Proceedings of the Scandinavian Conference on Image Analysis (SCIA'95)
06–11 June 1995, Uppsala, Sweden, p. 1061-1070.

A Generic Approach for Planar Patches Stereo Reconstruction

Jean-Philippe Tarel Jean-Marc Vezien
INRIA, Domaine de Voluceau, ROCQUENCOURT
78153 LE CHESNAY Cedex FRANCE
Tel: (33) 1 39 63 54 79
Jean_Philippe.Tarel@Inria.fr Jean_Marc.Vezien@Inria.fr

Abstract

Whereas edge-based stereo has traditionally received a lot of attention, relatively few region-based 3D reconstruction techniques have been developed, despite the significant advantages such global features present for representing indoor or urban environments. We propose a generic framework for the 3D reconstruction of planar patches from stereoscopic pairs of images, and explicit a planar equation recovery scheme, which can be applied to several stereo systems configurations. Another advantage of the formulation is its possible use on various cues: we propose two different applications, using moments of inertia, non-parametric photometric analysis. In each case, the robustness of the results is assessed, using both analytic data consistency check and tests performed on synthetic stereograms. Experiments on real data are also presented.

1 Introduction

Since the introduction of early stereo technics and paradigms [1], steady efforts were made to robustly recover depth from stereo. A lot of work was dedicated to the classical goal of robot autonomous navigation, but new applications in scene analysis, such as virtual environment construction or video-conferencing, require a relatively high understanding level, in terms of where visible *surfaces* are located, as opposed to less structured information such as 1D features (e.g. segments).

In that regard, the planar surfaces assumption, which holds reasonably well in indoor scenes (as long as humans and plants are not present) has received most of the attention so far. Aloimonos [2] presented moment-based relationships between the two projections of a planar region in a calibrated stereoscopic image pair, but used them in the derivation of invariants instead of explicit 3D reconstruction. The planar assumption is often used as a constraint for other techniques, such as stereo contour reconstruction [3] or MRF field-based correlation [4]. Kristensen [5] proposes to use it in collaboration with corner and line detectors, but relies on a depth from focus process to obtain robust reconstructions. It is worth noting that most works rely on the assumption that the stereo geometry is rectified (scanlines correspond to each other in both images), a condition which may not be readily available on a given vision system.

We propose a generic framework for the 3D reconstruction of planar patches from a single stereo pair. The method assumes that region segmentation has been

performed on both images, as well as the matching between 2D primitives, with methods such as described in [6] and [7]. The key point is that once region correspondences are obtained, **no additional stereo information - e.g. local feature matching - is required**.

Our procedure relies on the assumption that there exists an affine transformation between the left and right coordinates of the perspective projections of a 3D point. In section 2, we show that such a relationship exists for a rectified stereosystem, and allows for simple computations in a transformed coordinate system that we call the *disparity space*. In the case of a general transform between images (rotation and translation), we use a paraperspective model to approximate the perspective projection, and thus derive a locally affine transformation between the coordinates of the two projections of a 3D point lying on a planar patch. In this case, one assumes that each 3D surface is shallow with respect to the scene-observer distance.

Section 3 is dedicated to using the equations of section 2 to find the surface (planar) parameters which gives the best *global* correspondence between a pair of matched 2D regions. Here we put the emphasis on avoiding any local feature matching. We propose two different approaches: the first one is solely based on the shape of the regions and uses centered inertia-moments. The second takes into account both geometry and photometry by using correlation maps computed on the regions.

As pointed put by many authors, the main problem with region-based stereo algorithms is that occlusions often bias 3D global shape recovery. More generally, segmentation is often prone to errors. Yet very few attempts have been made to detect, take into account or even quantify these effects precisely. In section 4, we propose several ways of checking the coherency of the stereo data, which enable to decide whether the reconstruction is valid or not. When using geometrical cues, affine invariants computed on the matched regions prove to be robust enough to detect unreliable data (i.e. if the observed 2D geometry is *inconsistent* with the hypothesis of the projection of a planar patch). A generalized, intensity-weighted version of these invariants can be furthermore be applied when dealing with textured regions.

2 Problem Formulation

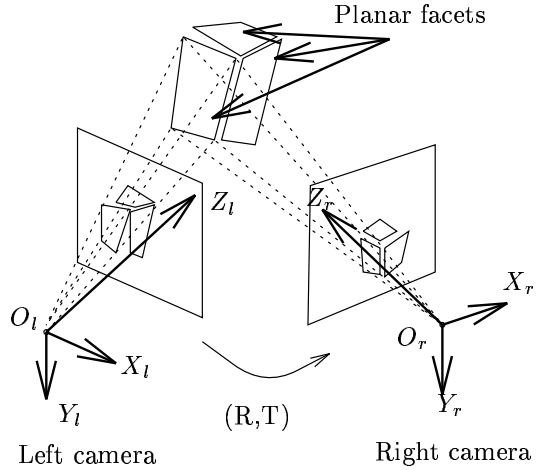


Fig. 1 - 3D stereo reconstruction of planar patches.

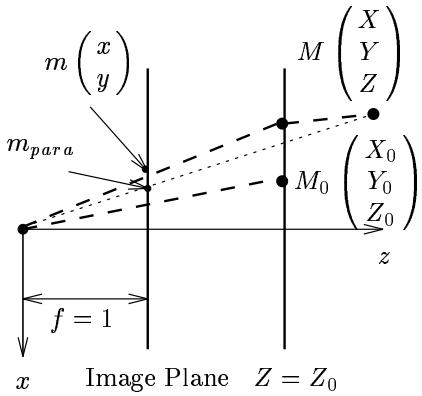


Fig. 2 - Paraperspective projection M_0 is the reference point (see text).

Let us consider the set-up of Fig. 1, where a stereo system (calibrated or not) is looking at a world made of planar patches. One can establish the equation linking

the two projections $m^l=(x^l,y^l)^t$ and $m^r=(x^r,y^r)^t$ of a given 3D point M lying on a plane $P: Z^l=p^lX^l+q^lY^l+c^l$ (or $Z^r=p^rX^r+q^rY^r+c^r$), as a function of the plane parameters. It can be easily shown that, without further constraints on the displacement (rotation R + translation T) between the cameras, the relationship between the coordinates of m^l and m^r is nonlinear and as such of little practical use.

Nevertheless, it can be drastically simplified in the case of a *rectified* stereo system, where $T=(T_x,0,0)^t$ and $R=Id$. In that case, $y^l=y^r=y$, $p^r=p^l=p$, $q^r=q^l=q$ and:

$$\begin{pmatrix} x^l \\ y^l \end{pmatrix} = \begin{pmatrix} x^r + \frac{T_x}{c^r}(1-px^r-qy) \\ y \end{pmatrix} \tag{1}$$

In other words, epipolar lines are rasterlines, and the relative displacement of the matching points along the epipolar lines is a linear function of x, y, p and q. At this point, it is convenient to map the Cartesian 3D space (X,Y,Z) into the coordinate space (x^r,x^l,y) which we call *disparity space*, and in which the 3D plane $P: Z^r=pX^r+qY^r+c^r$ is transformed into $P': ex^l+fx^r+gy+h=0$ (the mapping is homographic), with:

$$p=-\frac{e+f}{h} \qquad q=-\frac{g}{h} \qquad c^r=-T_x\frac{e}{h} \tag{2}$$

In the rectified case, it is thus **equivalent to recover planar patches equations in the disparity space or the real 3D space**.

In the general case, a local linearization of the pixel correspondence equation can still be achieved by approximating the true projection by a paraperspective model, defined by (see Fig. 2):

$$\begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{Z_0}\begin{pmatrix} X \\ Y \end{pmatrix} - \frac{Z-Z_0}{Z_0^2}\begin{pmatrix} X_0 \\ Y_0 \end{pmatrix} \tag{3}$$

where $M_0=(X_0,Y_0,Z_0)^t$ is a reference point. This model relies on the hypothesis that the projective projection $\frac{1}{Z}$ can be well approximated by the first term of its Taylor series around Z_0: in other words, the dimensions of an observed object are much smaller than its depth Z. This is sometimes refered as the *shallow object* hypothesis, which holds well for indoor scenes or satellite imagery.

Under these two conditions, the approximation holds, and one can derive the following approximation:

$$\begin{aligned} \begin{pmatrix} x^l \\ y^l \end{pmatrix} &= \frac{Z_0^r}{Z_0^l}\begin{pmatrix} (R_{11}-\frac{T_x}{Z_0^l}R_{31})-R_{33}\frac{Z_0^r}{Z_0^l}T_x\frac{p^r}{c^r} & (R_{12}-\frac{T_x}{Z_0^l}R_{32})-R_{33}\frac{Z_0^r}{Z_0^l}T_x\frac{q^r}{c^r} \\ (R_{21}-\frac{T_y}{Z_0^l}R_{31})-R_{33}\frac{Z_0^r}{Z_0^l}T_y\frac{p^r}{c^r} & (R_{22}-\frac{T_y}{Z_0^l}R_{32})-R_{33}\frac{Z_0^r}{Z_0^l}T_y\frac{q^r}{c^r} \end{pmatrix}\begin{pmatrix} x^r \\ y^r \end{pmatrix} \\ &+ \frac{Z_0^r}{Z_0^l}\begin{pmatrix} R_{13} \\ R_{23} \end{pmatrix} + \left(\frac{1}{Z_0^l}\left(2-R_{33}\frac{Z_0^r}{Z_0^l}\right)+\left(\frac{Z_0^r}{Z_0^l}\right)^2 R_{33}\left(\frac{p^r}{c^r}x_0^r+\frac{q^r}{c^r}y_0^r\right)\right)\begin{pmatrix} T_x \\ T_y \end{pmatrix} \end{aligned} \tag{4}$$

Again, we obtain an affine equation, with a linear dependence in the plane coefficients $\frac{p}{c}$ and $\frac{q}{c}$. For the choice of M_0 see next section.

Starting from equations (1) and (4), we will now derive global relationships between the two projections of a 3D planar patch in the images of a stereo pair, and use them to recover the position parameters (p,q,c).

3 3D Plane Computation

It is assumed here that two images of a stereo pair have been segmented into regions. A lot of segmentation methods exist, and we are currently using different algorithms developed by our team, based on region growing [8], region splitting [7] or energy minimization [6]. Segmented regions are subsequently matched, using epipolar constraints and image intensity based characteristics [9, 7]. The advantage of this approach comes from the fact that it is relatively easy to reliably extract and globally match extended 2D features, if the planar surfaces approximation holds, whereas obtaining exact pixel-to-pixel matching is much trickier. Therefore we will not try to use matching equations (1) or (4) at a local level, but to integrate over the whole regions. In this section, we propose three different techniques, directly derived from the planar-constrained correspondence equations, which rely on geometric and photometric cues, along with additional assumptions about the content of the scene. Results are shown in section 5.

3.1 Geometry-Based Reconstruction

Given a region R (a set of N_R connected pixels in an image), (i,j) two integers of $\mathbb{N}$, such as $i+j=n$, the inertia moment L_{ij} of order n is defined as $L_{ij}=\sum_{m(x,y)\in R} x^i y^j$. In particular $L_{00}=N_R$, and the center of inertia of R is $\vec{\mu_R}=(x_\mu,y_\mu)^t=\frac{1}{N_R}(L_{10},L_{01})^t$. We will use a more intrinsic representation, the *centered* moments: $C_{ij}=\sum_{m(x,y)\in R}(x-L_{10})^i(y-L_{01})^j$. In this section we use only moments of order 2 or lower.

• **Rectified geometry:**

Let us also define the cross-moment between the two matching regions R^l and R^r as: $\sigma_{rl}=\sum_{m(x^l,y^l)\in R^l}(x^l-x^l_\mu)(x^r-x^r_\mu)$. Starting from the disparity plane equation $ex^l+fx^r+gy+h=0$, and using the definition of C_{ij} and σ_{rl}, one easily derives the following set:

$$eC^l_{11}+fC^r_{11}+gC_{02}=0 \qquad eC^l_{20}+f\sigma_{rl}+gC^l_{11}=0 \qquad e\sigma_{rl}+fC^r_{20}+gC^r_{11}=0 \tag{5}$$

Eliminating σ_{rl} (we don't compute any pixel matching), one gets:

$$\frac{e}{f}=\pm\sqrt{\frac{{C^r_{11}}^2+C^r_{20}C_{02}}{{C^l_{11}}^2+C^l_{20}C_{02}}} \qquad g=-\frac{eC^l_{11}+fC^r_{11}}{C_{02}} \tag{6}$$

The normal vector $(e,f,g)^t$ being normalized, there are therefore two solutions, but a negative $\frac{e}{f}$ ratio will be rejected as it corresponds to the case where the two cameras look at different sides of the 3D plane (which is unlikely for real scenes).

• **General case:**

If one cannot work in a simplified geometry, a local linearization is performed through equation (4). M_0 is triangulated from matching points μ^r and μ^l, thus assuming they are the projections of the real center of gravity of the 3D region (which is a reasonable first estimate if the two cameras are not too far apart). We will denote by F_{para} the linear part of the left-to-right correspondence equations. It depends on $\frac{p}{c}$ and $\frac{q}{c}$, denoted P and Q respectively.

$$F_{para}=\begin{pmatrix}(R_{11}-\frac{T_x}{Z^l_0}R_{31})-R_{33}\frac{Z^r_0}{Z^l_0}T_xP^r & (R_{12}-\frac{T_x}{Z^l_0}R_{32})-R_{33}\frac{Z^r_0}{Z^l_0}T_xQ^r\\ (R_{21}-\frac{T_y}{Z^l_0}R_{31})-R_{33}\frac{Z^r_0}{Z^l_0}T_yP^r & (R_{22}-\frac{T_y}{Z^l_0}R_{32})-R_{33}\frac{Z^r_0}{Z^l_0}T_yQ^r\end{pmatrix}=\begin{pmatrix}A & B\\ C & D\end{pmatrix} \tag{7}$$

This time, the inertia matrix correspondence equation is:

$$\mathcal{C}^l=(\vec{m^l}-\vec{\mu^l})(\vec{m^l}-\vec{\mu^l})^t=F_{para}\mathcal{C}^rF^t_{para} \tag{8}$$

Matrices C are positive definite, and can be decomposed as $C = (U)^T \Psi^T \Psi U$, where U is a rotation and Ψ a diagonal matrix (this is obtained by standard eigenvalue decomposition). Replacing C^l and C^r in (8), one finally deduces that matrix $\mathcal{K}$, defined as $\mathcal{K} = (\Psi^l)^{-1} U^l F_{para} U^{r\,T} \Psi^r$ is necessarily a rotation, and as such verifies $\mathcal{K}_{11} = \mathcal{K}_{22}$ and $\mathcal{K}_{12} = -\mathcal{K}_{21}$. The elements $\mathcal{K}_{ij}$ are analytically obtained as first order polynomials in P^r and Q^r, so that these last two equalities form a system which solution is the best planar fit.

3.2 Correlation-Based Reconstruction

In this section, we assume that the surface considered is *stereo-lambertian*, meaning that the intensity of the projection of a 3D point in both images of the stereo system is practically the same. Provided that the region gray levels are not uniform, a correlation-based method can now be applied. As opposed to standard correlation approaches, the proposed algorithm does not operate locally but globally using all the region photometry.

Let us define the autocorrelation function of the bounded region as [10] $T(\vec{s}) = \sum_{(x,y)\in R} I(x+s_x, y+s_y) I(x,y)$. The mapping between matching points in the right and left images is affine (equation (1) with a rectified geometry and equation (4) with a paraperspective projection). Hence $\vec{s^l} = F_{para}\vec{s^r}$, where F_{para} is a function of the facet specification (p,q,c) (equation (7)). Consequently, we deduce that the dependency between the left and right autocorrelation is linear:

$$T^l(\vec{s^l}) = \sum_{R^l} I^l(\vec{u^l}) I^l(\vec{u^l} + \vec{s^l}) = \sum_{R^r} I^r(\vec{u^r}) I^r(\vec{u^r} + F_{para}\vec{s^l}) = T^r(F_{para}\vec{s^l}) \tag{9}$$

Autocorrelation images are numerically computed with a Fast Fourier Transform before normalization. As previously, centers of gravity provide the c patch position (see 3.1). The (p,q) orientation of the face is given by fitting the left autocorrelation map on the right one, by minimizing the least-square criterion:

$$\epsilon(p,q,c) = \sum_{\vec{s^l}\in C^l} \left[T^r(M\vec{s^l}) - T^l(\vec{s^l}) \right]^2 \tag{10}$$

The minimization problem is achieved using Powell's iterative algorithm. The initial estimate is given by a coarse 3D reconstruction method (or the geometric method exposed above) to ensure a fast and reliable convergence.

4 Checking data consistency

For each oh the two proposed methods of planar patch reconstruction, it is possible to test the relevance of its application on a given region pair.

4.1 Geometric Moments Invariants

In section 2, we produced affine mapping formulas between matching points in the right and left images. It follows that affine moment-based invariants exist between corresponding regions of a stereo pair [11]. We will limit ourselves to the three fundamental invariants of order no higher than 3, namely:

$$I_1 = (C_{20}C_{02} - C_{11}^2)/C_{00}^4 \tag{11}$$

$$I_2 = (C_{30}^2 C_{03}^2 - 6C_{30}C_{21}C_{03} + 4C_{30}C_{12}^3 + 4C21^3 C_{03} - 3C_{21}^2 C_{12}^2)/C_{00}^{10} \tag{12}$$

$$I_3 = (C_{20}(C_{21}C_{03} - C_{12}^2) - C_{11}(C_{30}C_{03} - C_{21}C_{12}) + C_{02}(C_{30}C_{12} - C_{21}^2))/C_{00}^7 \tag{13}$$

Consequently, before starting the reconstruction process, the computation of invariants allows to test the validity of our hypothesis, i.e. if the shape of a given match of regions is consistent with the planar patch projection hypothesis. Any violation of the hypothesis (due to segmentation errors, occlusions, or other unidentified causes) can produce a variation between the estimates of I_1, I_2 and I_3 computed on the images, and be used as a discrimination tool. To determine the efficiency of this consistency check, it is necessary to investigate the stability of moment-based invariants versus noise or viewpoint changes. For example, it is clear that the higher the

order, the more sensitive the invariant is. A systematic study using known synthetic data is presented in section 5.3.

4.2 Checking Correlation Information

We can also take into account both geometric and photometric informations at the same time, using the generalized centered moments, which are simply the inertia moments weighted by the intensity $I(x,y)$ of each pixel $m(x,y)$:

$$C'_{ij} = \sum_{m(x,y)\in R} I(x,y)(x - L_{10})^i (y - L_{01})^j \tag{14}$$

As in section 4.1, invariants are built out of the generalized moments and can be used in a similar way.

5 Experimental results

5.1 Synthetic scenes

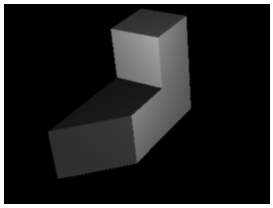
The proposed reconstruction methods are first tested with synthetic data, where both geometry and photometry are fully known. Synthetic scenes also allow to isolate errors produced by the reconstruction process, without interferences induced by the previous analysis processes (calibration, segmentation and matching).

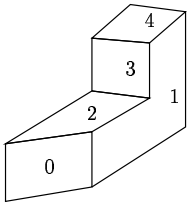
5.1.1 Paraperspective and rectified geometry

First, we verify that the paraperspective model (equation (4)) gives accurate results despite approximations. For example, for the synthetic object of Fig. 3-a. the system vergence is 10°, and the geometric reconstruction obtained in the paraperspective case is shown in Fig. 3-c.

The orientations of the facets relatively to one another are shown in Table 1, where they are compared to the real angles and to the values obtained with the rectified geometry as well.

This experiment, along with others not presented here, show that the method is more accurate in the rectified case, but the two algorithms tend to produce similar estimates if the vergence angle becomes small.





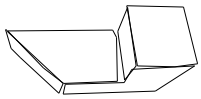


Fig. 3 : a - A synthetic object. b - Facets numbering used in table below. c - Top view of the reconstructed object using the geometry based method.

Facets pair	real	parapersp.	rectified
0–1	135.0°	132.6°	135.45°
0–2	90.0°	89.18°	89.61°
1–2	90.0°	85.55°	91.24°
1–3	90.0°	83.22°	87.79°
1–4	90.0°	85.09°	88.98°
2–3	90.0°	88.60°	94.33°
2–4	0.0°	2.96°	2.45°
3–4	90.0°	91.44°	95.20°

Table 1 - Real normal angle of pair of facets, and estimations with geometry based reconstruction in paraperspective and rectified geometry.

5.1.2 Correlation-Based reconstruction

By definition, the correlation-based method assumes the presence of a texture (for example, see Fig. 4 below). But as the geometry is also taken into account implicitely in the actual correlation computation, this method turns out to be the most accurate (but also the most computationally intensive).

Plane	Equation (p,q,c)	Error Angle
Real	(-0.004,-0.171,18.29)	
Geometric	(-0.014,-0.198,18.34)	1.6°
Correlation	(-0.023,-0.188,18.4)	1.45°

Fig. 4 - Stereo pair of a banknote and the recovered planar equations and errors.

5.2 Real Scene

Figure 5 shows the original image of a soccer ball and the results obtained with the geometry based reconstruction algorithm. Please note that the method is not disturbed by the fact the patches (hexagons and pentagons) are only approximately planar.

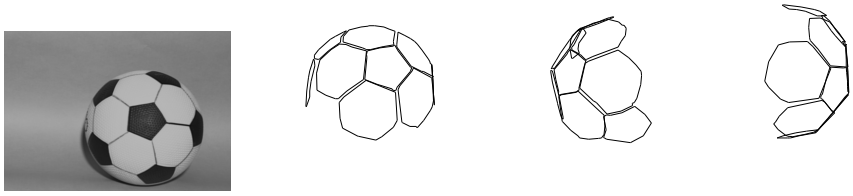


Fig. 5 - Soccer ball and planar patch based 3D reconstruction.

The experiments described above, along with others not presented here show that the general framework presented here is suited for planar patch reconstruction, and that furthermore the two proposed methods can deal with a large spectrum of experimental conditions.

5.3 Stability of the methods

We have chosen to test the robustness of our algorithms on known synthetic data, consisting of a succession of 8 stereo images of two planar patches, with increasing amounts of occlusion (Fig. 6 shows four images of the obtained image sequence).

5.3.1 Robustness to occlusion

The plot below shows the error of orientation of the occluded facet as a function of occlusion (we have plotted the corresponding occlusion-free experiment for comparison).

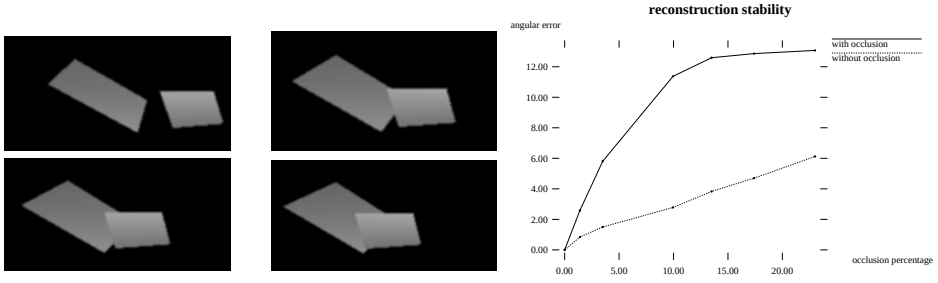
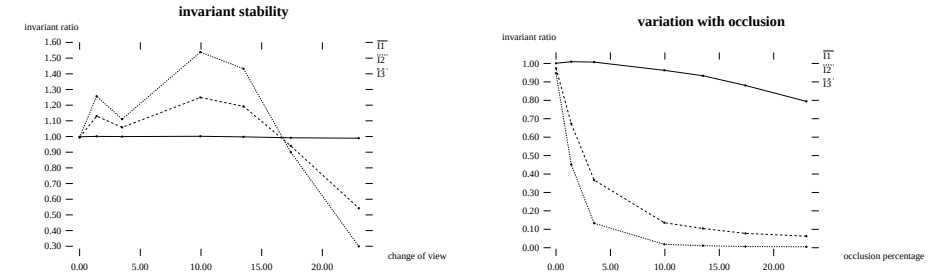


Fig. 6 - Four images of the occlusion test sequence, and the reconstruction stability results.

Experimentally, the bias produced by occlusions on the 3D patch orientation and pose is reasonable (about 12° for the data of Fig. 6), and about the same magnitude as noise and segmentation errors.

5.3.2 Invariance Checking

We use the same scene as previously to test the robustness of invariants defined in section 4.1. First, we remove the occluding plane and plot the 3 invariants in the different views of the sequence. Invariants of the first image are taken as references.



a- b-

Fig.7 - Ratio of the left and right invariants I_1, I_2 and I_3 computed on the successive viewpoints without occluding plane (a) and with occluding plane (b).

It is well known that moments become more susceptible to noise as their order increases: I_2 and I_3 clearly show insufficient stability to viewpoint changes (see Fig. 7–a). We therefore select I_1 as the more robust invariant. Fig. 7–b confirms this conclusion: I_2 and I_3 are very sensitive to big occlusion errors, whereas I_1 degrades nicely.

In this example, region couples with an invariant ratio out of the interval $[0.96, 1.04]$ corresponds to couples with more than 10% of their area occluded. Other experiments corroborate this result, and regions couples displaying more than 5 % invariant disparity should be considered unreliable. Weighted invariant I_1' holds well on textured surfaces: if one considers the two images of the banknote (Fig. 4), the relative error in the estimation in I_1' (using textural information) is 1.9 %, compared to 0.7 % if one considers I_1 (geometry only).

6 Conclusion

We proposed a generic framework for the 3D reconstruction of planar surfaces from their image projections on a stereo couple. The approach does not rely on any local feature matching and can make use of both geometric and photometric cues. It is relatively robust to segmentation errors or occlusions.

The work presented here currently relies on a region segmentation and matching process. Future extensions include the development of a segmentation-less stereo reconstruction scheme making use of similar equations as geometric and photometric constraints.

References

[1] David Marr. *Vision.* Freeman, San Francisco, 1982.

[2] John Y. Aloimonos. Perspective approximations. *Image and Vision Computing*, 8(3):179–192, August 1990.

[3] David B. Cooper, Yi-Ping Hung, and Gabriel Taubin. A new model-based stereo approach for 3d surface reconstruction using contours on the surface pattern. In *International Conference on Computer Vision*, pages 74–82, 1988.

[4] J. Subrahmonia, Y. P. Hung, and D. B. Cooper. Model-based segmentation and estimation of 3d surfaces from two or more intensity images using markov random field. In *Proceedings of the 10th International Conference on Pattern Recognition*, pages 390–397, June 1990.

[5] Steen Kristensen, Henrik Nielsen, and Henrik Christensen. Cooperative depth extraction. In *8th Scandinavian Conference on Image Analysis*, pages 321–328, 1993.

[6] A. Ackah-Miezan and A. Gagalowicz. Discrete models for energy minimizing segmentation. In *Proceedings of the 3rd International Conference on Computer Vision*, pages 200–207, 1993. 11-13 May, Berlin.

[7] S. Randriamasy and A. Gagalowicz. Region based stereo matching oriented image processing. In *Proceedings of Computer Vision and Pattern Recognition*, Maui, Hawaii, June 1991.

[8] André Gagalowicz and Olivier Monga. A new approach to image segmentation. In *Proceedings of the Eighth International Conference on Pattern Recognition*, pages 265–267, Paris, October 1986.

[9] Laurent Vinet, Peter T. Sander, Laurent Cohen, and André Gagalowicz. Hierarchical regions based stereo matching. In *Proceedings of the Sixth Scandinavian Conference on Image Analysis*, pages 71–78, Oulu, Finland, June 1989.

[10] Berthold Klaus Paul Horn. *Robot Vision.* MIT Press, Cambridge, 1986.

[11] Thomas H. Reiss. Recognizing planar objects using invariant image features. In *Lecture Notes in Computer Science.* Springer-Verlag, 1993.