On the Application of a Modified Self-Organizing Neural Network to Estimate Stereo Disparity

Y. V. Venkatesh, S. Kumar Raja, and A. Jaya Kumar

***Abstract*—We propose a modified self-organizing neural network to estimate the disparity map from a stereo pair of images. Novelty consists of the network architecture and of dispensing with the standard assumption of epipolar geometry. Quite distinct from the existing algorithms which, typically, involve area- and/or feature-matching, the network is first initialized to the right image, and then deformed until it is transformed into the left image, or *vice versa*, this deformation itself being the measure of disparity. Illustrative examples include two classes of stereo pairs: synthetic and natural (including random-dot stereograms and wire frames) and distorted. The latter has one of the following special characteristics: one image is blurred, one image is of a different size, there are salient features like discontinuous depth values at boundaries and surface wrinkles, and there exist occluded and half-occluded regions. While these examples serve, in general, to demonstrate that the technique performs better than many existing algorithms, the above-mentioned stereo pairs (in particular, the last two) bring out some of its limitations, thereby serving as possible motivation for further work.**

***Index Terms*—Correspondence problem, nonepipolar, occlusion, self-organizing map (SOM), stereo disparity estimation, stereo-pair analysis.**

I. Introduction

COMPUTATIONAL stereo refers to the problem of estimating the depth of objects in a physical scene from multiple 2-D images captured from different viewpoints. In the case of images taken by two cameras with parallel optic axes and displaced perpendicular to the axes, the depth can be extracted from the difference in the positions of pairs of pixels in the two images that *correspond* to a single physical point. This difference is called *disparity*. The problem of physical depth estimation is reduced to locating the match for each pixel of one image, a pixel in the other, and, hence, the name *correspondence problem*. In practice, however, the camera axes may not be parallel. In such cases, *calibration* has to be carried out to estimate the external geometrical parameters which are then used to rectify the images.

Manuscript received March 2, 2006; revised June 25, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Hassan Foroosh.

Y. V. Venkatesh is with the Department of Electrical and Computer Engineering, Faculty of Engineering, National University of Singapore, Singapore 117576 (e-mail: yv.venkatesh@gmail.com).

S. K. Raja is with the Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India (e-mail: kumar@darbar.ee.iisc.ernet.in).

A. J. Kumar is with the Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India, and also with the Department of Physics, Raman Research Institute, Bangalore 560080, India.

Digital Object Identifier 10.1109/TIP.2007.906772

When humans view stereo pairs appropriately (i.e., using, for example, a stereoscope for separately given left and right images or the red–green goggles for the red–green stereo composites or anaglyphs), they perceive the corresponding 3-D scenes. The problem under consideration refers to automating this phenomenon, and can be formulated as follows. ***Stereopsis*: For the stereo pairs under consideration, estimate the disparity maps, thereby providing a solution to the correspondence problem.** Stereopsis is, in general, associated with one or more of the following difficulties: noise and illumination changes (and specularity) as a result of which the feature values for the corresponding points in the left and right images differ; lack of *unique match* features in large regions; presence of salient structural features: discontinuities in depth at object boundaries, discontinuities in surface orientation ("creases"), and steeply sloping surfaces, occlusion, and half occlusion.

Since there have been, in recent times, many surveys of the literature on stereopsis (for instance, [1]–[3]), we examine only those references that relate to or can be compared with the present results.

II. Literature Survey

The problem of stereopsis is normally broken into three relatively independent parts: feature selection, correspondence, and disparity interpretation. Starting from the pioneering work of Marr *et al.* [4], many assumptions have been made, and constraints imposed on the variables, in order to arrive at a solution to the stereopsis problem. The explicit assumptions are: stereo pairs are epipolar and the epipolar lines are horizontally aligned, i.e., the correspondence points in the two images lie along the same scan lines; the objects have *continuity* in depth and, hence, in disparity; there is a one-to-one mapping of an image element from one image to the other (*uniqueness*); and there is an ordering of the matchable points [5], and the implicit assumptions are: every pixel or point in one image has a corresponding point in the other; and there is a fixed range for disparity. On the other hand, typical constraints (relaxed in some recent papers) are: corresponding points have similar intensities/features; surfaces are smooth; and disparity gradient is 1.

Traditionally, solutions to the correspondence problem have been explored using area-, feature-, and phase-based, and Bayesian approaches, even though a starting point for an investigation can be normally, expected to be pixel-based [4]. The use of pixel values is declared to be "generally not suitable as matching primitives for stereo because a given scene entity often produces pixels in the left and right images of different intensities" [6]. However, more recently, in the process of explicitly dealing with occlusions, pixel-based approaches have

been revived in generalized forms, since they provide a dense disparity map, require no feature extraction, avoid the adaptive windowing problem of area-based correlation methods, and eliminate the need for sophisticated interpolation techniques.

Area-based algorithms provide a dense disparity field but lack an explicit occlusion model [7]–[9]. *Feature-based* algorithms are expected to give an accurate location of discontinuities. Marr and Poggio [10] propose the choice of zero-crossings but the disparity map is sparse, necessitating an interpolation stage to fill in the values of disparity at intermediate points and leading to loss of accuracy in disparity. For modifications and refinements, see [11]–[15]. In the last two references, edge segments are assumed to be "abundant in the environment." Such an assumption does not seem to be helpful in images with sparse features as in wire frames. Further, even though [15] and others employ the self-organizing map and the perceptron, there is little likelihood of an explicit relationship between their results and the human visual model since the self-organizing map is used merely for grouping of (feature) data. Finally, the phase-based approach of [16] is implicit, i.e., disparity is expressed in terms of phase differences in the outputs of local, bandpass filters applied to a stereo pair. However, there exist mathematical limitations with respect to the permissible disparity range.

Recent algorithms deal with *occlusion* explicitly, incorporating some stereo cues to deal with occlusions. In [17], intensity windows rather than individual pixels are matched, and smoothing of discontinuities is done within the framework of regularization. For a Bayesian formulation, see [18]. Assuming that an intensity variation accompanies depth discontinuities, [19] proposes a cost function based mainly on intuition and "justified solely by empirical evidence," resulting in "crisper, more accurate depth discontinuities on a wider range of images, and with much less computation." It is found that such an assumption is not, in general, valid for textureless and wire-frame images. Even dynamic programming (DP) in combination with graph-theoretic formulations has also been employed, assuming epipolar geometry and imposing and uniqueness and ordering constraints [20]. By designing a suitable graph, [21] relates the stereo matching problem to the minimum-cut problem. The complex graph approach in [22] includes, in the energy function, an occlusion term which represents the penalty associated with the pixels that cannot be matched; and invokes the uniqueness constraint. However, it fails in the case of wire-frame images. In [23], a maximum-flow (graph-theoretic minimum-cut) approach is proposed along with a convex discontinuity cost, and, more importantly, dispenses with epipolar geometry and the traditional ordering constraint. It is found that graph-cut algorithms require more computational time than DP methods. For some recent extensions, see [24]. According to [25], experimental results show that the graph-cut algorithm has the most accurate performance, especially in low textures scenes. However, no results seem to be available for wire frames.

The motivation for the proposed approach arose from the acknowledged stereo-perception ability [26] of the human visual system (HVS). A natural question is: To what extent do the existing algorithms have any resemblance to the HVS? It is found that very few of them provide an approach to imitating the HVS. In this context, an unsolved problem is: Are there trainable artificial neural networks that can be employed for stereopsis? In view of the quite exciting pattern recognition results using neural networks and, in particular, self-organizing maps [27], we propose the latter, and explore its application to stereopsis.

We organize the remaining part of the paper as follows. We introduce, in Section III, the self-organizing map (SOM) of Kohonen [28] and its modification, leading to the modified self-organizing map (MSOM); analyse it in Section III-A as applied to the problem of stereopsis, and explain the procedure employed to deal with occlusion; and present the main contributions in Section III-B. We generate, in Section IV, some basic synthetic images needed for testing stereo algorithms. After comparing, in Section V, the performance of the proposed approach with some of the results of the literature, we summarize, in Section VI, the special features and merits of the proposed approach along with its critique; and conclude the paper in Section VII. For details, see [29].

III. Proposed Self-Organization Map-Based Approach

The SOM is a neural network that transforms a higher-dimension feature space to a one- or 2-D discrete map in a topologically ordered fashion. During the training, which is unsupervised, the neuron/node whose synaptic weight is closest (in a Euclidean sense) to the input feature vector is declared the winner. The winner's weight vector is updated so that it moves closer to the input vector in the weight space. The topologically neighboring neurons (in the *weight-space*) are also updated in a weighted manner. The *updating* at nth iteration for the kth-component of weight vector w^n, with input vector α, is given by $w_k^{n+1} \leftarrow w_k^n + \eta(n) g_n \left(w_k^w - w_k^n\right)(\alpha_k - w_k^n)$, where w^w is the weight vector of the winning node; η is the learning rate, and $g_n(x)$ is the neighborhood-function, defined below in (2), with the parameter $\sigma(n)$ controlling the influence of the winner on its neighbors: $g_n(x) = \exp\left(-x^2/(2\sigma^2(n))\right)$. If the network is isomorphic to the data set, i.e., the number of nodes is equal to the number of data points, then, on convergence, each node maps onto a unique data point; and the positions of the nodes represent the best topographical fit for the initial distribution of the nodes.

It is found that the SOM cannot be directly applied to stereopsis, and certain modifications are called for in order to take care of stereo constraints. We now explain how to modify the SOM to discover a possible isomorphic mapping between the two images, the novelty in modification consisting primarily in taking into account the stereo-matching constraints. This leads to the creation of the MSOM. The steps are as follows.

Step 1: Initialize a node for each pixel in the left image, with the corresponding coordinates and intensity as the initial weights. Let $\left(w_1^{ij}, w_2^{ij}, w_3^{ij}\right)$ denote the weights of the node corresponding to pixel at (i,j) with intensity w_3^{ij}. For each pixel in the right image, associate a feature vector whose elements are the coordinates and the intensity. These feature vectors serve as inputs to the network. Select a pixel at random from the right image, and feed the corresponding feature vector as input to the network. Let $(\alpha_1^{mn}, \alpha_2^{mn}, \alpha_3^{mn})$ be the input feature vector corresponding to pixel at (m,n). *Step 2*: Let

(p, q) be the index of the winning neuron in the network for (m, n)th input, then

$$(p, q) = \arg\min_{i,j} \sqrt{\sum_{k=1}^{3} \left(w_k^{ij} - \alpha_k^{mn}\right)^2}. \quad (1)$$

The winner node (p, q) corresponds to the minimal distance, and represents the pixel in the left image which could be a match for the (m, n)th pixel in the right image. *Step 3*: Let M be the height and N, the width of an image. Update only the first two components of all the neuron weight vectors as follows:

$$w_k^{ij} \leftarrow w_k^{ij} + h_k(i', j') g_k(\Delta I) \left(\alpha_k^{(m+i')(n+j')} - w_k^{ij}\right) \quad (2)$$

where

$$i' = i - p, \quad j' = j - q h_k(i', j') = \eta_k \exp\left(-\frac{i'^2 + j'^2}{2\sigma_{hk}^2}\right) \quad (3)$$

$$g_k(\Delta I) = \exp\left(\frac{-(\Delta I)^2}{2\sigma_{gk}^2}\right) \Delta I = \left(w_3^{pq} - w_3^{ij}\right) \quad (4)$$

for $k = 1, 2$, and $\forall\ i,\ m \in \{1, 2, \ldots, M\}$, $\forall\ j,\ n \in \{1, 2, \ldots, N\}$, η_k is the standard learning rate, and σ_{hk}, σ_{gk} are the neighborhood parameters, which control the rate of propagation of disparity in the topological neighborhood and within the range of the object, respectively. Repeat the above three steps N_p times where N_p is a predetermined number; typically, $N_p = 100 \times MN$. *Step 4*: The disparity vector $\left(d_x^{ij}, d_y^{ij}\right)$ for each pixel (i, j) is defined as: $d_y^{ij} = i - w_1^{ij}$; $d_x^{ij} = j - w_2^{ij}$, corresponding to vertical and horizontal disparity. Note that MSOM assumes that the mapping of each pixel from one image to the other is isomorphic. However, in practice, some sections of one image will be occluded in the other [29]. For such images, the MSOM needs to be further generalized.

A. Analysis of MSOM

The MSOM converts the stereo-correspondence problem into an estimation of an *onto* map from each pixel of one image to the *corresponding* pixel in the other image of the stereo pair. Let S_L represent the set of pixels in the left image, and S_R, the set of pixels in the right image; $S_x = \{[x_x, y_x]^t\}$, the set of position vector of the pixels, and $I_x(s)$, for $s \in S_x$, is its intensity. Then, the correspondence map is represented as the transformation, $\mathcal{F}_{RL} : S_R \rightarrow S_L$, which is estimated on the basis of the constraints related to photometry, continuity, localization, and occlusion briefly explained below.

Photometric constraint: Minimize $\sum_{s \in S_R} |I_R(s) - I_L(\mathcal{F}_{RL}(s))|$. For $[i, j]^t \in S_R$, the correspondence point is given by $\mathcal{F}_{RL}([i, j]^t) = [p, q]^t$, $[p, q]^t \in S_L$ which is the winner node. See (1). The updating of the winner node in (2) tends to reduce the distance between the correspondence-node and the input in the coordinate space, thereby *strengthening* the map. *Continuity Constraint:* The transformation should be one-to-one and topologically ordered. We incorporate continuity using, $\left\|\mathcal{F}_{RL}(s^1) - \mathcal{F}_{RL}(s^2)\right\| < \epsilon$; if $\|s^1 - s^2\| < \delta_1$, and $\left|I_L\left(\mathcal{F}_{RL}(s^1)\right) - I_L\left(\mathcal{F}_{RL}(s^2)\right)\right| < \delta_2$, where $s^1, s^2 \in S_R$ and $\epsilon,\ \delta_1,\ \delta_2 > 0$. In contrast with the algorithms of the literature, which define continuity in the coordinate space alone, we use continuity in 3-D feature space, with intensity as one of the components. This helps preserve the discontinuity in disparity based on the assumption that it is accompanied by a discontinuity in intensity. The MSOM implements this criterion using the modified updating (2). The terms g and h define the neighborhood in the intensity and coordinate spaces respectively. The schematic diagram in Fig. 1 shows the difference in the directions of updating in the SOM and MSOM, the latter according to (3). Note that MSOM produces a one-to-one map that is also topology-ordering, with (2) ensuring that the neighboring nodes take on similar disparity, and, thereby, improving continuity in disparity.

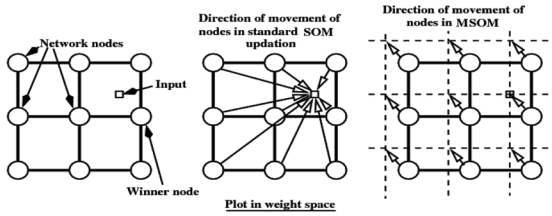


Fig. 1. Schematic illustration of the updating process.

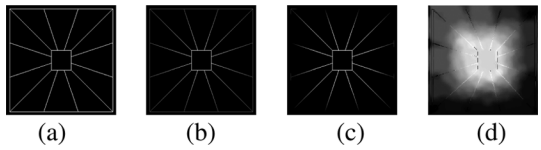


Fig. 2. Advantage of the g term in MSOM as applied to wire frames. (a) Left image of an wireframe stereo pair; (b) corresponding true disparity; (c) disparity map obtained with g term; (d) disparity map obtained without the g term.

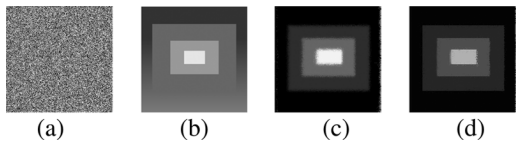


Fig. 3. Distortion effect of the g term in MSOM as applied to textured stereo pairs. (a) Left image of a random dot stereo pair; (b) corresponding true disparity map; (c) disparity map obtained with g term; (d) disparity map obtained without g term.

The continuity condition is implemented in the algorithm by neighborhood updating, using the h and g terms. The h term leads to the propagation of disparity, i.e., smoothing in the spatial neighborhood. The g term preserves the discontinuity in disparity across object boundaries with discontinuity in intensity and does not lead to smoothing or blurring of disparity boundaries. This is striking in the case of wire-frame stereo pairs where disparity can be estimated along the wires (see Fig. 2). A similar result is obtained for natural images (like Tsukuba), too [29]. On the other hand, in the case of high texture, the use of the g term leads to a distortion of the boundary in comparison with its omission (see Fig. 3).

In the standard SOM updating process, the winner and all its neighboring nodes move towards the input vector in the weight/feature vector space. However, for the correspondence problem, we require an isomorphic map. This inconsistency is demonstrated in the case of zero disparity (see Fig. 4). The updating (2) in MSOM ensures that the winner node moves towards the input

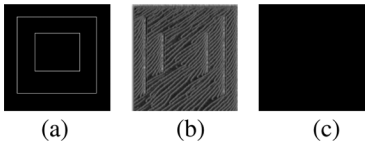


Fig. 4. SOM versus MSOM as applied to a wire frame. (a) Image used for both left and right image of stereo pair, i.e., disparity is zero everywhere; (b) disparity map obtained using standard SOM updation; (c) disparity obtained from MSOM (zero everywhere).

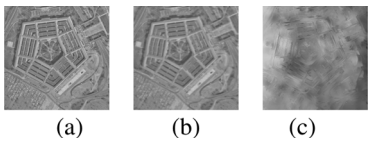


Fig. 5. Major difference between SOM and MSOM: No intensity upgradation in MSOM. (a) Right image of the pentagon stereo pair; (b) intensity weight of the network (right image) after 30 iterations with intensity updation; (c) estimated disparity map.

vector, *whereas the neighbors move to corresponding neighbors of the input vector*.

In contrast with the application (of the standard) SOM, the intensity feature in MSOM is not updated, since such an operation would change the photometric constraint, and produce an erroneous disparity (see Fig. 5). *Localization Constraint:* For stereo pairs containing periodic textures, if the periodicity of the texture is n pixels, then there is a match at every n pixel interval. If d is the actual disparity at that point, observations suggest that the perceived depth corresponds to a disparity of $\min(d \mod n, n - d \mod n)$. This shows that, out of all the maps possible, we need to select the one that gives minimum disparity. We minimize $\sum_{s \in S_R} \|s - \mathcal{F}_{RL}(s)\|$, thereby ensuring that, in the case of nonunique solutions, we choose the localized solution in the coordinate space. As far as we are aware, none of the algorithms in the literature consider this case (of periodic textures). Since there are multiple matches, uniqueness is not defined by the existing algorithms of the literature, and the range of disparity has to be known *a priori* in order to get a unique solution. Hence, the choice of disparity depends on the implementation. In contrast, MSOM needs no such prior knowledge of disparity range. *Occlusion Constraint:* Many stereo algorithms assume implicitly that there exists a correspondence for every pixel of one image in the other, and, hence, produce a false mapping for stereo pairs with half-occluded points. However, the MSOM inherently detects occluded pixels, and uses a modified pixel set to estimate disparity. Such a procedure is distinct from the various pre- and postprocessing approaches proposed in the literature in order to find the occlusion points. Details are left out for lack of space [29]. The MSOM tracks pixels that do not win, and then labels them as occluded pixels automatically. It is found that occlusion detection is successful for not only the simple rectangular fronto-planar objects [29] but also real textured images (Fig. 6).

B. Main Contributions of the Paper

The primary contribution is the MSOM which is a novel modification of SOM [28] to facilitate the latter's application to stereopsis. The special characteristics of the MSOM are: It does not assume epipolar constraint, is invariant to vertical disparities, and does not require a prior knowledge of the range of disparity. In fact, the search region could be taken as the whole image, and the disparity map obtained is unique. However, in practice, we employ a search window *only* to speed up the algorithm. The MSOM is superior to the existing algorithms in its ability to reliably estimate disparity in images having low textures. It can handle occlusions and salient features better than the earlier algorithms, and performs well on wire-frame images. It is robust not only to noise but to limited size variations and blurring of one image, and achieves global stereopsis automatically as a propagation of local correspondences. The resulting disparity maps are, in general, comparable to those in the literature but with less restrictive assumptions; and better than those in the literature as applied to the difficult bench-mark (synthetic) stereo pairs described in Section IV.

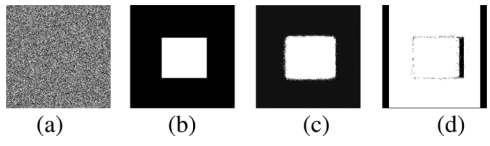


Fig. 6. Occlusion in a textured background. (a) Left image of random dot stereo pair; (b) corresponding true disparity map; (c) MSOM-computed disparity map; (d) corresponding occlusion map (black pixels are the occluded ones).

IV. Generation of Synthetic Stereo Pairs

We generate certain classes of elementary stereo pairs in order to help us identify limitations on the use of constraints in the stereo algorithms. 1) **Planar surfaces:** The images are of planar surfaces that are texture-mapped with different types of textures or may be textureless. For stereo constraints analysis, we examine two of its categories. Fronto-planar surfaces (FPS) are visible surfaces of objects that are parallel to the image plane, having a uniform disparity. The corresponding pixels of the object have a one-to-one map. The performance of the algorithms on these stereo pairs shows the use of continuity constraint. In the case of textureless objects, the ability of an algorithm to propagate disparity can be tested. Nonfrontal planar surfaces (NFPS) are planar surfaces that are at an angle with the image plane. The surface has a uniform disparity gradient. Here, the correspondence pixels do not have a one-to-one mapping; hence, they violate the uniqueness condition. The results of the algorithms on these stereo pairs will demonstrate the interaction between continuity and uniqueness constraints. The results of the algorithms also depend on the types of texture, based on the matching approach used. Since the types of textures are innumerable, we use random-dots with different sparsity to check on matching. As mentioned above, a periodic texture is a special case meant to test for localization constraint. 2) **Curved surfaces:** Here, the objects could be of any shape, but from the stereo point of view, the objects considered are convex and, hence, do not have self occlusions. Textureless surfaces do not give rise to curved surfaces. On the other hand, Lambertian surfaces (LCS) with intensity gradients produce curvatures [Fig. 10, top row (a), in Section V below]. We can evaluate not only the ability of the algorithm to use intensity for matching but smoothness criteria. These stereo pairs do not

Fig. 7. Panum's stereo pair. (a) Left image; (b) right image; (c) estimated disparity map using SOFM.

have extractable features like zero-crossings, and, hence, the feature-based approaches fail to produce disparity.

3) **Thin/fine objects wire-frames (WFS)**: One of the most significant features of wire-frame images [Fig. 10, bottom row (a), in Section V below] is that they contain no clues (like shading or texture), which are important pointers to depth information in (natural) monocular images. The wire-frame images have sparse match points, with disparity confined to the lines, and represent the narrow objects (in the image). Such images help us test the performance of a stereo algorithm on fine-structured objects. 4) **Objects with occlusion, half occlusion:** A discontinuity in disparity map along the epipolar line gives rise to half occlusions. To analyze the occlusion handling capability of the algorithm, we use the stereo pair with Panum's limiting case [4] (Fig. 7).

V. Implementation and Results

We apply the MSOM to standard, nonepipolar and noisy (elementary and standard) stereo pairs, and compare the results with those of sum-of-accumulated difference (SAD), cooperative stereo (CS) [7] and graph-cut-based (GC) [21], [30] approaches. In order to speed up the MSOM, the winner for each input is determined in a small neighborhood as against searching the whole image, and weight updation is carried out within the regions where the Gaussian tail tapers off to 0.001. Typical values of different parameters that are used for the experiments are as follows: For an image of size $N \times M$, number of iterations, $N_p = 100 \times MN$, depending on the sizes of the image and textureless region in it; neighborhood parameter of h, $\sigma_h = 10.0$ (which controls the spread in disparity); neighborhood parameter of g, $\sigma_g = 10.0$, depending on the intensity levels of the image; and learning constant, $\eta = 0.1$ (which controls the speed of convergence and the smoothness of the disparity map).

The choice of the values for N_p and η has been guided by those in the standard literature on neural networks. The others parameters were determined by trial and error. Typically, for images with texture and small depth gradient, $\sigma_h = 6$, $\sigma_I = 5$ and $\eta = 0.1$, and for images with low texture and high depth gradient, $\sigma_h = 3$, $\sigma_I = 5$ and $\eta = 0.05$. Compared to other standard algorithms, the MSOM is slow. No attempt was made to optimize the code since the primary goal of the study was to check on the general correctness and viability of the algorithm. If the processing window size is $n \times m$, the processing time is given by $t_{\text{tot}} = ((n*m) \times (N*M) \times N_{It}) * t_{\text{search}}$ where N_{It} is the number of iterations and t_{search} is the time required to search for the winning node. For an image-pair of size 256×256 with a search region of 20×10 pixels, the time taken to compute the disparity map on a Linux platform with Pentium-IV 1.4-GHz CPU and 256-MB RAM is, typically, 120 s. If we know the disparity range to be, say, 0–20, and there is no vertical disparity, a window of size 20×1 requires a computational time is 50 s. The speedup over using the whole image as a window is of the order of 60.

A. Performance Measures for Comparison

In order to compare the results of different algorithms, we adopt a method similar to that of Scharstein *et al.* [31]. In the case of stereo pairs with true disparity map, the estimated and the true disparity maps are normalized such that the pixels with lowest disparity are assigned zero value, and the highest disparity value is equal to the range of disparity. Then, the difference in the disparity value is computed at each pixel between the estimated and the true value. If the difference is greater than a threshold (of 2, chosen by trial and error), then the corresponding pixels are classified as "bad." The percentage of bad pixels in the image is used as the measure for comparison, and we call it percentage of bad pixels in disparity (PBD).

On the other hand, in the case of stereo pairs without the true disparity maps, only a qualitative evaluation can be made. However, in an attempt at some quantification, we employ the estimated disparity map to estimate the right image from the left image. Then, the nonoccluded pixels for which the intensity difference between the actual right image and the estimated right image pixels is greater than a threshold (of 50 out of 255) are classified as bad pixels. We, thus, obtain the percentage of bad pixels in intensity (PBI). It should be noted that this approach fails to give reliable measures in textureless regions and near the object boundaries.

For lack of space, a few experimental results are presented here [29]. However, the error measures for the other tested stereo pairs are also given in the tables. On the same PC as indicated above, typical computational times for an image of 256×256 pixels and a disparity range (0–20) are: SAD: 14 s (with block size of 20×20); CS: 40 sec (15 iterations); GC: 55 sec; and MSOM: 115 s (with search window size of 40×40 and 100 iterations).

B. Elementary Stereo Pairs

Results show that, in the case of textured images, if the intensity distributions in the object and in the background are similar, then the disparity estimate by MSOM at the boundaries is more noisy than that obtained by other algorithms. On the other hand, in the case of textureless images, SAD is unable to propagate disparity, in comparison with other algorithms. The MSOM is able to estimate disparity in the case of textureless object and background, whereas the other algorithms fail.

In Fig. 8, the FPS object has a periodic texture with periodicity of 15 pixels, and the object has a disparity of 10 pixels. Observation shows that the object is perceived as having a disparity of -5 pixels. The MSOM is able to compute the disparity of -5, because of the localization constraint, whereas, with other algorithms, the computed disparity depends on the search range.

If a binary random-dot stereogram is treated as a texture, the MSOM fails to give correct disparity. This is due to the lack of features for matching but can be resolved by blurring the images using a Gaussian mask so as to introduce some spatially

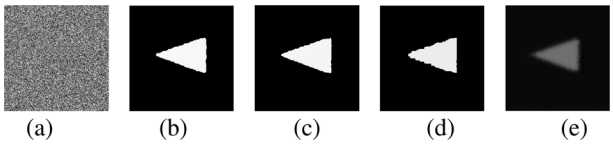


Fig. 8. Results for FPS with textured object (periodic texture) and background. (a) Left image; (b) the corresponding estimated disparity map using SAD; (c) cooperative stereo; (d) graph cut; (e) MSOM.

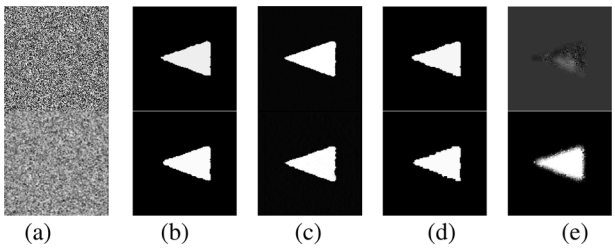


Fig. 9. Top row: Results for FPS with binary texture-mapped object and background. (a) (Left image) Corresponding disparity maps obtained from: (b) SAD; (c) cooperative stereo; (d) graph cuts; (e) MSOM. (Bottom row) Results for the Gaussian-blurred pair of the same FPS: (a) left image; corresponding disparity maps obtained from: (b) SAD; (c) cooperative stereo; (d) graph cuts; (e) MSOM.

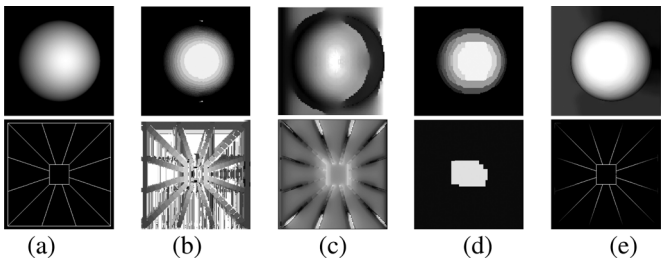


Fig. 10. Top row: Results for LCS. (a) Left image and corresponding disparity maps obtained from (b) SAD; (c) cooperative stereo; (d) graph cuts; (e) MSOM. (Bottom row) Results for WFS: (a) left image and corresponding disparity maps obtained from (b) SAD; (c) cooperative stereo; (d) graph cuts; (e) MSOM.

dependent intensity. The disparity results with and without blur are shown in Fig. 9.

Even in the absence of features, like edges or zero-crossings, the network produces an acceptable disparity map (Fig. 10). The disparity map produced by MSOM is smooth compared to other approaches. The superiority of the MSOM algorithm is demonstrated in the case of wire frames, for which the computed disparity is *indeed* confined to the lines, and the disparity map is close to the true-disparity. In contrast, the other algorithms give rise to *blocky* artifacts in the disparity map.

C. Standard Stereo Pairs

Qualitative comparison of the results shows that MSOM has the following distinct advantages: it can estimate disparity of fine objects, detect discontinuity in disparity effectively, preserve the shape of objects, and compute a smooth disparity map. The PBD and PBIs for the stereo pairs are tabulated in Table I.

1) Stereo Pairs With Known Disparity Map: Fig. 11 shows an estimated disparity map for a standard stereo pair (Tsukuba) which has only fronto-planar surfaces. The background is textured with small textureless patches in the foreground. Compared to other algorithms, the MSOM preserves the shape of the objects. However, it fails to estimate the disparity of the lamp post. The disparity values close to the outer boundary of the lamp and the camera are noisy, since the occluded parts have similar intensities.

TABLE I
PBD FOR ELEMENTARY STEREO-PAIRS. ^{a}O AND B IN THE BRACES STAND FOR OBJECT AND BACKGROUND. THE BAR INDICATES IT IS TEXTURELESS

Stereo-Pair [a]	SAD	CS	GC	MSOM
FPS(OB)	1.5575	1.4400	1.5443	2.9843
NFPS($O\bar{B}$)	8.2760	2.4000	8.7576	2.3107
FPS($\bar{O}\bar{B}$)	25.7365	22.4510	9.3866	4.3681
FPS($O'B$)	1.9597	2.0657	1.8008	9.7474
FPS$(OB)_{blur}$	1.2530	1.7578	1.4731	3.1647
WFS	68.9728	67.9764	7.9590	0.8209
LCS	12.7006	22.6512	18.3457	11.2762

TABLE II
PBD FOR THE NATURAL IMAGES WITH TRUE-DISPARITY MAPS

Stereo-Pair	SAD	CS	GC	MSOM
Tsukuba	8.3659	7.8902	5.4941	6.5701
Venus	24.7182	6.8944	5.6629	8.2961
Cones	14.9647	11.7814	10.9997	11.1556
Teddy	15.4708	16.0013	24.0699	12.1582

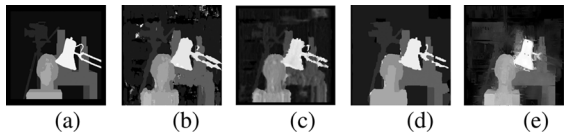


Fig. 11. Results for Tsukuba (natural) stereo pair: (a) Known disparity map and corresponding disparity maps obtained from (b) SAD; (c) cooperative stereo; (d) graph cuts; (e) MSOM.

TABLE III
PBI FOR THE NATURAL IMAGES WITH UNKNOWN TRUE-DISPARITY MAPS

Stereo-Pair	SAD	CS	GC	SOM
Pentagon	18.3659	8.7945	11.9412	10.3411
Renault	9.2362	6.1850	9.8001	7.0193

2) Stereo Pairs With Unknown Disparity Map: In the (highly textured) Pentagon stereo pair [29], the camera motion is not exactly horizontal but contains some rotation, violating the epipolar constraint. The MSOM gives the structure of the pentagon effectively and distinguishes the bridge from the ground. See Table III for PBIs of different algorithms.

3) Occluding Contours: Observations by Nakayama *et al.* [32], [33] show that the occluded contours or sections always take on such disparity values as to suggest that they are behind the occluding object. This is clearly visible in the Panum's case (Fig. 7), where a reversal of the stereo pair will change the contours that are matched. None of the existing algorithms take this into account.

4) Orientation Discontinuities: The MSOM can produce discontinuities in orientation only if there is a change in intensity. In certain cases, there are two possible disparity maps: one in which only the middle black line appears to be raised above the background; and the other in which the white plane forms a roof like structure. Our perception gives rise to the roof-like effect, but the MSOM gives the disparity map in which only the line

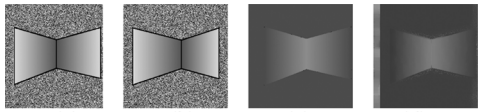


Fig. 12. Results for crease-stereo: (a) left image; (b) right image; (c) perceived disparity; (d) estimated disparity map using MSOM.

Fig. 13. Results for the Tsukuba stereo pair with a modified right image—set with size change. (a) Right image which is 90% of the original height; disparity maps obtained from (b) SAD; (c) cooperative stereo; (d) graph cuts; (e) MSOM.

TABLE IV
PBD FOR THE STEREO-PAIRS WITH NOISE

Stereo-Pair	SAD	CS	GC	SOM
Vertical shift	48.3688	28.6301	24.0831	7.4517
Vertical scale	54.0210	41.8365	25.0172	8.3523
Impulse noise	18.8757	10.1027	15.9569	17.8123
Blur	16.2073	13.9093	16.3475	17.1223
Contrast	36.0681	23.5982	71.3943	26.3433

Fig. 14. Results for modified right image of Tsukuba pair—set with contrast change. (a) Contrast enhanced version; and disparity maps from (c) SAD; (d) cooperative stereo; (e) graph cuts; (f) MSOM.

is raised [29]. However, the MSOM can estimate disparity correctly in the presence of discontinuity in orientation, accompanied by texture or intensity gradient as in Fig. 12.

The MSOM algorithm, in which the planes are forced to take on fronto-planar form, performs well on textured vertical and horizontal slant surfaces. However, it fails to produce gradient in textureless surface unless there is good support from the boundaries [29].

5) Other Stereo Pairs: The first is nonepipolar. In Fig. 13, the right image of the tsukuba stereo pair is altered to create nonepipolar stereo pairs. The results in the figure show the superior performance of our approach compared to other algorithms. The PBD for the MSOM (Table IV) shows that it is robust with a nonepipolar constraint. For the second, noisy pair, the robustness of the algorithms against noise in the stereo pair is analyzed by estimating disparity map with added noise in the right image. The resulting disparity-maps with impulse noise, and blurred right image are found in [29].

However, since the pixel-by-pixel intensity is used to find the closest match and for updating weights, if the two images of the stereo pair have different contrasts, the MSOM fails to give an acceptable result (see Fig. 14).

6) Specular Surfaces: The disparity map, estimated for a typical stereo pair with specular reflection, is found to be erroneous in the region of reflection [29].

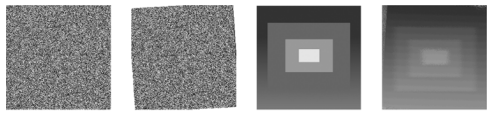


Fig. 15. Results for a modified right image—set with rotation. (a) Left image of the gray-scale random dot (RDS), (b) right image of the stereo pair, rotated by 15°, (c) true disparity map, (d) MSOM output.

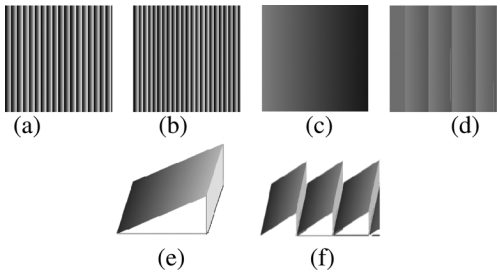


Fig. 16. Results for periodic, slant plane stereo pair: (a) Left image, (b) right image of the stereo pair, (c) true disparity map, (d) MSOM output, (e) true depth-map, (g) perceived depth map (same as MSOM-output!).

VI. UNIQUE FEATURES AND CRITIQUE OF MSOM

The MSOM is distinct from a mere application of the standard SOM, with no need for interpolation to propagate the disparities along horizontal lines, since the MSOM is able to estimate disparities in these regions by local cooperation. The disparity maps compare favorably with those of the literature, and, in the difficult cases of nonepipolar geometry and occlusion, are superior to those of the literature.

The earlier psychovisual experiments by Julesz (see [26]) seem to suggest human visual ability to fuse stereo pairs subjected to rotations of upto 15°. In Fig. 15, the left image of the RDS pair is rotated about the center of the image. This gives rise to an additional gradient in vertical disparity and the horizontal disparity. Most of the existing algorithms assume that epipolar geometry can be estimated from the images, using feature points like lines or corner points. However, for Fig. 15, it is not possible to find the epipolar lines by any method available in the literature.

As far as occlusion is concerned, the MSOM can track the pixels that do not win, and then label them as occluded pixels automatically. Compared to other algorithms, the MSOM has inherent global interaction. In the periodic, slant-plane stereo pair of Fig. 16, applied on a slant plane, the left and right images have the same texture but with different periods. Experiments show that we perceive a sawtooth shaped object [Fig. 16(g)], instead of a slant plane [Fig. 16(f)]. The phenomenon can be explained by using the coordinate constraint. This seems to be the first ever such observation; and it is significant that MSOM is able to predict the perceived object. For wire-frame stereo pairs, MSOM can estimate disparity along the wires because of the g term.

In the case of stereo pairs of different sizes with texture, there should be a limit on the amount of difference in the sizes of the left and right images. If the image is low textured, the MSOM fails. Further, since the matching criteria are based on intensity

comparison of individual pixels, the MSOM fails to give an acceptable disparity map for stereo pairs in which the right (or left) image has a different contrast from the correct version.

VII. Conclusions

We have proposed a novel modification of the self-organizing map for estimating the disparity map from a stereo pair of images. After initializing the network to one of the images, we compute the amount of deformation required to transform it into the other image; the deformation itself constitutes a measure of disparity. It is conjectured that this deformation is closely related to the phase difference between Gabor-mask outputs of Qian *et al.* [34] who model the neurons in the visual cortex using Gabor functions. The MSOM has many special properties: no assumption of epipolar geometry of the images, no limit on disparity but accommodates salient features (like discontinuous depth values at boundaries and surface wrinkles) and half occlusions. It performs better than many existing algorithms on synthetic and natural stereo pairs (including wire frames). Examples are given to illustrate not only its superiority but also its limitations.

Acknowledgment

The authors would like to thank the expert referees whose critical comments and suggestions have led to this significantly improved version of the paper.

References

[1] U. R. Dhond and J. K. Aggarwal, "Structure from stereo—A review," *IEEE Trans. Syst. Man, Cybern.*, vol. 19, no. 6, pp. 1489–1510, Nov./Dec. 1989.

[2] M. Z. Brown, D. Burschka, and G. D. Hager, "Advances in computational stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 8, pp. 993–1008, Aug. 2003.

[3] Y. Liu and J. K. Aggarwal, *Local and Global Stereo Methods, in Handbook of Image and Video Processing*, A. L. Bovik, Ed. New York: Elsevier, 2005, pp. 297–308.

[4] D. Marr, *Vision*. San Francisco, CA: Freeman, 1982.

[5] J. Little and W. E. Gillett, "Direct evidence for occlusion in stereo and motion," *Image Vis. Comput.*, vol. 4, no. 8, pp. 328–340, 1990.

[6] J. P. Frisby and S. B. Pollard, *Computational Issues in Solving the Stereo Correspondence Problem in Computational Models of Visual Processing*, M. S. Landy and J. A. Movshon, Eds. Cambridge, MA: MIT Press, 1991, pp. 331–357.

[7] C. L. Zitnick and T. Kanade, "A cooperative algorithm for stereo matching and occlusion detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 7, pp. 675–684, Jul. 2000.

[8] S. Yoon, S. K. Park, S. Kang, and Y. K. Kwak, "Fast correlation-based stereo-matching with the reduction of systematic errors," *Pattern Recognit. Lett.*, vol. 26, no. 14, pp. 2221–2231, Nov. 2005.

[9] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proc. 3rd Eur. Conf. Computer Vision*, 1994, pp. 150–158.

[10] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proc. Roy. Soc. Lond. B*, vol. 204, pp. 301–328, 1979.

[11] H. Sunyoto, W. van der Mark, and D. M. Gavrila, "A comparative study of fast dense stereo vision algorithms," in *Proc. IEEE Intelligent Vehicles Symp.*, 2004, pp. 319–324.

[12] M. H. Lin and C. Tomasi, "Surfaces with occlusions from layered stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 8, pp. 1073–1078, Aug. 2004.

[13] G. Pajares, J. M. Cruz, and J. Aranda, "Stereo matching based on the self-organizing feature map algorithm," *Pattern Recognit. Lett.*, vol. 19, pp. 319–330, 1998.

[14] J. M. Cruz, G. Pajares, and J. Aranda, "A neural network approach to the stereovision correspondence problem by unsupervised learning," *Neural Netw.*, vol. 8, no. 5, pp. 805–813, 1995.

[15] J. M. Cruz, G. Pajares, J. Aranda, and J. L. F. Vindel, "Stereo matching technique based on the perceptron criterion function," *Pattern Recognit. Lett.*, vol. 16, pp. 933–944, 1995.

[16] D. J. Fleet, A. D. Jepson, and M. R. M. Jenkin, "Phase-based disparity measurement," *CVGIP: Image Understand.*, vol. 53, no. 2, pp. 198–210, 1991.

[17] D. Geiger, B. Ladendorf, and A. Yuille, "Occlusions and binocular stereo," *Int. J. Comput. Vis.*, vol. 14, no. 3, pp. 211–226, 1995.

[18] P. N. Belhumeur and D. Mumford, "A bayesian treatment of the stereo correspondence problem using half-occluded regions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1992, pp. 506–512.

[19] S. Birchfield and C. Tomasi, "Depth dicontinuities by pixel-to-pixel stereo," *Int. J. Comput. Vis.*, vol. 35, no. 3, pp. 269–293, 1999.

[20] G. Kraft and P. P. Jonker, "Real-time stereo with dense output by a simd-computed dynamic programming algorithm," in *Proc. Int. Conf. Parallel and Distributed Processing Techniques and Applications*, Las Vegas, NV, Jun. 2002, vol. III, pp. 1031–1036.

[21] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 11, pp. 1222–1239, Nov. 2001.

[22] V. Kolmogorov and R. Zabih, "Computing visual correspondence with occlusions using graph cuts," in *Proc. 8th Inf. Conf. Computer Vision*, 2001, vol. 2, pp. 508–515.

[23] S. Roy, "Stereo without epipolar lines: A maximum-flow formulation," *Int. J. Comput. Vis.*, vol. 34, pp. 147–161, 1999.

[24] L. Hong and G. Chen, "Segment-based stereo matching using graph cuts," in *Proc. IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, 2004, vol. 1, pp. 74–81.

[25] R. Szeliski and R. Zabih, "An experimental comparison of stereo algorithms," in *Proc. Int. Workshop Vision Algorithms*, 1999, pp. 1–19.

[26] J. P. Frisby, *Seeing: Illusion, Brain and Mind*. Oxford, U.K.: Oxford Univ. Press, 1979.

[27] Y. V. Venkatesh and N. Rishikesh, "Self-organizing neural networks based on spatial isomorphism for active contour modelling," *Pattern Recognit.*, vol. 33, pp. 1239–1250, 2000.

[28] T. Kohonen, *Self-Organization and Associative Memory*. Berlin, Germany: Springer, 1989.

[29] Y. V. Venkatesh, A. Jayakumar, and S. K. Raja, "On the application of a self-organizing neural network to stereo-image pair analysis," Tech. Rep, Electrical Engineering Dept., Indian Inst. Sci., Bangalore, 2005.

[30] S. Roy and I. J. Cox, "A maximum-flow formulation of the n-camera stereo correspondence problem," in *Proc. Int. Conf. Computer Vision*, Bombay, India, 1998, pp. 492–499.

[31] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1, pp. 7–42, 2002.

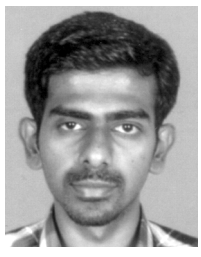
[32] B. L. Anderson and K. Nakayama, "Towards a general theory of stereopsis: Binoclular matching, occluding contours, and fusion," *Psych. Rev.*, vol. 101, pp. 414–445, 1994.

[33] K. Nakayama and S. Shimojo, "Davinci stereopsis: Depth and subjective occluding contours from unpaired image points," *Vis. Res.*, vol. 30, no. 11, pp. 1811–1825, 1990.

[34] N. Qian and Y. Zhu, "Physiological computation of binocular disparity," *Vis. Res.*, vol. 37, no. 13, pp. 1811–1827, 1997.

Y. V. Venkatesh, photograph and biography not available at the time of publication.

S. Kumar Raja, photograph and biography not available at the time of publication.



A. Jaya Kumar received the B.E. degree from Regional Engineering College, Surathkal, India, in 2003, and the M.Sc. (Res) degree from Indian Institute of Science, Bangalore, India, in 2005, both in electrical engineering. He is currently pursuing the Ph.D. degree in theoretical physics at the Raman Research Institute, Bangalore.

His interests include geometric problems in the fields of computer vision and physics.