A robust coarse-to-fine least-squares stereo matching for automatic terrain 3-D reconstruction

L. Alparone, F. Argenti, V. Cappellini

Dipartimento di Ingegneria Elettronica, University of Florence,
via S. Marta, 3, 50139 Florence, ITALY.
Telephone: +39-55-4796-380/372. Facsimile: +39-55-494569.
E-Mail: alparone@cosimo.ing.unifi.it

ABSTRACT

Least squares methods are widely used for digital stereo matching and 3-D reconstruction from stereo pairs. A geometric transformation is used to shift, rotate and stretch the right-image search area until fitting its left-image target area in the least square sense. However, preliminary knowledge of the disparity field between the two stereo images is critical for the performance of the algorithm. Robustness, as well as accuracy, may be enhanced when adopting a coarse-to-fine approach. In this work, a multi resolution representations (Gaussian pyramid) of the stereo images and application of the least squares matching at each pyramid layer is proposed. Convergence results to be quicker and less critical, due to the progressively decimated disparity field. Also computation is speeded up, as search areas are roughly decimated as well. Finally, the correlation coefficient is jointly employed at the full resolution level to detect mismatches caused by the presence of local minima of the residual, thus reducing the mismatch error probability.

1. INTRODUCTION

Techniques for the automatic extraction of three-dimensional measures, from digitized stereo pairs of two-dimensional pictures, have recently received a great deal of interest in several application fields of computer vision such as robotics, biomedicine, and photogrammetry and digital cartography from remotely sensed imagery (Ackermann, 1984; Claus, 1984; Pertl, 1985; Day, 1989).

The construction of a 3-D model of a scene from a stereo pair comprises three main aspects (Cappellini, 1991):

1) definition of a parametric model relative to the acquisition geometry (*calibration* of the two cameras);

2) matching of couples of corresponding points (*homologous* points) in the images of the stereo pair (*stereo-matching*);

3) determination of the 3-D coordinates of each point from the results of steps 1) and 2) (*3-D reconstruction*).

The first step requires knowledge, for each camera, of focal length, absolute position of the optical centre, and direction of the view axis. The problem is usually solved from the knowledge of the 3-D absolute positions of a certain number of points (Ground Reference Points).

For the second step, which is the object of this work, different approaches may be followed, depending on applications and on accuracy requirements. *Area-based* techniques try to estimate the point *disparity*, that is the difference between the positions of the same 3-D point when projected onto the stereo images, due to the relief of the scene, from the similarity of the areas around homologous points.

Conversely, *feature-based* techniques are founded on the matching of edges and contours of the scene, whose points are labelled as homologous. They are generally used for *close-range* vision applications, as in robotics, and whenever computational velocity is a more recommendable requirement than accuracy is.

Eventually, once steps 1) and 2) have been accomplished, the 3-D reconstruction is a simple geometric procedure: determining the intersection of two straight lines each passing through the optical centre of the respective camera and the matched point in the corresponding view; however, the accuracy of its results is very sensitive to the accuracy of both the calibration parameters and the correlation measures. Only step 2) is peculiar of digital stereo matching and requires digitization of the stereo pair;

1) and 3) are currently available in automatic form in analytic systems (Capanni, 1990).

Although originally developed for digital stereophotogrammetry (Argenti, 1990), the approach followed in this work could be successfully proposed for *close-range* stereoscopy, in which less severe precision is demanded, as an alternative to *feature-based* matching techniques, by quitting the refinement procedure at an intermediate step.

2. LEAST SQUARES STEREO MATCHING

The basic idea of the use of LS for correlation consists in finding the optimal matching of two homologous areas of the stereo pair by minimizing the sum of the squared grey-level differences (Pertl, 1985; Gruen, 1985; Rosenholm, 1987[1], 1987[2]).

As the images of the stereo pair are two different perspective views of the same scene, they differ both in geometry, due to the relief, and in radiometry, because of different exposures and relative positions of the cameras with respect to light sources. A geometric and a radiometric transformation are therefore considered between homologous areas for best matching. A complete bilinear affine transformation (six parameters accounting of rotation, scaling and shift) and a linear nonhomogeneous radiometric transformation (two parameters for luminance offset and contrast) have been used to align, rotate and stretch the right-image area until it fits the corresponding left-image area.

Let (x_l, y_l) and (x_r, y_r) be the coordinates and $g_l(x_l, y_l)$ and $g_r(x_r, y_r)$ the grey level values in the two windows of the left and right images to be matched, respectively. The geometric transformation is a bilinear nonhomogeneous affine map, that is

$$x_r = a_0 + a_1 x_l + a_2 y_l$$
$$y_r = b_0 + b_1 x_l + b_2 y_l \quad (1)$$

Also a linear nonhomogeneous transformation is used for the radiometric fitting, i.e.

$$T_R[g_r(x_r, y_r)] = h_0 + h_1 g_r(x_r, y_r) \quad (2)$$

Accounting also for the noise images are affected by yields the complete eight-parameters transformation between the two areas

$$g_l(x_l, y_l) + n_l(x_l, y_l) = T_R[g_r(x_r, y_r) + n_r(x_r, y_r)] \quad (3)$$

The residual between the original left and transformed right image window W is therefore

$$v_l(x_l, y_l) = n_l(x_l, y_l) - h_1 nr(x_r, y_r) = \\ = T_R[g_r(x_r, y_r)] - g_l(x_l, y_l) \quad (4)$$

The parameters h_0, h_1, a_0, a_1, a_2, b_0, b_1, and b_2 are to be chosen so as to minimize the overall residual within the target window

$$R_W = \sum_W [v_l(x_l, y_l)^2] \quad (5)$$

Although T_R is practically applied to the observed noisy data, as in (3), in the theoretical model it should be related to the *noise-free* radiometric data.

The problem may be solved through classical iterative algorithms based on the partial derivatives of (4) with respect to the eight parameters (Pertl, 1985). At each step a window of the original right area is transformed and resampled. In practice, the geometric transformation (1) is inverted and a sub-pixel position of the input window is found whose greylevel is retrieved by applying the radiometric transformation (2) to the bicubic interpolation of its 16 neighbour pixel greylevels.

In some applications the digital stereo pair is rectified (i.e. locally rotated and resampled) to yield horizontal epipolar lines (Rosenholm, 1987[1]). The six parameters of the geometric transformation are reduced to three, since the two parameters accounting for rotation are missing as already considered in the rectification, and only shifts along the epipolar lines are allowed, while stretches are not affected. Hence, the *on-line* computational burden results to be halved relatively to the affine map, since rectification may be done *off-line*.

3. MULTIRESOLUTION REPRESENTATION

Gaussian pyramid (GP) is a multiresolution image representation obtained by recursively lowpass filtering and down-sampling the two-dimensional data set by a factor two (Burt, 1983).

Let $\{G_0(i, j), i = 0, ..., M-1, j = 0, ..., N-1\}$, $M = p \times 2^K$ and $N = q \times 2^K$, be the input image, p and q being any pair of positive integers. Let us define as *Generalized Gaussian Pyramid*, G_0 and the set

$$G_k(i,j) = \sum_{m=-L}^{L} \sum_{n=-L}^{L} W(m,n)\, G_{k-1}(2i+m, 2j+n) \qquad (6)$$

for $i = 0, ..., (M/2^k)-1$, $j = 0, ..., (N/2^k)-1$, and $k = 1, ..., K$, where k identifies the current level of the pyramid and $K > 0$ the top level or *root*, of size p x q.

The term *Gaussian* pyramid finds an origin in Marr's vision theory (Marr, 1982), representing a whole representation of the imaged scene at different spatial resolutions, as performed by the human visual system. The attribute of *generalized* has been inserted to account that in the original definition only Gaussian-shaped kernels were used for reduction (Burt, 1981, 1983).

Purpose of the spatial filter $W(m, n)$ is introducing a low-pass effect to prevent spatial frequency aliasing from being generated by decimation. $W(m, n)$ has support regions of size $(2L + 1)$ x $(2L + 1)$ and is usually taken to be separable as product of two one-dimensional odd-sized symmetric kernels, namely

$$W(m, n) = W_y(m)W_x(n)$$

In the following it is assumed that

$$W_x(n) = W_y(m) = w(n)$$

Due to its favourable properties, the *5-taps* parametric kernel, introduced by Burt (Burt, 1981), stated in the z domain as

$$H(z) = a + 1/4(z^{-1} + z) + (1/4\text{-}a/2)(z^{-2} + z^2) \qquad (6)$$

has been intensively used for pyramids (Burt, 1983).

Considerations on the frequency response properties of pyramid generating filters, as well as on computational complexity, suggest using *half-band* filters (Meer, 1987), whose even-order taps, except zeroth, are null. However, the half-band requirement is not tight for reduction; in fact, the pyramid generating filter must achieve a tradeoff between delivering the maximum amount of spectral energy to upper levels, simultaneously reducing the spurious energy due to aliasing. Burt's kernel is not half-band except for $a = 0.5$, value which is not optimal for yielding maximum-energy reduced versions. The parametric frequency response of (6) is shown in Figure 1, for different values of a; note the absence of ripple and the parametric dependence of the passband. The value $a = 0.375$ corresponds to a Gaussian-shaped kernel (Burt, 1981). However, the optimum value of a has been experimentally found to be close to *0.6*, for a wide class of images we have dealt with, corresponding to the flattest response near the *DC*.

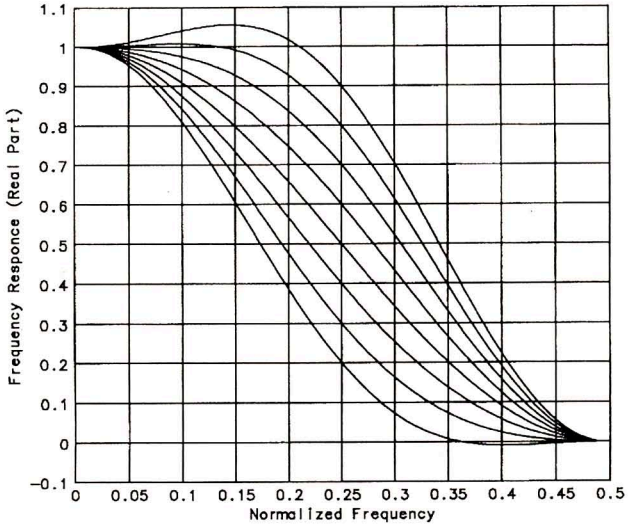

Figure 1 - Frequency response of Burt's parametric kernel; a ranges in .35 ÷ .7, .05 steps left-to-right.

4. MULTIRESOLUTION STEREO MATCHING

The main advantage of the use of GP representations of the stereo images consists in having information about the entire images condensed in the root. If δ is the distance from the initial estimate and the correct matching point at pyramid base, then it is reduced to $\delta/2^K$ at the root.

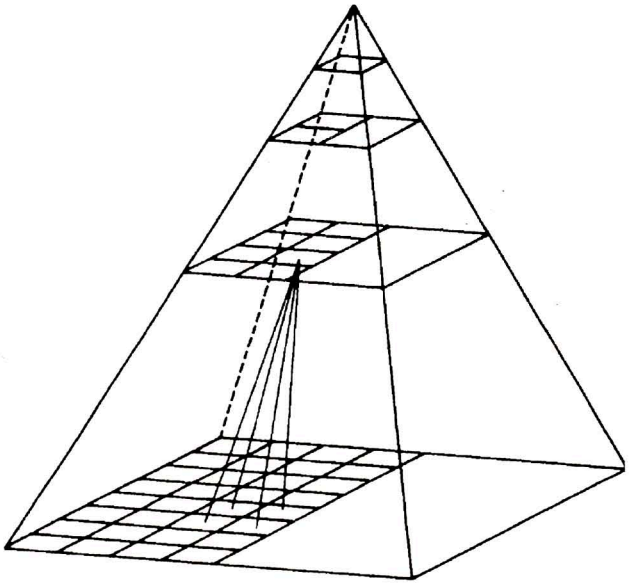

Figure 2 - Pyramid hierarchy for multiresolution LS stereo matching.

Our approach to LS stereo matching consists in applying the LS method at progressively finer levels of the GP, from the root K downward, with start point at level k given by the convergence point at level $k + 1$, rescaled by two (see Figure 2). In this way only the start point at the lowest resolution level K is to be specified, whose choice is less critical for the convergence, due to the decimated disparity field. Also the search and target areas have been progressively shrunk, although not proportionally to the decimation factor to comprise a suitable amount of significant space details, when approaching the top level of the GP. Moreover, computation cost of the multi-step scheme is lower than the classic LS, since smaller target and search areas are considered at each step.

A further advantage of the joint use of multiresolution approach and LS matching is that local (relative) minima of the square residual (5) gradually vanish, as the resolution decreases; hence the actual global (absolute) minimum is easier to be found out at lower resolution and refined afterwards, while an iterative search on a full resolution area is likely to mistake the assessment of the right match point. A similar approach have been previously applied to cross correlation for two-step template matching (Gosthasby, 1984) and to stereo matching of urban images (O'Neill, 1992).

5. EXPERIMENTAL RESULTS

Experiments to assess the robustness, efficiency and accuracy performance of the coarse-to-fine LS algorithm have been carried on from aerial stereo photographs. The calibration parameters of the two cameras has been provided apart from this work, from the terrain coordinates of reference points.

Figure *3a-b* show two overlapped fragments from a pair of stereo plates acquired by two monochrome video cameras equipped with SONY ICX 021-L CCD sensors (arrays of *h 500* x *v 582*, pixel's size *17μm* x *11μm*) and digitized at *8 b/pel* by means of one 4-buffer *512* x *512* (1 Mbyte RAM) frame grabber card. The resolution of the optic-video-grabber system corresponds to *1500 dpi* (dots-per-inch) on the photographic plates, which represent a small agricultural region with vineyard and wood. In order to produce a 3-D diagram of the scene, a regular grid of *12* x *12* points taken at a distance of 25 pixels from one other has been chosen in the left image. The search in the right image was started at level $K = 4$ of the pyramid (*32* x *32* root image) with an assumed zero disparity, i.e.

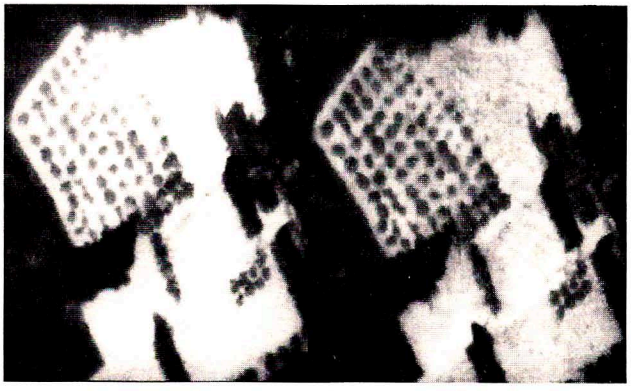

Figure 3 - Digital stereo pair: left and right.

iterations start with the target area centred in the middle point of the search area.

Figure 4 presents a 3-D diagram of the relief of the scene, achieved through a surface interpolation based on the spatial coordinates of the grid points recognized as correctly matched (*16* points over *144* have been discarded). The parallax measurement accuracy of the automatic correlation method is absolutely comparable with that of a skilled operator ($5 \div 10$ *μm*) in most of the points.

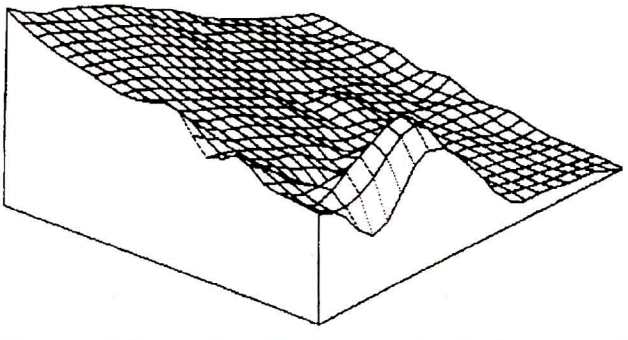

Figure 4 - 3-D reconstruction of the central part of the scene of figure 3.

Automatic recognition of gross errors, i.e. mismatch errors, is provided by the correlation coefficient (CC), that is the normalized covariance between the greylevels of the left area and of the (resampled) right area, around the pair of homologous points found out by the LS algorithm. This strategy ensures a lower percentage of mismatches than the RMS of the residuals (5) does, as it accounts for shape similarity and therefore better discriminates *absolute* minima from *relative* minima within the search area.

Figure 5 reports the plot of number of points versus CC for correctly matched and mismatched points, relatively to a sample set of 2397 points. It is manifest that CC, although

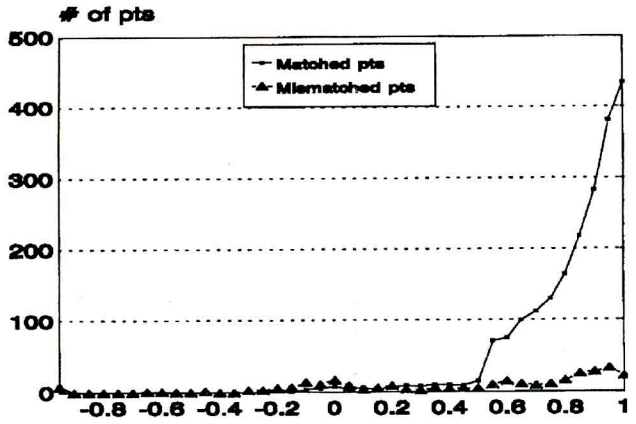

Figure 5 - Number of matched and mismatched points varying with the correlation coefficient CC between minimum MSE left and right areas.

it is not an accurate disparity measure when used for matching (Cappellini, 1991), allows *good* and *bad* points to be discriminated by thresholding it to about *0.5*.

Computing times, on a *20MHz 386PC* with a standard *floating point* co-processor were *10* seconds per point, in average, for full resolution matching of a digital stereopair, without previous rectification (Argenti, 1990). Recent advances in computer facilities now enable the proposed method to be suitable for fully automatic *on-line* stereoplotters.

6. CONCLUDING REMARKS

A robust and accurate a multi-step algorithm, based on a multiresolution representation of the digitized stereo pair and on application of a least squares correlation method at each pyramid layer, has been proposed for digital stereo matching.

The major assets of the coarse-to-fine approach to LS stereo matching are the following:

- yielding parallaxes measures at the same *sub-pixel* accuracy as for standard LS methods;
- providing a quick estimate of parallaxes that may be refined up to the requested degree of accuracy, depending on the spatial resolution;
- working also without any *a priori* knowledge of the disparity field of the stereo pair;
- expediting convergence and reduces running times, as the search area is recursively analyzed;
- performing automatic detection of gross mismatch errors through a joint use of correlation coefficient in a neighbourhood of the previously found match points.

Consequently, the above method could be proposed for production of low/medium-scale digital cartography from remotely sensed airborne or spacecraft (SPOT satellite) imagery.

ACKNOWLEDGMENTS

The authors wish to gratefully acknowledge the valuable support of Mr. G. Capanni and Mr. F. Flamigni of *GALILEO SISCAM Spa,* who provided calibration data and offered full availability of the *DIGICART 40™* stereoplotter for manually measuring reference parallaxes.

REFERENCES

Ackermann F., 1984. "Digital image correlation: Performance and potential application in photogrammetry". *Photogrammetric Record*. 11 (64): 429-439.

Argenti F. & Alparone L., 1990. "Coarse-to-fine least squares stereo matching for 3-D reconstruction". *Electr. Letters*. 26 (12): 812-813.

Burt P. J., 1981. "Fast filter transforms for image processing". *Computer Vision, Graphics, and Image Processing*. 16: 20-51.

Burt P. J. & Adelson E. H., 1983. "The Laplacian Pyramid as a Compact Image Code". *IEEE Trans. Communications*. 31 (4): 532-540.

Capanni G., Flamigni F. & Argenti F., 1990. "Digital stereo-compilation on analytical plotter Digicart 40. Principles of work, some results and practical applications". *SPIE Proceedings "Close-Range Photogrammetry meets machine vision"*. 1395: 924-931.

Cappellini V., Alparone L., Carlà R., Galli G., Langé P., Mecocci A. & Menichetti L., 1991. "Digital processing of stereo images and 3-D reconstruction techniques". *International Journal of Remote Sensing*, 12 (3): 477-490.

Claus M., 1984. "Digital terrain models through digital stereo correlation". *Photogrammetria*, 39, 183-192.

Day T. & Muller J-P., 1989. "Digital elevation model production by stereo-matching spot image-pairs: a comparison of algorithms". *Image and Vision Computing*, 7 (2): 95-101.

Gosthasby A., Gage S. H. & Bartholic J. F., 1984. "A Two-Stage Cross Correlation Approach to Template Matching". *IEEE Trans. Pattern Anal. Machine Intell.*, 6 (3): 374-378.

Gruen A., 1985. "Adaptive Least Squares Correlation – A Powerful Image Matching Technique". *South African Journal of Photogrammetry, Remote Sensing and Cartography*, 14 (3), 175-187.

Marr D., 1982. *Vision*. Freeman, New York.

Meer P., Baugher E. S. & Rosenfeld A., 1987. "Frequency Domain Analysis and Synthesis of Image Pyramid Generating Kernels". *IEEE Trans. Pattern Anal. Mach. Intell.* 9 (4): 512-522.

O'Neill M. A. & Denos M. I., 1992. "Practical approach to the stereo matching of urban imagery". *Image and Vision Computing*, 10 (2): 89-98.

Pertl A., 1985. "Digital image correlation with an analytical plotter" *Photogrammetria* 40 (1): 9-19.

Rosenholm D., 1987[1]. "Least squares matching method: some experimental results". *Photogrammetric Record.* 11 (67): 493-512.

Rosenholm D., 1987[2]. "Multi-Point Matching Using the Least Squares Techniques for Evaluation of Three Dimensional Models". *Photogrammetric Engin. and Remote Sensing.* 53 (6): 621-626.