Bulletin of the Psychonomic Society
1987, 25 (2), 92-94

On the coarse-to-fine strategy in stereomatching

K. PRAZDNY
Artificial Intelligence Center, FMC Corporation, Santa Clara, California

The coarse-to-fine matching strategy is a computationally efficient search technique often used in stereomatching or motion detection. The coarse-to-fine search is based on the assumption of spectral continuity and does not work when different spatial frequencies carry unrelated information, as frequently occurs in images containing transparent or lacy surfaces. Simple demonstrations are used to illustrate and discuss this issue.

The correspondence problem is the problem of identifying image locations corresponding to the same cause (e.g., a point or region in the 3-D environment). Such problems arise frequently in machine vision, for instance in matching images obtained by different sensors (Hall, 1979), images in long-range apparent motion (Anstis, 1978), or images in stereopsis (Julesz, 1971). The correspondence problem is thus an abstract computational problem common to many different areas (e.g., matching problems arise in audition when left and right ear information is used to localize the sound source in the 3-D space). The difficulty of this task in stereopsis has been made clear by Julesz's discovery (1960) of the random-dot stereogram in which there is a multitude of false point-to-point matches at nearly every point in the image.

An interesting and computationally efficient strategy for solving the correspondence problem is a hierarchical algorithm known as coarse-to-fine search (Rosenfeld & Vanderbrug, 1977; Wong, Hall, & Rouge, 1976). With this strategy, the original scene is reduced to a set of images that are decreasingly lower in resolution and smaller in size; matches at the lower resolutions are used to select the most promising test locations at the next higher resolution. At search level L, except at the highest level (with the lowest resolution and smallest image size), a search is made only at best test locations found at the previous $(L-1)$ level. As the search level decreases and the spatial frequency (and image size) increases, increasingly fewer test locations need to be considered. Moravec (1977, 1980) used a simpler version of this method: only the most promising test location at each level was used to guide the search at the next level. Following some earlier work (Felton, Richards, & Smith, 1972), a similar technique has been used by Marr and Poggio (1979) in stereomatching. In this implementation, the images were not reduced in size but only in spatial resolution to simulate the spatial-frequency channels in human vision. The search radius, however, varied with the resolution: the lowest resolution was associated with the largest search interval and the highest spatial-frequency channel with the smallest search interval. Marr and Poggio (1979) thus attempted to solve the often massive ambiguity by essentially avoiding it. At the lowest resolution, there are fewer primitives to match, and matching can be attempted (for a given ambiguity tolerance) over large distances. If the spatial-frequency channels can be thought of as bandpass filters (and the images as bandlimited signals), then, statistically, the zero-crossings in the filtered images can be neither too close to nor too far from their neighbors. If these conditions are met, then one can restrict the matching range to within approximately the order of channel resolution.[1] Matches between structures in a coarser channel are used to align the two eyes using vergence eye movements (i.e., a global shift) to bring into register an adjacent higher spatial-frequency channel[2] with a smaller disparity range. By aligning the images using the disparity detected at this finer level, the next finer level can be brought into the correspondence range and so on until one reaches the highest spatial frequency channel (i.e., the most accurate disparity estimates).[3] This is the strategy of *progressive refinement*, which "homes in" on the most accurate disparity estimates through the sequence of ordered, ever-decreasing disparity ranges and filter sizes.

There are, however, two major problems with this approach. The first is specific to Marr and Poggio's (1979) implementation and has to do with the vergence eye movements as the mediator of the search selectivity. The problem is that a vergence eye movement is a global, resource-limited process: it can select or mediate disparity computation within only one region at a time. This problem can be circumvented by assuming that the shifts are local and parallel, that is, not dependent on the eye movements. The second problem is more general and has to do with the basic (and rarely mentioned) assumptions of the coarse-to-fine search. The strategy requires that information in adjacent spatial frequency bands is related, in the sense that they are generated by the same physical surface. In other words, the features must persist across spatial frequencies, or equivalently, the spatial averaging process must meaningfully "summarize" the fine-scale detail. This requirement is satisfied only within the boundaries of a relatively smooth and opaque object but does not hold, in general. The inheritance of disparities

The author's mailing address is: Artificial Intelligence Center, FMC Corporation, P. O. Box 580, Santa Clara, CA 95052.

from coarser to finer channels will fail at and near object boundaries and in transparency situations. Consider, for example, a grass surface viewed through a pattern of a picket fence. The (low) spatial frequencies associated with the vertical bars of the fence are not related to the (high) spatial frequencies of the grass texture. Similarly, the disparities of the fence are not related in any way to the disparities of the grass surface. In general, high spatial frequencies cannot be expected to inherit the disparities detected in the low spatial-frequency bands.

A laboratory demonstration of an effect similar to that of the picket fence is shown in Figure 1.[4] This display is essentially a classical random-dot stereogram portraying a central cyclopean square. Low spatial frequencies unrelated to the random dots are introduced by sinusoidally modulating the brightness of the white dots to produce monocularly visible vertical bars. In this way, the disparities of the bars can be manipulated independently of the disparities of the random-dot background and the cyclopean square (by varying the phase angle of the sinusoid in the left with respect to the right image). In Figure 1, the vertical bars have positive disparity (+25 pixels) while the cyclopean square is defined by negative disparity (−15 pixels) relative to the zero-disparity background and the bounding rectangle.[5] Thus the disparity difference between the coarse and fine structures is 40 pixels. As can be seen in Figure 2, the cyclopean square does not exist in the low and the vertical bars do not exist in the high spatial-frequency bands. Observe that the coarse structures due to the random-dot pattern are completely masked by vertical bars: the lower two spatial-frequency bands in Figure 2 do not contain any clue to the presence of the cyclopean square.[6] Because of the discontinuity between the disparities carried by different spatial-frequency bands, the progressive refinement strategy of the coarse-to-fine search cannot extract the disparities carried by the fine-scale detail (high spatial-frequency bands). The situation is further illustrated in Figure 3.

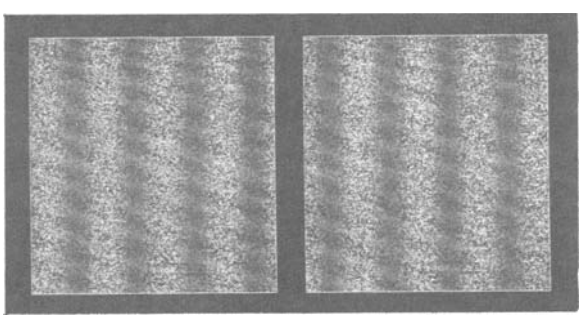

Figure 1. Coarse spatial structures independent of fine scale can be created, for example, by sinusoidally modulating bright picture elements of a conventional random-dot stereogram [p(white) = p(black) = 0.5]. Free fusion by crossing one's eyes in front of the display will produce a percept of a central protruding square and a set of black bars behind zero disparity plane. Low spatial frequencies (vertical bars) are independent of the fine-scale detail. Similarly, the disparities carried by them are independent of the disparities of the random-dot texture. This situation arises also, for example, in high-altitude photography of land partially obscured by cloud cover.

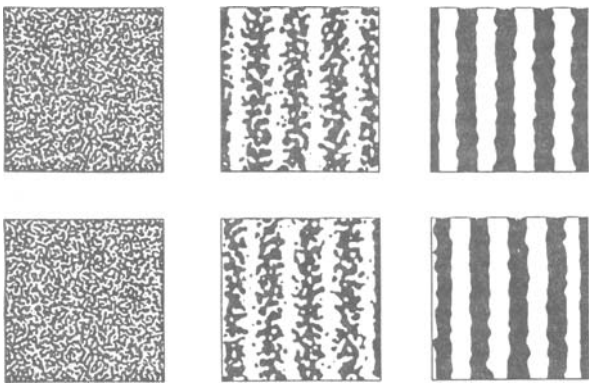

Figure 2. The sign bit representation of the difference of gaussians (DOG) filtered half images (positive and negative values are black and white, respectively) of the stereogram in Figure 1. The filter central frequencies differ by 1 octave. The width of the central excitatory part of the filter (w) is 6, 12, and 21 pixels (from left to right, respectively). The low spatial frequencies are dominated by coarse structures of the vertical bars. In other words, low spatial-frequency regions (Columns 2 and 3) contain no disparity information about the cyclopean square.

CONCLUSION

Hierarchical coarse-to-fine search has several highly desirable features. It is computationally efficient and allows decisions about the approximate location of the match to be continually updated and refined as the computation proceeds, unlike the more conventional correlation techniques in which no useful information about match location is available and no decision can be made until the whole correlation surface is computed for all permissible displacements. The critical aspect of the coarse-to-fine strategy is its proposition that ambiguity can be avoided and the search space restricted by coupling the match range to spatial frequency. Thus, a trade-off is being made between resolution and range. The coarse-to-fine technique based on spatial-frequency filtering relies, however, on the continuity of disparity (or displacement) across the spatial-frequency spectrum. This condition is not satisfied in many situations where different spatial-frequency bands may carry different and unrelated disparities.

In general, discontinuous disparity fields can be handled in two ways. One is to rely on the vergence eye movements: only a limited range of small disparities would be processed at each vergence angle. Disparities obtained in this way would have to be integrated over many vergence settings to obtain a complete disparity description of a scene. Another unsolved problem here is the vergence eye movement control.

The second way is to rely on the disparity distribution regularities present even in discontinuous distributions.

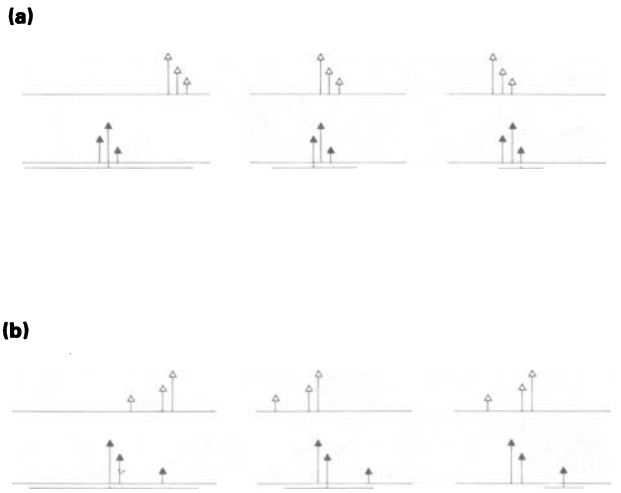

Figure 3. (a) The coarse-to-fine strategy of progressive refinement can correctly detect disparities due to a single opaque surface because the disparities of structures in adjacent spatial-frequency bands do not differ significantly. The disparities between the coarsest structures (large arrows) are used to approximately align the two images (Column 2), and this brings the next finer structures (middle arrows) into the correspondence range (dotted lines). The disparities between structures at this level (middle arrows) are in turn used to control the alignment so that structures in the next smaller channel (small arrows) come into the correspondence range (Column 3). (b) The strategy of progressive refinement fails when the disparity field is not continuous (i.e., in the presence of transparent surfaces) and different (unrelated) disparities are carried by different spatial frequency bands. When the images are matched on the basis of the coarsest channels (large and middle arrows) carrying positive disparities (Columns 2 and 3), the disparities between the finest structures (smallest arrows) are increased (Columns 2 and 3). Because the smaller channels are assumed to have smaller disparity ranges, the disparity gap cannot be bridged using signals from coarse channels.

Using this approach, a class of stereomatching algorithms can be formulated based on the notion of local coherence (Prazdny, 1985): a discontinuous disparity field can be viewed as a superimposition of several interlaced continuous disparity fields, each corresponding to a piecewise smooth surface. This formulation enables construction of different classes of algorithms based on the notion of local support and information orthogonality. In this formulation, two disparities are either similar, in which case they facilitate each other because they probably contain information about a common source, or dissimilar, in which case they are informationally orthogonal and should not inhibit each other because they probably carry information about different sources.

REFERENCES

ANSTIS, S. M. (1978). Apparent motion. In. R. Held, H. Leibowitz, & H. L. Teuber (Eds.), *Handbook of sensory physiology* (pp. 655-673). Berlin: Springer Verlag.

FELTON, B., RICHARDS, W., & SMITH, R. A. (1972). Disparity processing of spatial frequencies in man. *Journal of Physiology*, **225**, 349-362.

HALL, E. L. (1979). *Computer image processing and recognition*. New York: Academic Press.

JULESZ, B. (1960). Binocular depth perception of computer generated patterns. *Bell Systems Technical Journal*, **38**, 1001-1020.

JULESZ, B. (1971). *Foundations of cyclopean perception*. Chicago: Chicago University Press.

MARR, D., & POGGIO, T. (1979). A theory of human stereopsis. *Proceedings of the Royal Society of London*, **B204**, 301-328.

MORAVEC, H. P. (1977). *Towards automatic visual obstacle avoidance*. Paper presented at the 5th International Joint Conference on Artificial Intelligence, Cambridge, MA.

MORAVEC, H. P. (1980). *Obstacle avoidance and navigation in the real world by a seeing robot rover* (Tech. Rep. No. CMU-RI-TR-3). Robotics Institute, Carnegie-Mellon University, Pittsburgh, PA.

PRAZDNY, K. (1985). Detection of binocular disparities. *Biological Cybernetics*, **52**, 387-395.

ROSENFELD, A., & VANDERBRUG, G. J. (1977). Coarse-fine template matching. *IEEE Transactions on Man, Machine & Cybernetics*, **7**, 104-107.

WONG, R. Y., HALL, E. L., & ROUGE, J. (1976, December). *Hierarchical search techniques for image matching*. Proceedings IEEE Conference: Decision Control, Clearwater, FL.

NOTES

1. More precisely, to $\pm w/2$ where w is the width of the central excitatory region of the difference of gaussians (DOG) receptive field.

2. The central frequencies of the neighboring channels in their implementation differ by approximately 1 octave. The disparity ranges differ thus by the factor of 2.

3. The ambiguity in random-dot stereograms is not entirely avoided: local cooperative effects (*pulling*) were necessary to obtain unique matches.

4. The half image is 256×256 pixels. The cyclopean rectangle is 75×75 pixels. The stereogram was constructed for free fusion by crossing the eyes in front of the display.

5. The disparities are stimulus disparities. Positive disparity means that the right image is shifted to the right relative to the left image.

6. We can readily check (e.g., by superimposing a transparency of the upper sign bit representations of the DOG convolutions on the lower representation in Figure 2 and shifting it to obtain image alignments minimizing in some way the mismatch between the two images) that the two coarse channels signal only the (approximately) +25-pixel disparities while the fine channel contains only the zero disparity of the background and the −15-pixel disparity of the cyclopean square. Computational experiments indicate that the vertical bars in Figure 1 are detected by a DOG filter with a central excitatory region width (w) of about 22 pixels. Assume that the bars (Figure 2, Column 3) are somehow approximately aligned. This enables the medium structures (Figure 2, Column 2) to come into the correspondence range ($w=11$ pixels). Aligning the images using matches in this spatial frequency band *will not* bring the next finer channel ($w=6$ pixels) into the correspondence range of the smaller filters (Figure 2, Column 1). The gap of about 40 pixels between the disparities in the coarse and the fine channel cannot be bridged using the matches in the coarse channels as the control signal.

(Manuscript received for publication September 22, 1986.)