This means that one have to specify some projective element, like the epipole.

In this work, we propose an automatic solution based on the replication of the epipolar geometry that links two model views, considered as an elementary displacement step. This allows the user to move the virtual camera "to the left and a little bit upward," for example.

Our method starts by segmenting images of a static scene in background and foreground, basing on motion parallax, using a statistical feature-based method for dominant motion estimation. This extends [18], where independent motion was used to segment moving objects from background. In this work, objects are static and we only exploit camera motion and parallax to recover the background. Previous works on motion segmentation using a parametric model for the dominant motion include [7, 9, 19, 12, 20].

From this segmentation we are able to recover the *relative affine structure* [24] for the foreground points, and to build a mosaic of the background. Following [2] we use two homographies to represent the epipolar geometry that links pairs of views. Finally, we synthesize novel views with an original method based on step-wise replication of the epipolar geometry acquired from few model or "seed" views.

The rest of the paper is structured as follows. In

[1] We are working in a projective frame that is linked to the Euclidean frame by an unknown projective transformation.

the next section we review some background material necessary to make the paper self-consistent. Then, we outline our method, which will be described in details in Sections 4, 5, 6, 7, and 8. In particular Section 8 deals with the method for specifying novel views. Some results are shown in Section 9, and, finally, conclusions are drawn in Section 10.

2 Background

In this section we review some background notions needed to understand the paper. A complete discussion and formulation of the "plane+parallax" theory can be found in [24, 25]. A more general reference on the geometry of multiple views is [6].

Two views of a planar set of points are related via a homography, i.e, a non-singular linear transformation of the projective plane into itself. The most general homography is represented by a non-singular 3×3 matrix H.

If $p_i \in I_1$ and $p_i^{'} \in I_2$ are projection in two different views I_1 and I_2 of the same 3-D point P_i belonging to some plane Π, we have

$$p_i^{'} \cong H_\Pi p_i \tag{1}$$

where H_Π is the homography induced by plane Π, $\cong$ means "equal up to a scale factor" and points are expressed in homogeneous coordinates.[2] The matrix H_Π has eight degrees of freedom, being defined up to a scale factor: four corresponding points in the two views define a homography.

For a general 3-D point P_i, we have

$$p_i^{'} \cong H_\Pi p_i + k_i v^{'} \tag{2}$$

where $v^{'}$ denotes the epipole in the second view, and k_i is the *relative affine structure*, which is proportional to the distance of the point P_i from the plane Π (denoted by "a" in Fig. 1).

This equation tells us that points are first transferred as if they were lying on the reference plane Π, and then their position gets corrected by a displacement $k_i v^{'}$, called *parallax*, in the direction of the epipole, with magnitude proportional to the relative affine structure. If $P_i \in \Pi$ then $k_i = 0$ and Eq. (2) reduces to Eq. (1).

[2]Points in the image plane are denoted as $p = (x_1, x_2, x_3) \cong (\frac{x_1}{x_3}, \frac{x_2}{x_3}, 1)$ with $(u, v) = (\frac{x_1}{x_3}, \frac{x_2}{x_3})$ being the corresponding Cartesian coordinates.

For Eq. (2) to hold, it must be suitably normalized, because both the homography matrix H_Π and the epipole are defined only up to a scale factor. To this end, a point p_0 is chosen and scale factors are fixed so as to satisfy:

$$p_0^{'} \cong H_\Pi p_0 + v^{'}. \tag{3}$$

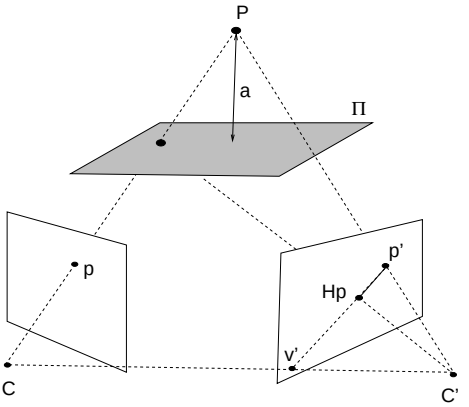

Figure 1: Relative affine structure. The segment joining p′ and Hp is the parallax for point P.

A very important property is that the relative affine structure is independent of the choice of the second view. Therefore, arbitrary "second views" can be synthesized by specifying a plane homography and the epipole. This leads to the following view synthesis algorithm [25]:

1. given two views in correspondence $p_i' \leftrightarrow p_i', \;\; i = 0 \ldots n$;
2. compute the plane homography H_Π (it is always possible, given 3 arbitrary points and the epipoles);
3. scale H_Π to satisfy $p_0' \simeq H_\Pi p_0 + v'$;
4. solve for the relative affine structure k_i in $p_i' \simeq H_\Pi p_i + k_i v'$;
5. given a new epipole v'' and a new homography H_Σ;
6. points are transferred in the third view with $p_i'' \simeq H_\Sigma p_i + v'' k_i$

Two problems are to be addressed here: i) how to compute correspondences, and ii) how to specify a new epipole v'' and a new homography H_Σ.

Relative affine structure can be used similarly for *mosaicing*. Given a sequence of images in full correspondence, one can compute, for each view, the

homography and the relative affine structure with respect to a reference view, and then warp it to the reference view using Eq. (2).

3 Outline of our method

As remarked in the previous section, view synthesis must address the problems of computing correspondences and specifying new views.

As for the first problem, we divide it in computing a dominant homography that caters for the motion of the majority of the pixels (usually the background), and a residual parallax. The latter can be in turn computed as an homography, or as a sparse field. New views are specified by replicating the epipolar geometry that links two or more model (or "seed") views.

Let us consider the case where two model views I_1 and I_2 are available; the extension to the case of more than two views is straightforward. The processing pipeline includes the following stages, each of which is described in a separate section of the paper:

1. **Dominant motion recovery**: computation of H_d, the homography that aligns the background of I_1 and I_2.
2. **Foreground segmentation**: pixels in I_1 are labeled as belonging to background or foreground.
3. **Relative affine structure recovery**: computation of k_i for each foreground pixel in I_1 wrt image I_2.
4. **View Synthesis**: transfer points from I_1 to the new image I_3 using the relative affine structure k_i and the new viewpoint.

4 Estimating the dominant motion

The homography of the background plane is obtained as the one that explains the motion of the majority of the pixels in the image: the *dominant motion*. We are here implicitly assuming that the background is approximately planar, or that its depth variation is much smaller than its average distance from the camera. We use a feature-based technique: first we extract and match corners obtaining a certain number of candidate conjugate pairs. Then we compute the homography with a robust parameter estimation technique that disregards wrong conjugate pairs (*outliers*), which are caused either by a wrong matching or by a correct matching of foreground points.

Matching is done with a weighted area-based correlation algorithm that takes into account both distance and gray-level similarity between two image windows.

The homography is computed using the Random Sample Consesus (RANSAC) algorithm of Fischler and Bolles [5], a minimal subset random sampling search technique. Rather than maximising the amount of data used to obtain an initial solution and then identifying outliers, as small a subset of the data as is feasible is used to estimate model parameters. The objective function to be maximized is the number of data points (*inliers*) having absolute residuals smaller than a predefined value. By virtue of this prespecified inlier band, RANSAC can find structures formed by substantially fewer than half the data [26].

RANSAC proceeds by repeatedly constructing solutions from randomly sampled minimum subsets and evaluating them in terms of the amount of data that is consistent with the resulting model (*consensus set)*. This process is repeated enough times to ensure that, within some level of probability, at least one of the subsets will contain only inliers. Eventually, the solution with the largest consensus set is accepted.

Assuming the proportion of outliers in the data is ϵ, the number of trials m required by RANSAC to arrive at a consensus with probability γ can be estimated as [13]:

$$m = \frac{\log(1-\gamma)}{\log(1-(1-\epsilon)^p)} \tag{4}$$

where p is the size of the sampled subset.

In our case, a "data point" is a match, and the size of the minimal subset is four, as instantiating an homography needs four matching points. The error used to determine the consensus set is the distance between the point transferred by the current homography and its true conjugate.

5 Foreground segmentation

One of toughest problem to address is to separate foreground from background. By warping I_1 with the dominant homography H_d, we obtain another image I_w that (ideally) matches I_2 in those points that lie on the background plane. Therefore, the

segmentation that divides the foreground from the background can be determined by examining the difference between I_w and I_2 (change detection). This operation represents a critical phase of the acquisition process. Simple image differences does not suffices, because i) there are some residual differences not imputable to parallax, due to photometric distortion, image noise, imprecision of the homography, rounding errors, aliasing; and ii) in areas where the foreground objects have a uniform color differences are zero. To make change detection more robust we use a local window to compare the intensity distribution around the pixel, instead of just the pixel itself. Following [10] we use the *likelihood ratio* defined as

$$\lambda = \frac{\left[\frac{\sigma_1+\sigma_2}{2} + \left(\frac{\mu_1-\mu_2}{2}\right)^2\right]^2}{\sigma_1\sigma_2} \qquad (5)$$

where μ and σ denote the mean gray value and the variance for the window around the pixel. Thresholding is then applied to the value of λ at each pixel, and the resulting binary image is processed using morphological filtering to remove isolated points and to fill small holes [3]. A manual intervention may be required to trim the parameters (namely: window size, threshold, number of iteration of morphological operators), especially when uniform regions are present.

6 Epipole recovery

The epipole can be computed from the Fundamental matrix [16], from the optical flow [27], from the virtual parallax [3] and in many other ways. Our system uses the relationship that arises from two homographies (see for example [6]).

Let H_d be the homography of dominant plane and suppose to find another homography H_s mapping 4 coplanar points on the foreground. As any homography maps one epipole to the other[4], the epipoles satisfy: $v_1 \cong \mathrm{H_d} v_2$ and $v_1 \cong \mathrm{H_s} v_2$, hence

$$v_2 \cong \mathrm{H_d^{-1}} v_1, \qquad (6)$$

and

$$v_1 \cong \mathrm{H_s H_d^{-1}} v_1. \qquad (7)$$

[3] We used the MATLAB "clean" and "fill" morphological operators.

[4] The line passing through the optical centers of the two cameras will intersect any plane of the projective space in two points, whose projection in the conjugate image plane are the epipoles.

The matrix $\mathrm{H}_\sigma = \mathrm{H_s H_d^{-1}}$ has three eigenvectors: two of them represent the line of intersection of the plane that induces $\mathrm{H_d}$ with the plane that induces $\mathrm{H_s}$, the third is the epipole. Its associated eigenvalue is distinct from the other two, that are equal.

7 Relative affine structure recovery

In order to find the relative affine structure we distinguish two different cases. The first is when foreground is (approximately) planar. In this case we fit a homography to the foreground points as we did for the background (Sec. 4).

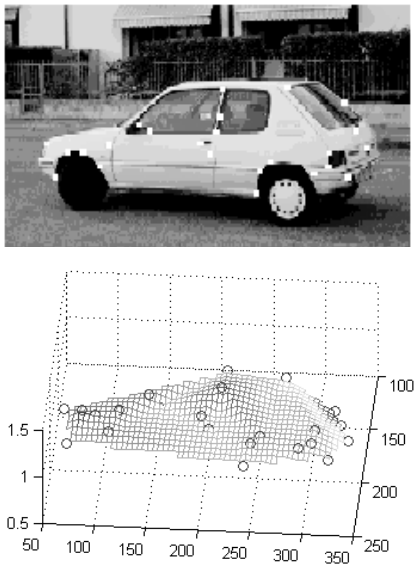

Figure 2: Example relative affine structure reconstruction. Top: image with with foreground features highlighted. Bottom: surface interpolating the values of the relative affine structure at the feature points.

When, on the contrary, the foreground is a free form surface, we recover the relative affine structure of as many points as possible and then interpolate. In order to match foreground points, we first warp I_1 with H_s, the homography that maps four foreground points (used for computing the epipole), obtaining I_1'. Then we extract and match corners in the foreground region of I_1' and I_2 as we did in Sec. 4. Matching now is easier because it is restricted to the foreground region and, in that region, I_1' is closer to

I_2 than I_1. Indeed, I_1' would perfectly overlap I_2 in the foreground area, if the foreground was planar.

Finally we compute the value of k_i for each conjugate pair $(p_i; p_i^{'})$:

$$k_i = \frac{(H_d p_i \times p_i^{'})^T (p_i^{'} \times v^{'})}{\| p_i^{'} \times v^{'} \|^2}. \qquad (8)$$

Equation (8) can be derived from Eq. (2), given that $p_i, p_i^{'}, v^{'}$ are collinear, since they belong to the same epipolar line.

Once k is computed we obtain a surface by fitting the set of k with a suitable function (see Fig. 2).

8 The synthesis

Having extracted from the model images all the information that are required, we can now use the synthesis equation

$$p_i'' \simeq H_\Sigma p_i + v'' k_i \qquad (9)$$

to construct a synthetic view, but first we need to specify H_Σ and v'', that are projective elements.

As we pointed out in the Introduction, for a view synthesis technique to be useful and usable, there must be a natural way of specifying new view points, and entering the epipole and the homography is certainly not what one would define to be "natural." In this work we propose a solution based on the replication of the epipolar geometry that links the model views, considered as an elementary displacement step. The user just need to specify, in a graphical way, the direction toward which the virtual camera must move, and the system computes automatically the required epipole and homography.

It is worth noting here that two homographies can be used in the definition of a novel viewpoint: one homography is used as H_Σ in Eq. (9) and both are used to compute the epipole v'' as in Sec. 6.

The idea is based on the following observation: suppose there are two planes, which induce homographies H_d and H_s from image I_1 to I_2. Let us synthesize a novel view I_3 in such a way that the same two planes induce homographies H_d^{-1} and H_s^{-1} respectively from I_1 to I_3. Then the image pair (I_3, I_1) is related by the same epipolar geometry than the image pair (I_1, I_2). This follows from the fact that two homographies completely determine the Fundamental matrix [6, 2], that encodes the epipolar geometry.

Please note also that the observation above is equivalent to say that the rigid displacement between view point 3 and 1 is the same as the rigid displacement G between view point 1 and 2. In other words, with reference to view point 1, view point 3 is given by G^{-1}. Therefore, if one uses the homography pair H_s, H_d he/she obtains a new viewpoint displaced by G from the original, otherwise, if one uses H_s^{-1}, H_d^{-1}, he/she obtains a new viewpoint displaced by G^{-1} from the original. When we move away from the reference views, homographies must be composed accordingly.

This allows one to replicate the unknown rigid displacement of two model views, using it as the atomic step in the definition of new view points for the view synthesis.

With only two views we can generate an arbitrary number of synthetic views, that extrapolate in discrete steps the basic displacement along two opposite directions (Fig. 3).

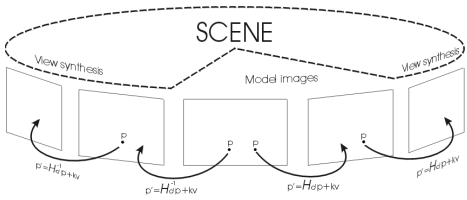

Figure 3: Extrapolation of the displacement between two model views.

Each pair of model images determines a direction along which view extrapolation can take place. With at least three model images, we can obtain synthetic images from above, below, left, right and combination of them, as shown in Fig. 4.

This technique can be extended to image interpolation. What we need in order to define a novel view is an homography pair. We can obtain it by "interpolating homographies". There is no way, to the best of our knowledge, to do this analytically [5]. Therefore we resort to a technique based on disparity interpolation. We take a number of points p_i on a regular grid. We transform them with the homography, thereby obtaining a list of conjugate points p_i'. Then we compute $p_i'' = tp_i + (1-t)p_i'$ for some

[5] The straightforward interpolation of the matrices entries does not work, because when interpolating two matrices which have singular values with different sign, one can obtain a singular matrix, which is not an homography.

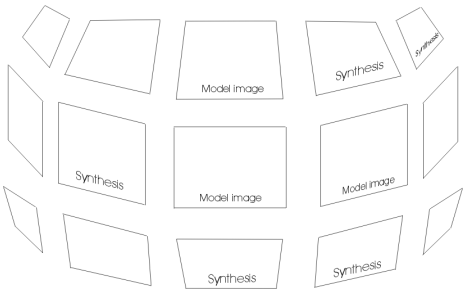

Figure 4: Three model images allows to span the viewing sphere.

value of $t \in [0, 1]$. Finally, we compute the new homography from the correspondences $p_i \leftrightarrow p_i''$.

Image warping was performed using destination scan and bilinear interpolation for background and planar foreground, and source scan and pixel "splatting" for free-form foreground. The splat size is fixed, even if it could have been made dependent on the pixel parallax [23]. Pixels with larger k (relative affine structure) overwrites pixels with smaller values.

9 Results

Several tests have been done using 640×480 gray scale images, taken with a digital camera. Figures 5 and 6 show some examples of view synthesis. The bottom rows of each figure are synthetic images obtained with our technique, using a plane for describing the foreground.

In Fig. 5 is probably better evident the fact that the novel viewpoints are extrapolations of the model ones along the directions given by the "seed" displacements between the model viewpoints.

In Fig. 6 one can appraise the fact that a mosaic of the background is used in the synthesis (indeed the shape of the background in the synthetic images is not a parallelogram).

Unlike Fig. 5, there are also black pixels in the synthetic images of Fig. 6, owing to the smaller baseline that did not allowed to "see" every point of the scene,

More examples can be found on the web at http://vips.sci.univr.it/˜fusiello/demo/synth.

10 Conclusion

An image-based method for synthesize novel images was introduced. Camera calibration is not required and no knowledge of its motion is needed. The position of the virtual camera is specified in an uncalibrated manner by replicating the epipolar geometry that links the model views, considered as an elementary displacement step. The virtual viewpoint is not constrained to lie in between the positions of the real cameras.

Our method starts by segmenting images of a static scene in background and foreground, basing on motion parallax, using a statistical feature-based method for dominant motion estimation. From this segmentation we recover the relative affine structure and, finally, we synthesize the novel views.

The method works well for texturized scenes, where a dominant background plane exist.

Future work will address the issue of homography interpolation (is the method based on disparity interpolation physically valid? is there an analytical form?), will improve the segmentation and consider more experiments with free-form surfaces instead of planes.

References

[1] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1034–1040, 1997.

[2] B. S. Boufama. The use of homographies for view synthesis. In *Proceedings of the International Conference on Pattern Recognition*, pages 563–566, 2000.

[3] B.S. Boufama and R. Mohr. Epipole and fundamental matrix estimation using virtual parallax. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1030–1036, 1995.

[4] O. D. Faugeras and L. Robert. What can two images tell us about a third one? In *Proceedings of the European Conference on Computer Vision*, pages 485–492, Stockholm, 1994.

[5] M. A. Fischler and R. C. Bolles. Random Sample Consensus: a paradigm model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, June 1981.

Figure 5: Model images (top row) and synthetic images (bottom row).

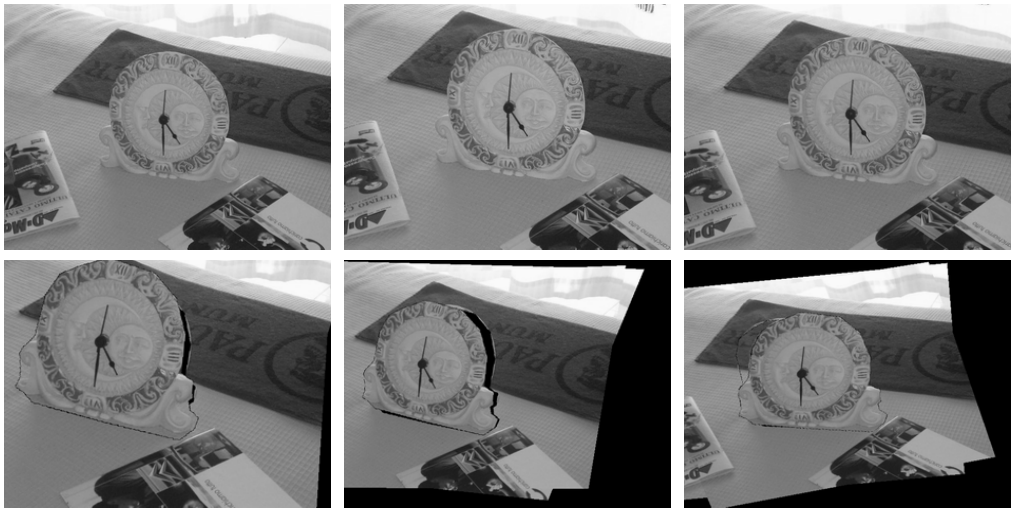

Figure 6: Model images (top row) and synthetic images (bottom row).

[6] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000.

[7] M. Irani, P. Anandan, J. Bergen, R. Kumar, and S Hsu. Efficient representations of video sequences and their applications. *Signal processing: Image Communication*, 8(4):327–351, May 1996.

[8] M. Irani, P. Anandan, and D. Weinshall. From reference frames to reference planes: Multi-view parallax geometry and applications. In *Proceedings of the European Conference on Computer Vision*, University of Freiburg, Germany, 1998.

[9] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1):5–16, February 1994.

[10] R. Jain, R. Kasturi, and B.G. Schunk. *Machine Vision*. Computer Science Series. McGraw-Hill International Editions, 1995.

[11] Sing Bing Kang. A survey of image-based rendering techniques. Technical Report CRL 97/4, Digital Cambridge Research Laboratories, August 1997.

[12] R. Kumar, P. Anandan, M. Irani, J. Bergen, and K. Hanna. Representation of scenes from collections of images. In *IEEE Workshop on Representations of Visual Scenes*, Cambridge, MA, June 1995.

[13] A. J. Lacey, N. Pinitkarn, and N. A. Thacker. An evaluation of the performance of ransac algorithms for stereo camera calibration. In *British Machine Vision Conference*, 2000.

[14] S. Laveau and O. Faugeras. 3-D scene representation as a collection of images and foundamental matrices. Technical Report 2205, INRIA, Institut National de Recherche en Informatique et an Automatique, February 1994.

[15] Jed Lengyel. The convergence of graphics and vision. *IEEE Computer*, 31(7):46–53, July 1998.

[16] Q.-T. Luong and O. D. Faugeras. The fundamental matrix: Theory, algorithms, and stability analysis. *International Journal of Computer Vision*, 17:43–75, 1996.

[17] Leonard McMillan and Gary Bishop. Plenoptic modeling: An image-based rendering system. In *SIGGRAPH 95 Conference Proceedings*, pages 39–46, August 1995.

[18] F. Odone, A. Fusiello, and E. Trucco. Layered representation of a video shot with mosaicing. *Pattern Analysis and Applications*, 5(3), August 2002. To appear.

[19] H. Sawhney and S. Ayer. Compact representations of videos through dominant and multiple motion estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):814–830, August 1996.

[20] H. S. Sawhney, S. Ayer, and M. Gorkani. Model-based 2-D 3-D dominant motion estimation for mosaicing and video representation. In *Proceedings of the IEEE International Conference on Computer Vision*, Cambridge, 1994.

[21] Daniel Scharstein. Stero vision for view synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 852–858, 1996.

[22] Steven M. Seitz and Charles R. Dyer. View morphing: Synthesizing 3D metamorphoses using image transforms. In *SIGGRAPH 96 Conference Proceedings*, pages 21–30, August 1996.

[23] J. Shade, S. Gortler, L. He, and R. Szeliski. Layered depth images. In *SIGGRAPH 98 Conference Proceedings*, 1998.

[24] A. Shashua and N. Navab. Relative affine structure: Theory and application to 3-D reconstruction from perspective views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 483–489, 1994.

[25] A. Shashua and N. Navab. Relative affine structure: Canonical model for 3D from 2D geometry and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):873–883, September 1996.

[26] C. V. Stewart. Robust parameter estimaton in computer vision. *SIAM Review*, 41(3):513–537, 1999.

[27] A. Verri and E. Trucco. Finding the epipole from uncalibrated optical flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 987–991, Bombay, India, 1998.