RECOVERING DEPTH FROM STEREO USING ART NEURAL NETWORKS

Stylianos Markogiannakis and Elias S. Manolakos

Communications and Digital Signal Processing (CDSP)Center
Electrical and Computer Engineering Department
Northeastern University, Boston, MA 02115, USA
Email: {stelios,elias}@cdsp.neu.edu

ABSTRACT

One of the long standing problems in passive stereo vision is that of constructing a range map using only two images providing two views of the same 3-D world scene. It amounts to identifying pairs of corresponding pixels, one pixel in each image, that are associated with the same point on the 3-D world. We are introducing ART-1 neural networks as a primitive capable for addressing the stereo correspondence problem. Using a multi-pass approach it is possible to increase gradually the density of matched points, while at the same time false matches are filtered by requiring close agreement between disparity estimates in a neighborhood. At the end, a reasonably dense disparity map is obtained, to the extend that it allows the reconstruction of the scene using interpolation methods. Our scheme has been tested on stereograms, virtual world scenes generated by computer programs (where the "ground truth" is known) as well as real-world scenes. In all cases the scene reconstructions are shown to be quite realistic.

1. INTRODUCTION

The *Adaptive Resonance Theory* (ART) was originally developed by G. Carpender and S. Grossberg [1] in an attempt to model the way biological organisms learn and behave. During the last ten years, specific architectures of ART neural networks, in many of their forms, have been found useful in numerous scientific and engineering domains including speech and image processing, temporal pattern recognition, neurobiology, control, telecommunications etc.

In this paper we show how ART-1 can be used as a computation primitive in order to generate accurate *disparity* (positional difference) maps that may be used for range estimation (via triangulation) and scene reconstruction (via interpolation). The proposed ART-1 based stereo correspondence algorithm integrates feature-based (since it uses edge maps) and area-based processing (for comparing patterns) with learning (to store patterns in an orderly fashion) and local associative recall (to retrieve the best candidate for matching). The disparity map's density can be increased by gradually reducing the value of the vigilance parameter from one pass to the next (thus relaxing pattern similarity conditions) while the requirement for close agreement of matches in a region (global consistency) can be handled by simple neighborhood voting that also minimizes false matches.

2. ART-1 FOR STEREO CORRESPONDENCE

After performing edge detection the image is segmented into strips consisting of a few scanlines that cover the whole length of the edge maps. Each strip is scanned using small-size windows (also called "tiles") that overlap maximally. The tiles generated from a strip of the left image are presented as inputs to the ART neural network. During training, the vigilance parameter (ρ) is set to its maximum value of one so that each window creates a class of its own. In addition, each committed neuron on the layer F2 of ART is excluded from further competition for the duration of the training. The net effect is that each window will reserve a neuron on F2 which, due to the artificial ordering used during competition, will also reflect the position of the center pixel of this window to the left image.

During recall, we expect that if a tile is to be matched, the corresponding window should belong to an appropriate section of the strip taken from the left image. This section is highlighted and competition is taking place only among the F2 neurons in this section. Once a match with a presented tile from the right image has been identified, the search resumes for the next tile and a new section of F2 neurons is highlighted.

The actual training, as implemented in this work, does not need to follow all the steps of the algorithm in [1] [2]. Since we know the position where we want to store the pattern (tile) we just need to use the weights updating formulas for the predetermined F2 neuron. Due to this simplified training, same patterns in different image locations will be stored in different non-adjacent neurons. This deviation from the use of ART as a pattern clustering mechanism was introduced in order to be able to learn the spatial relationship of tiles within a strip.

Once the patterns are learned, an initial recall cycle is performed to start the matching process. The vigilance is set to a high value (thus demanding close similarity). Focusing on a certain epipolar line, the strip from the right image, that corresponds to that line, is segmented the same way as the strip from the left image. Every tile extracted from the right image strip is presented in succession to the NN and a match that satisfies the vigilance criterion may be achieved. If it does, then the difference of the two indices (the position of the window in the right strip and the index of the winning F2 neuron) provides the disparity for the center pixel of the presented tile that is stored in memory and comprises the starting point for the generation of the

sparse range map.

If a maximum parallax is assumed, then all the neurons with indices that will generate larger disparities are "soft suppressed" and they do not participate on the current competition. Also, since the disparity values should be positive, all the neurons that will generate negative disparities are "hard suppressed" and they remain inactive for the rest of the cycle. This ensures that the consistency constraint will be obeyed. When all windows from the right image strip have been presented the cycle ends and the same procedure is repeated for the subsequent strips.

After the whole image has been presented, the vigilance level is dropped and the first strip is reintroduced much the same way as before. The only difference now is that for a match to be declared the calculated disparity must fall close to the disparity value suggested by the neighboring pixels. If this test is not passed the winning neuron is soft- suppressed and a new competition is engaged until a winner is found, or all the neurons that may participate are exhausted. This mechanism preserves the continuity constraint and ensures that reducing the vigilance will not render the system unusable by matching patterns almost at random. The multiple-pass scheme of dropping the vigilance and re-scanning the image can be used to populate a sparse range map while also removing at the same time false positive matches that may emerge due to texturing.

3. RESULTS OF EXPERIMENTS

In this section we present the various types of stereo pairs used to test the proposed system and discuss its response in relation to the special characteristics that these input pairs possess.

3.1. Random-dot Stereograms

Random-dot stereograms are image pairs that can be easily created by first generating a binary image (the reference) consisting of black dots (edgels) placed randomly. The other image is created by shifting regions of the reference by various degrees to the right, and refilling the resulting open areas with random dots. The illusion of depth is then achieved when the observer's fusion mechanism processes simultaneously the two images. Julesz [3] proposed random-dot stereograms as means to investigate certain aspects of the human vision. For our experiments we created random-dot stereograms of size 128x128 pixels for the well-known "Wedding Cake" and the "Tower of Hanoi" with three and five range levels respectively. The right image was produced after successive right shifts of different regions of the reference (left) image by 5 pixels (2 pixels) for each range level in the "Wedding Cake" set ("Tower of Hanoi" set) respectively.

The results of these experiments are summarized in Table 1 where we show quantitatively the performance of the correspondence system for various pixel densities. The table provides the number of black dots (features) for each image pair (left/right), the number of points (black or white) matched, and from them the percentage of points that were calculated with wrong disparities (false matches). In addition, intensity coded disparity images are provided for each pair in Figure 1, where we plot all the matched points. Most of the false positives lay at the boundaries of the objects where the discontinuities occur as it can be clearly seen in Figure 1(F) where we plot all the matched points registered in wrong levels. From Table 1 and Figure 1 it can be noticed that there is a large number of pixels without features that were matched. This is so because pairs of white dots (non-edgels) may also generate matches, as long as the tiles that contain them pass the vigilance test and the calculated disparity is compatible with the neighborhood beliefs.

Scene	Density	Black-dots	Matched	% False
Cake	50%	8227/8213	12982	2.79
"	30%	4892/4874	12922	2.54
"	10%	1641/1631	12584	2.74
"	3%	463/462	7041	3.10
Tower	50%	8198/8215	14017	5.12

Table 1: Number of features (black-dots) in the left/right image, points matched and percentage of false matches for various pixel densities. All images are 128x128 pixels.

"Cake"	14% noise		18% noise	
Density	Matched	False	Matched	False
50%	11135	222	10000	201
30%	10871	266	8106*	189
10%	7685	262	750*	28

*Difficult to subjectively perform fusion.

Table 2: Effect of impulse noise on number of matched points and false positives for various black-dot densities.

The effect of noise on the matching process is summarized in Table 2 where we show the number of points matched for various black-dot densities and various percentages of impulsive noise. It is evident that the performance of the system degrades gracefully as the noise levels increase without considerate increase of false matches. Figure 2 shows the 3-D reconstruction for the "Wedding Cake" with 18% impulse noise and with 50% pixel density. It is interesting to note the effects of noise by comparing Figure 1(D) with Figure 2. All the pixels corrupted by noise were not considered for disparity evaluation since they violate the compatibility constraint. Nevertheless, major portions of the objects are recovered at the correct locations with correct disparities. The false matches, again, are mostly located at the object boundaries and reconstruction is faithful.

From Tables 1 and 2, it is evident that the matching process was very successful in most cases and that it is not severely compromised at the discontinuities. In the absence of noise the matching success rate is very high and the false matches rate very low, even at low density levels, therefore the interpolated surfaces resemble very closely to the actual ones. The presence of noise will sparsify the range map but does not necessarily lead to increasing rates for false matches. Calculated disparity estimates for matched points are correct (with high probability) even if noise levels are high and there are not too many feature points available (density is low). In regions of the image where the distribution of matched points is uniform the reconstructions

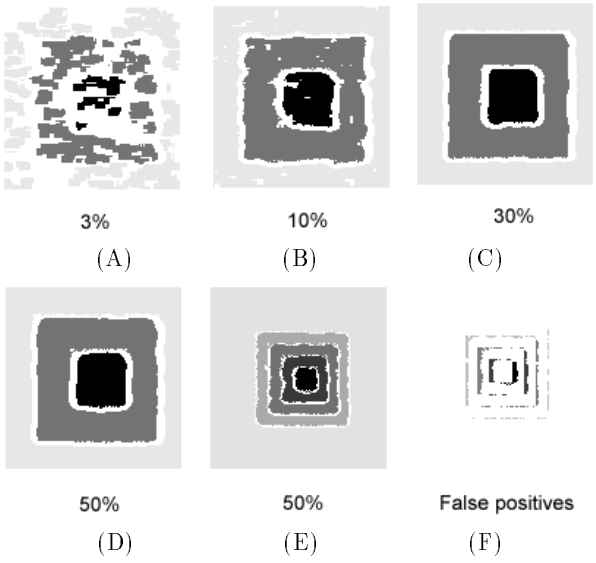

Figure 1: Intensity coded disparity maps (darker is closer) for Wedding Cake with various pixel densities (A,B,C,D) and Tower of Hanoi with 50% density (E). (F) shows the location of the false positives for (E).

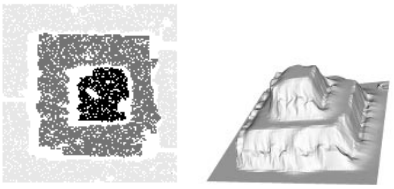

Figure 2: "Wedding Cake", density 50%, noise 18%: (A) intensity coded range map, (B) 3-D reconstruction.

reveal the general structure of the scene. Also, it is worth noting that the increased number of false positives in the case of the "Tower of Hanoi" set is primarily due to the fact that there are many discontinuities whose disparities differ only by 2 pixels (the neighborhood voting tolerance). The average execution time of the algorithm for these 128x128 images on a Sun Ultra 2 workstation was approximately 2 min.

3.2. Synthetic scenes

Our next step was to test the system using more realistic image pairs, that would allow us to evaluate various other aspects but in a controlled manner. So we created a number of artificial 3-D scenes using *Rayshade*, a ray-tracing tool available in the public domain that is very flexible and can produce high quality renderings. Camera placement, lighting, textures, materials and background can be defined by the user and real life like scenes can be easily created. Also, objective distance measurements from the cameras can be obtained for every visible point in the scene.

Due to lack of space we report here only the results

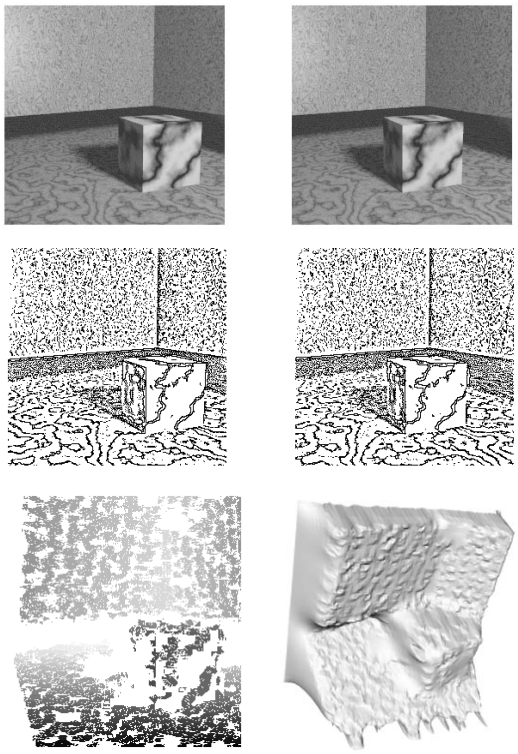

Figure 3: Top: Left/Right view of "Cube near interesting corner"; Middle: the corresponding edge maps; Bottom: intensity coded disparity map and a view of the reconstructed scene after interpolation.

from one stereo-pair. For more details the interested reader is referred to [4]. The interesting features of the scene in Figure 3 are the various normals of the surfaces. These surfaces provide ranges that vary smoothly in all directions. Also, the "walls" are detached from the "floor" so as to provide a view of the floor as it extends below them.

For a particular matched point, let us consider $e \equiv 100 \cdot |z - z_m|/z$ to be the relative error, where z is the actual depth given by Rayshade and z_m is the one calculated our the system. For the 256x256 images of Figure 3 our system matched 27,409 points and the mean and standard deviation of the relative error were $\overline{e} = 4.55\%$ and $\sigma_{n-1} = 8.35$ respectively.

As can be asserted from Figure 3 the depth recovery is quite good. The relative position of the objects is correctly estimated and surface discontinuities are determined successfully. As long as there are features to be matched on all surfaces of the objects, the matching mechanism can provide range estimates that lead to a realistic reconstruction of the scene after interpolation. Interestingly enough the reconstruction picks up the extension of the floor below the walls. In addition, although no special provision was taken to handle surface discontinuities, as long as the tile we are trying to match has enough neighborhood support to be considered a successful match, the two sides of the discontinuity can be recovered with ambiguity no greater

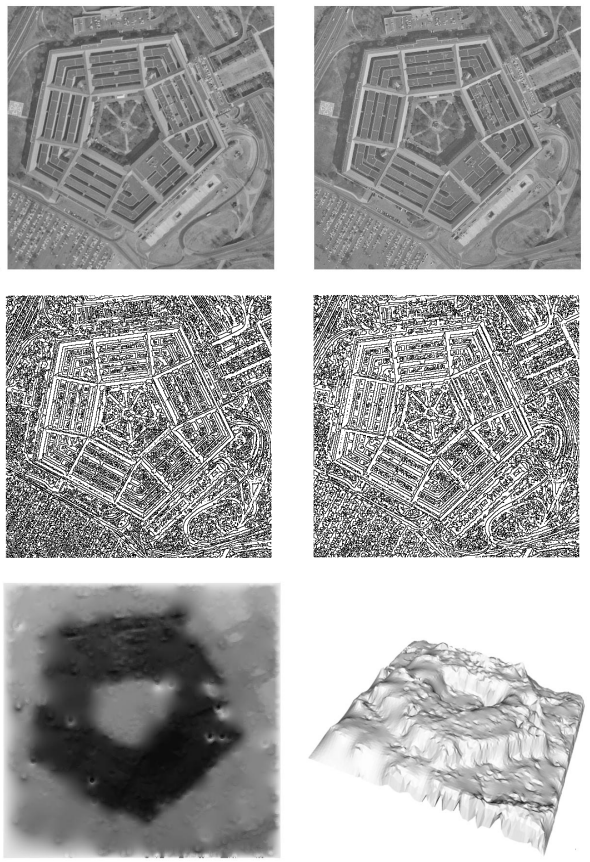

Figure 4: Top: Aerial stereo pair of the Pentagon; Middle: the corresponding edge maps; Bottom: the interpolated intensity coded disparity map and reconstruction.

than 1/2 tile.

3.3. Real World Scenes

A few sets of well-known stereo pairs available in the public domain were also used to test the proposed system [4]. The only parameter known in these images is their epipolar registration. We report here the results obtained using the widely known "Pentagon" pair. It is a set of aerial stereo photographs of size 480x480 that resembles closely the object setup on the "Wedding Cake". There is no registration of the walls of the building and there are many objects in the scene to provide a rich set of features to be used for matching. The image exhibits a few difficulties, namely many edges are near parallel with the epipolar lines, there is photogrammetric variation between the two images and there are a number of ribbon-like structures.

In the top two rows of Figure 4 we show the two photographs and the corresponding edge maps that differ significantly among the two pictures, especially in the upper two quadrangles of the building. Also the patterns of features extracted from small objects differ significantly as in the case of the parking lot at the lower left part of the scene. The reconstructed scene is shown at the bottom row of panels in Figure 3. An interesting area, the bridge above the freeway at the upper-right region of the images, was recovered despite the very small disparities in this region, as it can be seen from the interpolated intensity coded range map.

4. CONCLUSIONS

In this paper we have discussed how ART-1 integrates into a single primitive all the elements needed to address effectively the challenging problem of stereo correspondence. The experimental results presented are very encouraging. From the random-dot stereograms it can be seen that the patterns of features can be matched successfully, even when corrupted by noise or even if they are sparse. From the artificial scenes, it can be summoned that even when the patterns of features change slightly between the two images, due to stereo registration, the depth can still be recovered. Finally, from the real-world scenes, it can be concluded that scene reconstruction and depth estimation do not suffer big losses neither due to the physical acquisition systems' imperfections and their geometries, nor due to the photogrammetric variations between the two views.

The processing time ranges between 2 min (for images of size 128x128) and about 25 min (for images of size 512x400). These execution times along with the demonstrated capability to simulate efficiently large ART networks in parallel [4] [5] suggest that near real-time performance is feasible when using a ring of processors. Slight variations in execution times for images of the same size are attributed to the differences in the complexity of their edge maps.

The system can be extended to handle more complex potentially real-valued features if Fuzzy-ART or ART-2 type of networks are utilized. Other extensions currently under investigation include coupling of the edge-map density with the number of passes (vigilance levels) used and incorporation of local interpolation during the matching.

5. REFERENCES

[1] G. Carpenter and S. Grossberg. "A massively parallel architecture for a self- organizing neural pattern recognition machine", *Computer Vision Graphics and Image Processing*, 37:54–115, 1987.

[2] J. A. Freeman and D.M. Skapura. *Neural Networks: Algorithms, Applications, and Programming Techniques.* Addison-Wesley Publ., 1991.

[3] B. Julesz. Binocular depth perception of computer-generated patterns. *Bell System Technical Journal*, 39:1125–1162, September 1960.

[4] S. Markogiannakis. *"Stereo Correspondence using ART neural networks" Algorithms and Parallel Architectures"*. Engineer's Degree thesis, ECE Dept. Northeastern University, Aug. 1997.

[5] S. Markogiannakis and E. Manolakos. "Parallel implementation of ART-1 neural networks on processor ring architectures", chapter in *Parallel Architectures For Artificial Neural Networks – Paradigms and Implementations*, IEEE CS Press, to appear, 1998.