In IEEE Computer Society Conference on Computer Vision and Pattern Recognition 1999 (CVPR'99), June 23-25, Fort Collins, Colorado, USA

Multi-scale Feature Selection in Stereo

Hiroshi Ishikawa
Department of Computer Science, Courant Institute of Mathematical Sciences
New York University, 251 Mercer Street, New York, NY 10012
ishikawa@cs.nyu.edu
http://cs.nyu.edu/phd_students/ishikawa/

Abstract

In binocular stereo matching, points in left and right images are matched according to features that characterize each point and identify pairs of points. When one tries to use multiple features, a difficult problem is which feature, or combination of features, to use. Moreover, features are difficult to cross-normalize and so comparisons must take into account not only their output, but also their distribution (*their output for different parameters*). *We present a new approach that uses geometric constraints on the matching surface to select optimal feature or combination of features from multiscale-edge and intensity features. The approach requires the cyclopean coordinate system to set mutually exclusive matching choices. To obtain the matching surface, we solve a global optimization problem on an energy functional that models occlusions, discontinuities, and inter-epipolar-line interactions.*

1. Introduction

Binocular stereo is the process of obtaining depth information from a pair of left and right images. The fundamental issues in stereo are: (i) how the geometry and calibration of the stereo system are determined, (ii) what primitives are matched between the two images, (iii) what a priori assumptions are made about the scene to determine the disparity, (iv) how the disparity map is computed, and (v) how the depth is calculated from the disparity.

Here we assume that (i) is solved, and hence the correspondence between epipolar lines (see Figure 1) in the two images are known. Answering question (v) involves determining the camera parameters, triangulation between the cameras, and an error analysis, for which we refer the reader to Faugeras [8].

Many work focused on (iii) and (iv), including Julesz [12]; Pollard, Mayhew, and Frisby [21]; Grimson [10]; Okutomi and Kanade [20]; Ayache [1]. Various algorithms, as in the cooperative stereo (Marr and Poggio [16]), have proposed a priori assumptions on the solution, including *smoothness* to bind nearby pixels and *uniqueness* to inhibit multiple matches. Occlusions and discontinuities must also be modeled to explain the geometry of the multiple-view image formation (Belhumeur and Mumford [2]; Geiger, Ladendorf, and Yuille [9]; Ishikawa and Geiger [11]). Another aspect of stereo geometry is the interdependence between epipolar lines. This topic is often neglected because of a lack of optimal algorithms. Recently, Roy and Cox [22]; followed by Ishikawa and Geiger [11]; and Boykov, Veksle, and Zabih [4] introduced network-flow algorithms that can cope with this interdependence. We used the algorithm introduced in [11] that accounts for occlusions, discontinuities, and epipolar-line interactions, in computing the optimal solution.

Our focus in this paper is (ii). In order to find corresponding points in the two images, an algorithm must have some notion of similarity, or likelihood that a pair of points correspond to each other. To estimate this likelihood, various features are used, e.g., intensity difference, edges, junctions, and correlation. Since none of these features is clearly superior to others in all circumstances, using multiple features is preferable to using single feature if one knows which feature, or what combination of features, to use when. However, features are difficult to cross-normalize. How can we compare the output

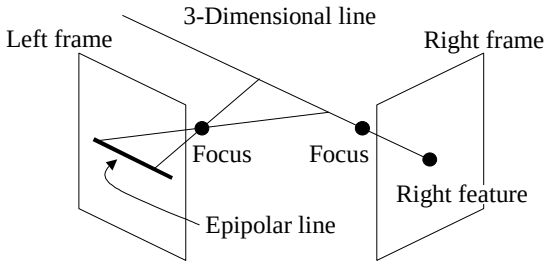

Figure 1. A pair of frames (eyes) and an epipolar line in the left frame.

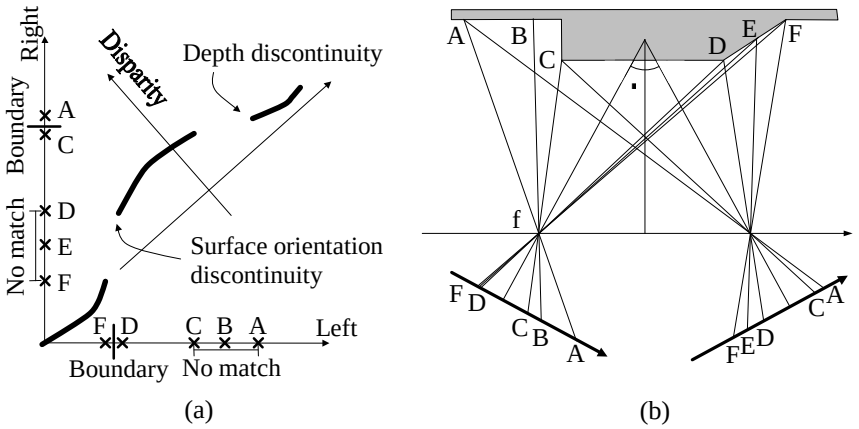

Figure 2. (a) A polyhedron (shaded area) with self-occluding regions and with a discontinuity in the surface-orientation at feature D and a depth discontinuity at feature C. (b) A diagram of left and right images (1D slice) for the image of the ramp above. Notice that occlusions always correspond to discontinuities. Dark lines indicates where match occurs.

from an edge matching with the one from correlation matching? We would like not to have to cross-normalize the output of the features, and still be able to use multiple features.

We present a new approach that uses geometric constraints for matching surface to select, for each set of mutually-exclusive matching choices, optimal feature or combination of features from multiscale-edge and intensity features.

2. Selection Set

Throughout this paper, we assume two gray-scale images $I_L(l,y)$ and $I_R(r,y)$ are given and an epipolar line appears in the images as a horizontal line $y = c$ for some constant c in both images. We suppose a local feature energy function f is defined in terms of (l,r,y), and $f(l,r,y)$ is smaller if points (l,y) and (r,y) are more likely to match. Stereo matching can be approximately seen as finding a surface embedded in this *l-r-y* space M so as to minimize the total sum of the energy under some constraints (See Figure 2).

Suppose we have k local feature energy functions $f_1, f_2, \cdots, f_k$. On what criterion should we choose from these functions? Obviously, one wants to choose the best function. However, different features are good at different situations. For instance, edges and other sparse features are good for capturing abrupt change of depth and other salient features, but can miss gradual depth change that can be captured by using dense features. However, what one *cannot* do is to choose functions at each point in *M*, since the values of different local energy functions are in general not comparable. In general, the same local function must be used over the set from which a selection is made. In other words, across these sets of selections, different functions can be used. Then, what is the set of selections in the stereo matching case? We utilize *monotonicity constraint* to answer this question. Monotonicity constraint can be stated as "the order of neighboring points from left to right does not change between images," or, equivalently, "l and r coordinate component of the tangent vector at each point that lines in a epipolar plane must have a non-negative ratio." This is not strictly true, but a reasonable approximation in many situations.

Figure 3 shows an epipolar slice of the matching space *M*. The surface that represents the matching appears as a curve here. In this figure, the monotonicity constraint means that at each point of the matching curve, the tangent vector of the curve must reside in the "light cone".

We introduce a different coordinate system, which is sometimes called the cyclopean coordinates:

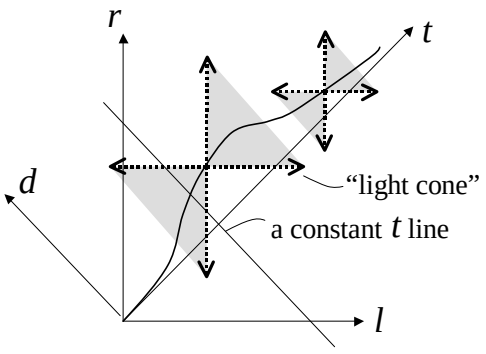

Figure 3. An epipolar slice of the matching space. The matching surface appears as a curve here. Monotonicity constraints means that this curve cross any constant t line

$$t = 1 + r,$$
$$d = 1 - r.$$

In this coordinate system, the monotonicity constraint implies that the matching curve crosses each constant-t- line at exactly one point. This means that on each of such lines the matching problem poses a selection of a point (match) from the points on this line. Thus, we can choose one particular local energy function on this line and safely choose a different one on another line. In the following, we will call these lines "selection lines."

The partition of M into selection lines is minimal in the sense that for any sub-partition the selection of the energy function cannot be local to each partition. There are, however, other minimal partitions of M with this local-selection property. For instance, M can be partitioned into other "space-like" lines with l to r tilt different from $-1{:}1$, as far as the ratio is negative.

3. Selection Rule

On each selection line, we are free to choose any local energy function. Note that the information we can easily utilize for the selection is limited. For instance, we cannot use any information concerning which matching surface is eventually selected, as that would lead to a combinatorial explosion. Here, we propose to employ a least "entropy" rule. It chooses the energy function most "sure" of the match on each selection line. After all, an energy function that does not discriminate between one match from another is of no use. In the other extreme, if we have the ground truth, an energy function that gives the true match the value of zero and positive infinity to other match is obviously the best. In other words, the energy function knows for sure which match to choose. This intuition leads us to evaluate how "sure" each energy function is.

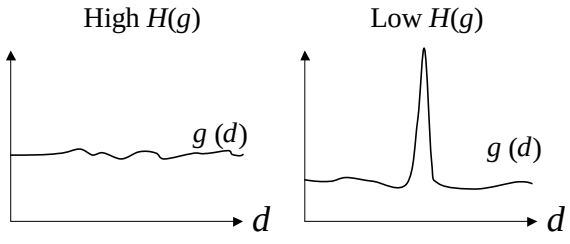

Figure 4. Functional H_t on function g. It measures the degree of concentration of the value of g.

Let us define an "entropy" functional for a positive-valued function g on $\{d = D_0, D_0 + 1, \cdots, D_1\} \times \{t\}$ as:

$$E_t(g) = \sum_{d=D_0}^{D_1} g(d,t),$$
$$H_t(g) = -\sum_{d=D_0}^{D_1} \frac{g(d,t)}{E_t(g)} \log \frac{g(d,t)}{E_t(g)}$$

This functional H_t gives a degree of concentration of the function g: it is smaller when g is more concentrated. See Figure 4. The more peaked the function is, the less value the functional gives. Now, we use this functional to choose preferable local energy function on a selection line. To use this functional for our purpose, where we need a dipped function rather than a peaked one, we invert the function and feed the result to the functional. That is, we choose the function f with least value of $H_t(f^{\max} - f)$, where $f^{\max}$ is the maximum value of f on the selection line.

$$f_t = \arg\min_i H_t\left(f_i^{\max} - f_i\right)$$

This selection rule prefers a function that has a distinguished dip, which means, in our situation, one or few disparity values have an advantage over other values. This way of selection allows avoiding irrelevant measures locally and ensures the most confident selection of the disparity on each selection line.

4. Optimization

Having selected the optimal local energy function on each selection line, we now solve a global energy functional. We used the maximum-flow-based algorithm introduced in [11]. The algorithm models occlusions and discontinuities, and uses smoothness assumption both across and along epipolar lines. In addition, it enforces monotonicity constraint while globally optimizing the energy functional, a crucial feature for our purpose. It represents the solution surface by a cut of a directed graph.

5. Experiment

We implemented the proposed algorithm using the following features:

- Intensity. This is a simple difference between the points, i.e.,
$$f_{\mathrm{I}}^2(l,r,y) = \{I_{\mathrm{L}}(l,y) - I_{\mathrm{R}}(r,y)\}^2$$

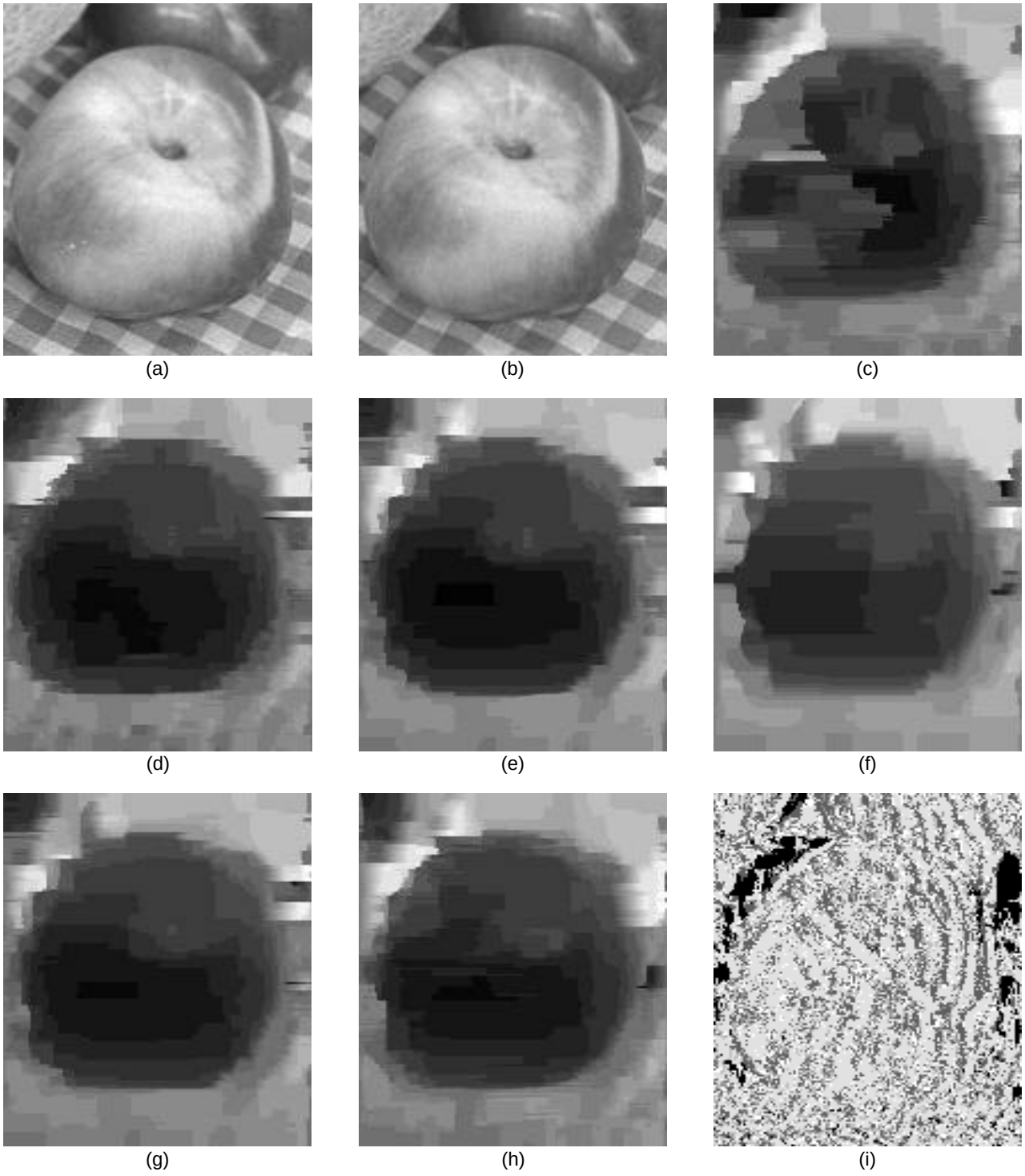

Figure 5. (a),(b) a sample image pair **Apple**, 135×172 pixel 8-bit gray-scale images. Results (disparity maps) are shown using intensity square difference f_{I}^{2} (c); wavelet edge features f_{E}^{s} with scale s = 1 (d), s = 2 (e), and s = 4 (f); multi-scale edge f_{E} (square difference of sum of wavelet coefficients for s=1,2,4) (g); and minimum-entropy selection from the five energies (h). The gray level in (i) shows which of five energy functions is used in (h) at each point. Black point represents occluded point, where no match was found, resulting in no corresponding t defined for the I-coordinate. Other gray values are in the order of (c) to (g), i.e., darkest: intensity f_{I}^{2}, lightest: multi-scale edge f_{E}.

- Wavelet edge. We used the derivative of Gaussian wavelet that detects an edge in vertical direction at various scale s:

$$f_{\mathrm{E}}^{s}(l,r,y) = \left| W_s I_{\mathrm{L}}(l,y) - W_s I_{\mathrm{R}}(r,y) \right|,$$

where

$$W_s I(x,y) = I * \psi_s(x,y),$$
$$\psi_s(x,y) = s^{-1}\psi(s^{-1}x, s^{-1}y),$$
$$\psi(x,y) = 2\pi^{-1}\exp(x^2 - y^2)x.$$

We refer the reader to Mallat [13] Chapter 6 for details on multi-scale edge detection.

- Multi-scale edges consistent across the scale. This is a measure of presence of an edge across the scale.

$$f_{\mathrm{E}}(l,r,y) = \left| \sum_s W_s I_{\mathrm{L}}(l,y) - \sum_s W_s I_{\mathrm{R}}(l,y) \right|.$$

In Figure 5, a comparison of result by these energy functions is shown. Intensity (c) gives the poorest result in this example. Wavelet edges (d), (e), and (f) for s = 1,2,4 are better, yet with black artifact on upper right, as with the multi-scale edge (g). The result where entropy minimization rule is used with these five functions is shown in (h). An illustration of which energy is used where is shown in (i).

Figure 6 shows the result stereo pair Pentagon, 508 $\times$512 pixels 8-bit gray-scale images, three wavelet coefficients for the left image, and the result. To demonstrate the quality of the result, a 3D rendering of the depth map is also shown.

6. Conclusion

In this paper, we have described a novel approach in stereo vision to select optimal feature locally, so that the chosen local energy function gives the most confident selection of the disparity from each set of mutually exclusive choices. This approach is independent of the prior model or optimization algorithm as far as the monotonicity or similar constraints are enforced.

References

[1] N. Ayache. *Artificial Vision for Mobile Robots*. MIT Press. Cambridge, Mass., 1991.

[2] P. N. Belhumeur and D. Mumford. A bayesian treatment of the stereo correspondence problem using half-occluded regions. In *Proc. Conf. on Computer Vision and Pattern Recognition*, Champaign, IL, 506–512, 1992.

[3] A. Blake and A. Zisserman. *Visual Reconstruction*. MIT Press, Cambridge, Mass., 1987.

[4] Y. Boykov, O. Veksle, and R. Zabih. "Markov random fields with efficient approximations." In *Proc. IEEE Conf. CVPR*, Santa Barbara, Calif., June 1998, 648–655.

[5] P. Burt and B. Julesz. A disparity gradient limit for binocular fusion. *Science*, 208:615–617, 1980.

[6] B. Cernushi-Frias, D. B. Cooper, Y. P. Hung, and P. Belhumeur. Towards a model-based bayesian theory for estimating and recognizing parameterized 3d objects using two or more images taken from different positions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-11:1028–1052, 1989.

[7] A. Champolle, D. Geiger, and S. Mallat. Un algorithme multi-échelle de mise encorrespondance stéréo basé sur les champs markoviens. In *13th GRETSI Conference on Signal and Image Processing*, Juan-les-Pins, France, Sept. 1991.

[8] O. Faugeras *Three-Dimensional Computer Vision*. MIT Press. Cambridge, Mass., 1993.

[9] D. Geiger and B. Ladendorf and A. Yuille Occlusions and binocular stereo. *International Journal of Computer Vision*, 14, 211-226 March, 1995.

[10] W. E. L. Grimson. *From Images to Surfaces*. MIT Press. Cambridge, Mass., 1981.

[11] H. Ishikawa and D. Geiger. "Occlusions, discontinuities, and epipolar lines in stereo." In *Fifth European Conference on Computer Vision*, Freiburg, Germany, Springer-Verlag. LNCS 1406, 232–248, June 1998.

[12] B. Julesz. *Foundations of Cyclopean Perception*. The University of Chicago Press, Chicago, 1971.

[13] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press., 1998.

[14] J. Malik. On Binocularly viewed occlusion Junctions. In *Fourth European Conference on Computer Vision*, vol.1, pages 167–174, Cambridge, UK, 1996. Springer-Verlag.

[15] S.B. Marapane and M.M. Trivedi. Multi-primitive hierarchical (MPH) stereo analysis. IEEE PAMI, 16(3):227-240, March 1994.

[16] D. Marr and T. Poggio. Cooperative computation of stereo disparity. *Science*, 194:283–287, 1976.

[17] D. Marr and T. Poggio. A computational theory of human stereo vision. *Proceedings of the Royal Society of London B*, 204:301–328, 1979.

[18] K. Nakayama and S. Shimojo. Da vinci stereopsis: depth and subjective occluding contours from unpaired image points. *Vision Research*, 30:1811–1825, 1990.

[19] Y. Ohta and T. Kanade. Stereo by intra- and inter-scanline search using dynamic programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-7(2):139–154, 1985.

[20] M. Okutomi and T. Kanade. A multiple-baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-15(4):353–363, 1993.

[21] S. B. Pollard, J. E. W. Mayhew, and J. P. Frisby. Disparity gradients and stereo correspondences. *Perception*, 1987.

[22] S. Roy and I. Cox. A Maximum-Flow Formulation of the N-camera Stereo Correspondence Problem In *Proc. Int. Conf. on Computer Vision, ICCV'98*, Bombay, India 1998.

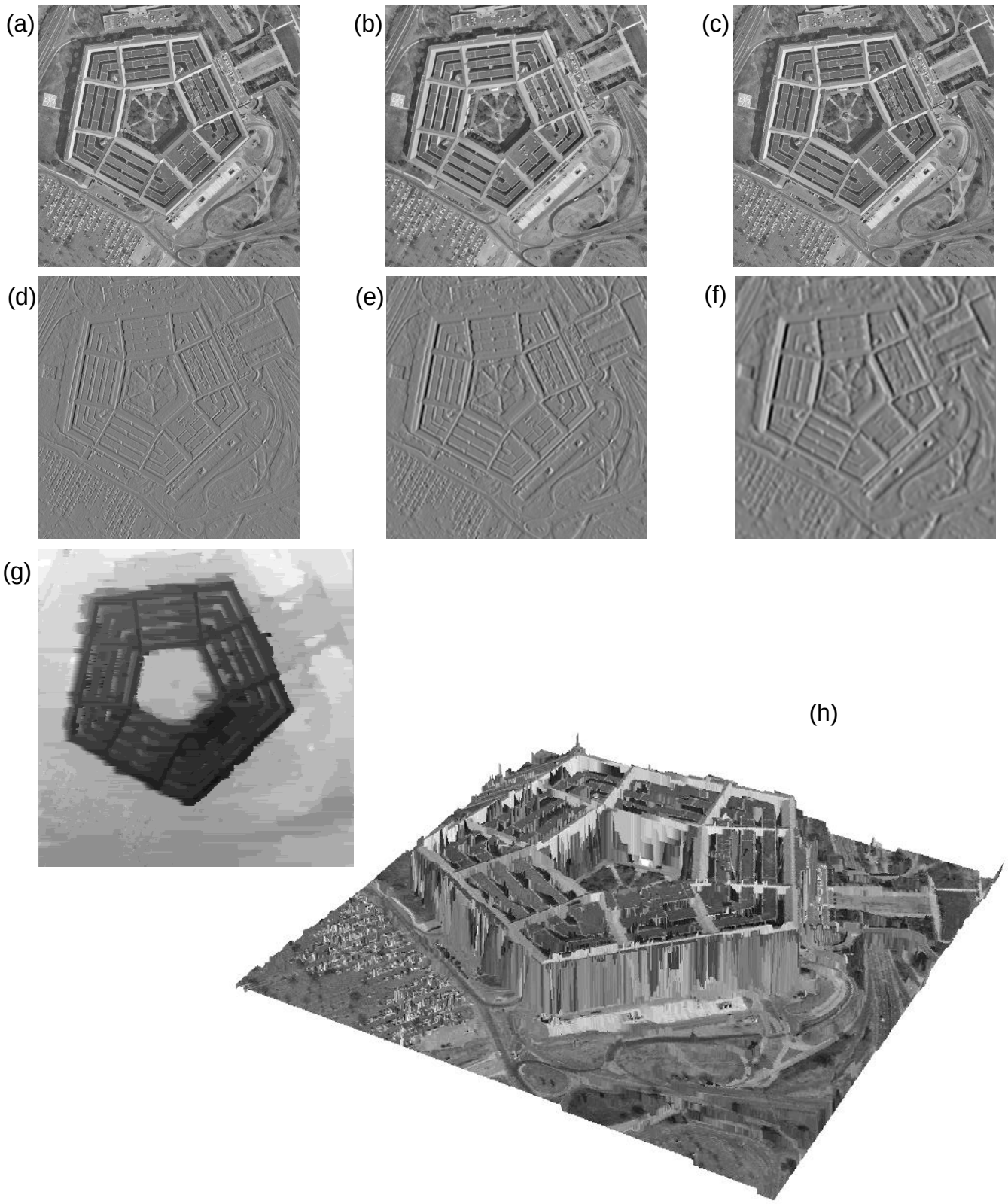

Figure 6. (a),(b) a sample image pair Pentagon, 508×512 pixels 8-bit gray-scale images. (c) The third left image is provided for cross fusers. (d) Edge detector response for the left image with $s=1$, (e) $s=2$, and (f) $s=4$. (g) The Disparity map detected. Disparity ranges from -5 to 16. (h) A 3D rendering of the result.