SIMULATING DYNAMICAL SYSTEMS FOR EARLY VISION

Babette Dellen[1,2], Florentin Wörgötter[3]
[1]*Bernstein Center for Computational Neuroscience,Max-Planck Institute for Dynamics and Self-Organization, Bunsenstrasse 10, Göttingen, Germany*
[2]*Institut de Robòtica i Informàtica Industrial (CSIC-UPC), Llorens i Artigas 4-6, 08028 Barcelona, Spain*
[3]*Bernstein Center for Computational Neuroscience, University Göttingen, Bunsenstrasse 10, Göttingen, Germany*
bkdellen@bccn-goettingen.de, worgott@bccn-goettingen.de

Keywords: Early Vision, Stereo Matching, Energy Minimization, Dynamical Systems.

Abstract: We propose a novel algorithm for stereo matching using a dynamical systems approach. The stereo correspondence problem is first formulated as an energy minimization problem. From the energy function, we derive a system of differential equations describing the corresponding dynamical system of interacting elements, which we solve using numerical integration. Optimization is introduced by means of a damping term and a noise term, an idea similar to simulated annealing. The algorithm is tested on the Middlebury stereo benchmark.

1 INTRODUCTION

In stereo vision, 3D information is reconstructed from stereo image pairs, i.e. two images of the same scene taken from a different viewpoint. Algorithmic solutions to this problem are not only of interest for the field of computer vision [Scharstein and Szeliski, 2002], but also for related fields, such as computational neuroscience [Roe et al., 2007]. Different approaches have been compared in a study by Scharstein and Szeliski (2002). In general, we distinguish between local algorithms and methods based on global optimization. Local methods are mainly characterized by their matching cost computation and cost aggregation step, while global algorithms formulate a global energy function which is then minimized. This energy minimization problem is known to be NP hard. The algorithms are distinguished based on the minimization procedure used. Common methods are simulated annealing [Marroquin et al., 1987, Geman and Geman, 1984, Barnard, 1989], graph cuts [Scharstein and Szeliski, 2002, Boykov et al., 2001], and max flow [Roy, 1999]. If global optimization is reduced to independent scanlines, methods such as dynamic programming or scanline optimization can be used to compute a solution in polynomal time [Scharstein and Szeliski, 2002].

In this paper, we propose a novel framework for computing approximate solutions to the energy minimization problem on the example of early stereo vision. From the energy function, a system of ordinary differential equations, determining the temporal evolution of the system, can be derived. Each pixel represents a "mass point", moving along a single dimension with an amplitude encoding the disparity estimate (or label) of the pixel. Each mass is moving under the influence of a data force, which is derived from the image data, and interacts with its neighbors via an interaction force. The resulting system of differential equations is solved using a Runge Kutta method of 4th order with fixed step size. A damping force ensures that the dynamical systems settles at a stable state.

2 THE MODEL SYSTEM

2.1 Stereo Vision as Energy Minimization

The general framework we consider can be defined as follows. Let P be the set of pixels in an image. The goal is to find a disparity z_p for each pixel $p \in P$ which minimize a gobal energy

$$E(z_p) = E_{\text{data}}(z_p) + \sum_{q \in N(p)} E_{\text{int}}(z_p, z_q) \quad , \qquad (1)$$

where $N(p)$ is the neighborhood of pixel p. The data term E_{data} measures how well the disparity values are in agreement with the input data. The interaction term $\sum_{q\in N(p)} E_{\text{int}}(z_p, z_q)$ encodes the smoothing assumptions of the algorithm.

In this work, the data energy is derived from the stereo image I_{left} and I_{right}

$$E_{\text{data}}(z_p) = k|I_{\text{left}}(x_p + z_p, y_p) - I_{\text{right}}(x_p, y_p)| \quad (2)$$

using absolute differences and a parameter k.

We further assume symmetric interactions between two pixels p and q with

$$E_{\text{int}}(z_p, z_q) = f(z_p - z_q) \quad . \quad (3)$$

The function f will be specified later on.

2.2 Dynamical Systems Formulation

The energy function corresponds to a system of interacting elements moving under the influence of a data force

$$F_p^{\text{data}}(z_p) = -\nabla E_{\text{data}}(z_p) \quad , \quad (4)$$

where $\nabla E_{\text{data}}(z_p)$ is the gradient of the data potential, and an interaction force (on pixel p)

$$F_{p,q}^{\text{int}}(z_p, z_q) = -\nabla E_{\text{int}}(z_p, z_q) \quad . \quad (5)$$

A schematic of the model is shown in Fig. 1.

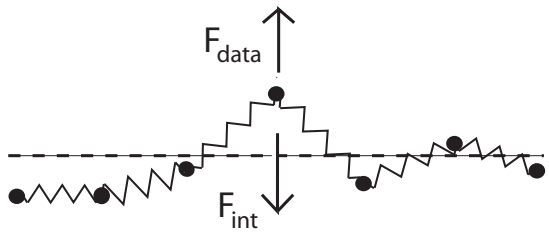


Figure 1: Schematic of the dynamical system. The pixels "mass points" are connected via elastic interaction forces F_{int}. The image data excerts a data force F_{data} on each mass point, which tends to move the mass towards the position which corresponds to a minimum data cost.

We define the interaction energy implicitly by choosing a discontinuity preserving interaction force with

$$F_{p,q}^{\text{int}}(z_p, z_q) = \kappa(d_{\max} - |z_p - z_q|)(z_p - z_q)/d_{\max} \quad (6)$$

if $|z_p - z_q| \leqq d_{\max}$ and zero otherwise. The parameter $d_{\max}$ defines the maximum disparity and κ determines the maximum amount of smoothing.

The dynamics of the system is described by a system of ordinary differential equations

$$dz_p/dt = v_p \quad (7)$$

$$dv_p/dt = F_p^{\text{data}}(d_p) + \tau - \gamma v_p - \sum_{q\in N(p)} F_{p,q}^{\text{int}}(z_p, z_q) \quad , \quad (8)$$

where τ is a noise term and γv_p a damping term with damping constant γ. These additional forces have been added to move the dynamical system towards a local minimum.

2.3 Finding a Local Minimum

The system of differential equations is solved using a fourth order Runge Kutta technique with a step size of 0.1, starting from random initial conditions. Cooling is introduced through the damping force and the noise term τ. With the course of time, we decrease the noise according to

$$\tau = p_r(n_i - t)/n_i \quad (9)$$

where n_i is the total number of iterations and t is the current iteration number. The number p_r is drawn from a Gaussian distribution with a standard deviation of 5 pixels. We further found it advantageous to decrease the smoothing parameter accordingly as well, such that

$$\kappa = \kappa_n(n_i - t)/n_i \quad . \quad (10)$$

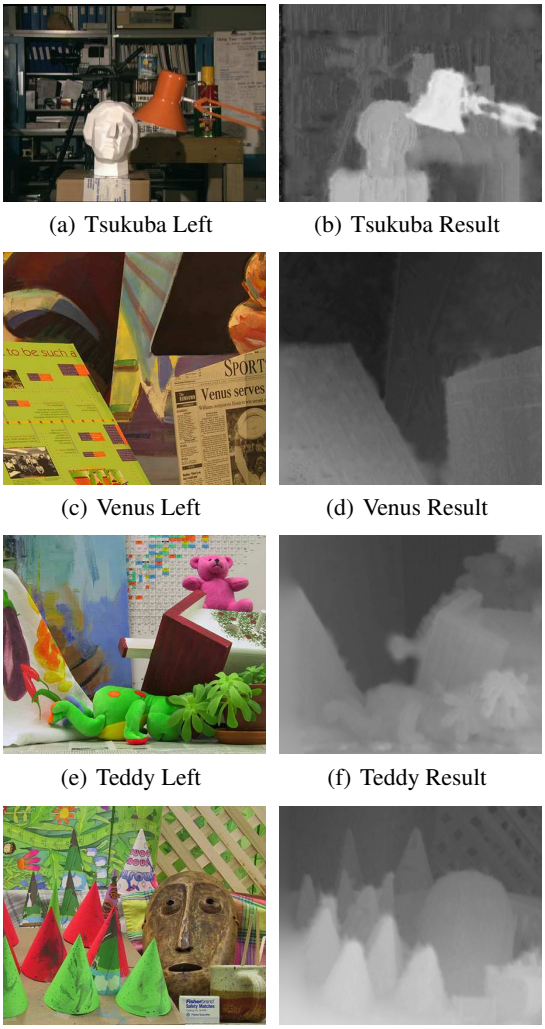
2.4 Boundary Conditions

The amplitude of the dynamical variable z_p is restricted to predefined disparity range. We realize this boundary conditions by including a potential barrier with $E(z_p) = c$ if $z_p > d_{max}$ or $z_p < 0$. The parameter c is chosen to be larger than the maximum absolute difference between image pixels. Further if during the computations $z_p > d_{max} + 1$, we push the value back to $z_p = d_{max} + 1$. The same strategy is used if $z_p < -1$. Then the value is pushed back to $z_p = -1$.

3 RESULTS

We evaluated the performance of the algorithm using the Middlebury stereo benchmark (www.middlebury.edu/stereo) [Scharstein and Szeliski, 2002], containing four stereo pairs, Tsukuba, Venus, Teddy, and Cones. The parameters were kept constant for all stereo pairs with $\kappa_n = 10$, $\gamma = 0.2$, and $k = 0.1$. On the grid, each pixel was allowed to interact with its left, right, up, and down nearest neighbor.

The results of the algorithm are presented in Fig. 2. Since the disparities are formulated as continuous variables, the algorithm returns subpixel disparities. The resulting disparity maps capture the basic structure of the scene. Depth discontinuities



(a) Tsukuba Left (b) Tsukuba Result

(c) Venus Left (d) Venus Result

(e) Teddy Left (f) Teddy Result

(g) Cones Left (h) Cones Result

Figure 2: Disparity results (b,d,f,h) for the Middlbury stereo data set (a,c,e,g). The algorithm returns a dense disparity map with captures the basic 3D structure of the scene. Near occlusion edges, however, errors are visible as can be seen for the lamp arm (b).

are mostly resolved, however, at regions with occlusions, the correct disparity at boundaries could not always be found, for example the lamp arm in Tsukuba (Fig. 2b), or for the Teddy stereo pair (Fig. 2f). Decreasing κ_n may decrease these undesired blurring effects. However, a lower smoothing parameter may also increase the convergence time.

We ranked the method the Middlebury stereo evaluation benchmark. In Fig. 3, a table of results for an error threshold of 0.75 pixels is shown. On average, the results of the method are comparable with those of other stereo algorithms such as dynamic programming and scanline optimization. In its current stage, the performance of the alorithm is inferior to graph cuts.

The algorithm was run for $n_i = 8000$ iterations, however, the result are usually stable after $4000 - 6000$ iterations. The computation for 1000 iterations take ≈ 3.4 min, using non-optimized code. This is in the range of computational times of other two-dimensional algorithms for energy optimization, i.e. graph cuts and simulated annealing.

4 CONCLUSIONS

4.1 Summary

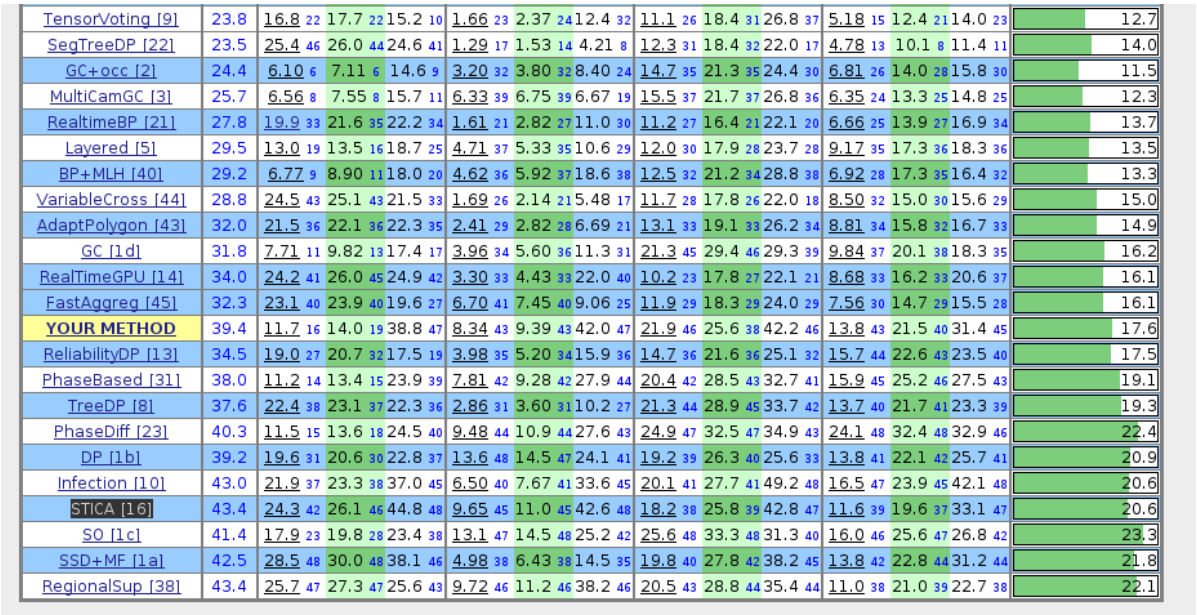
We proposed a dynamical systems approach to the energy minimization problem in early stereo vision. The stereo problem is first posed as an energy function minimization problem. Then, we derive a system of ordinary differential equations describing the dynamical system corresponding to the energy function. We explicitly calculate the development of the dynamical system using a Runge Kutta technique. The inclusion of a damping force ensures that the system converges to a local minimum. Overall, the algorithm delivers satisfying results for the images tested, using the Middlebury stereo database. The basic 3D structure of the scene could be captured correctly. Its performance is comparable to other approaches based on global optimization [Scharstein and Szeliski, 2002], except graph cuts [Boykov et al., 2001], which shows better results.

Energy functions for stereo vision have also been minimized by applying a gradient descent method to the associated Euler-Lagrange partial differential equation [Alvarez and Sánchez, 2000, Maier et al., 2003]. Since the solutions to the Euler-Lagrange equation are equivalent to Newton's laws of motion (in classical mechanics), possible relations to our method should be investigated in the future.

4.2 Future Work

In the future, the performance of the algorithm may be improved by modifying the parameters of the system, the energy function itself, or by increasing the number of nearest neighbors. The performance could also be improved by increasing the number of iterations. However, this would also increase the computation time.

Occlusions, which occur in almost all images, are not handled by the algorithm. This causes errors in the disparity estimation in particular near object boundaries. These problems could be decreased by incor-



TensorVoting [9]	23.8	16.8	22	17.7	22	15.2	10	1.66	23	2.37	24	12.4	32	11.1	26	18.4	31	26.8	37	5.18	15	12.4	21	14.0	23	12.7
SegTreeDP [22]	23.5	25.4	46	26.0	44	24.6	41	1.29	17	1.53	14	4.21	8	12.3	31	18.4	32	22.0	17	4.78	13	10.1	8	11.4	11	14.0
GC+occ [2]	24.4	6.10	6	7.11	6	14.6	9	3.20	32	3.80	32	8.40	24	14.7	35	21.3	35	24.4	30	6.81	26	14.0	28	15.8	30	11.5
MultiCamGC [3]	25.7	6.56	8	7.55	8	15.7	11	6.33	39	6.75	39	6.67	19	15.5	37	21.7	37	26.8	36	6.35	24	13.3	25	14.8	25	12.3
RealtimeBP [21]	27.8	19.9	33	21.6	35	22.2	34	1.61	21	2.82	27	11.0	30	11.2	27	16.4	21	22.1	20	6.66	25	13.9	27	16.9	34	13.7
Layered [5]	29.5	13.0	19	13.5	16	18.7	25	4.71	37	5.33	35	10.6	29	12.0	30	17.9	28	23.7	28	9.17	35	17.3	36	18.3	36	13.5
BP+MLH [40]	29.2	6.77	9	8.90	11	18.0	20	4.62	36	5.92	37	18.6	38	12.5	32	21.2	34	28.8	38	6.92	28	17.3	35	16.4	32	13.3
VariableCross [44]	28.8	24.5	43	25.1	43	21.5	33	1.69	26	2.14	21	5.48	17	11.7	28	17.8	26	22.0	18	8.50	32	15.0	30	15.6	29	15.0
AdaptPolygon [43]	32.0	21.5	36	22.1	36	22.3	35	2.41	29	2.82	28	6.69	21	13.1	33	19.1	33	26.2	34	8.81	34	15.8	32	16.7	33	14.9
GC [1d]	31.8	7.71	11	9.82	13	17.4	17	3.96	34	5.60	36	11.3	31	21.3	45	29.4	46	29.3	39	9.84	37	20.1	38	18.3	35	16.2
RealTimeGPU [14]	34.0	24.2	41	26.0	45	24.9	42	3.30	33	4.43	33	22.0	40	10.2	23	17.8	27	22.1	21	8.68	33	16.2	33	20.6	37	16.1
FastAggreg [45]	32.3	23.1	40	23.9	40	19.6	27	6.70	41	7.45	40	9.06	25	11.9	29	18.3	29	24.0	29	7.56	30	14.7	29	15.5	28	16.1
YOUR METHOD	39.4	11.7	16	14.0	19	38.8	47	8.34	43	9.39	43	42.0	47	21.9	46	25.6	38	42.2	46	13.8	43	21.5	40	31.4	45	17.6
ReliabilityDP [13]	34.5	19.0	27	20.7	32	17.5	19	3.98	35	5.20	34	15.9	36	14.7	36	21.6	36	25.1	32	15.7	44	22.6	43	23.5	40	17.5
PhaseBased [31]	38.0	11.2	14	13.4	15	23.9	39	7.81	42	9.28	42	27.9	44	20.4	42	28.5	43	32.7	41	15.9	45	25.2	46	27.5	43	19.1
TreeDP [8]	37.6	22.4	38	23.1	37	22.3	36	2.86	31	3.60	31	10.2	27	21.3	44	28.9	45	33.7	42	13.7	40	21.7	41	23.3	39	19.3
PhaseDiff [23]	40.3	11.5	15	13.6	18	24.5	40	9.48	44	10.9	44	27.6	43	24.9	47	32.5	47	34.9	43	24.1	48	32.4	48	32.9	46	22.4
DP [1b]	39.2	19.6	31	20.6	30	22.8	37	13.6	48	14.5	47	24.1	41	19.2	39	26.3	40	25.6	33	13.8	41	22.1	42	25.7	41	20.9
Infection [10]	43.0	21.9	37	23.3	38	37.0	45	6.50	40	7.67	41	33.6	45	20.1	41	27.7	41	49.2	48	16.5	47	23.9	45	42.1	48	20.6
STICA [16]	43.4	24.3	42	26.1	46	44.8	48	9.65	45	11.0	45	42.6	48	18.2	38	25.8	39	42.8	47	11.6	39	19.6	37	33.1	47	20.6
SO [1c]	41.4	17.9	23	19.8	28	23.4	38	13.1	47	14.5	48	25.2	42	25.6	48	33.3	48	31.3	40	16.0	46	25.6	47	26.8	42	23.3
SSD+MF [1a]	42.5	28.5	48	30.0	48	38.1	46	4.98	38	6.43	38	14.5	35	19.8	40	27.8	42	38.2	45	13.8	42	22.8	44	31.2	44	21.8
RegionalSup [38]	43.4	25.7	47	27.3	47	25.6	43	9.72	46	11.2	46	38.2	46	20.5	43	28.8	44	35.4	44	11.0	38	21.0	39	22.7	38	22.1

Figure 3: Ranking from the Middlebury stereo database for an error threshold of 0.75 pixels. The results are comparable to those obtained using dynamic programming or scanline optimization.

porating an occlusion detection method to our algorithm [Egnal and Wildes, 2002].

The speed of the algorithm could be improved using a coarse-to-fine approach. A parallel implementation would be feasible as well because of the local character of the algorithm, e.g. using a graphics processing unit.

The proposed model might be of interest for models of human stereo vision. Interacting neuronal elements might be able to utilize damping and/or noise to optimize their responses.

ACKNOWLEDGEMENTS

This work has received support from the BMBF funded BCCN Göttingen and the EU Project Drivsco under Contract No. 016276-2.

REFERENCES

Alvarez, L. and Sánchez, J. (2000). 3-d geometry reconstruction using a color image stereo pair and partial differential equations. *Cuadernos del Instituto Universitario de Ciencias y Tecnologas Ciberneticas*, 6:1–26.

Barnard, S. T. (1989). Stochastic stereo matching oer scale. *International Journal of Computer Vision*, 3(1):17–32.

Boykov, Y., Veksler, O., and Zabih, R. (2001). Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(11):1222–1239.

Egnal, G. and Wildes, R. (2002). Detecting binocular half-occlusions: Empirical comparisons of five approaches. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(8):1122–1133.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distribution, and the baysian restoration of images. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Maier, D., Rössle, A., Hesser, J., and Männer, R. (2003). Dense disparity maps respecting occlusions and object separation using partial differential equations. In Sun, C., Talbot, H., Ourselin, S., and Adriaanen, T., editors, *Proc. VIIth Digital Image Computing: Techniques and applications*, pages 613–622.

Marroquin, J., Mitter, S., and Poggio, T. (1987). Probabilistic solution of ill-posed problems in computational vision. *Journal of the American Statistical Association*, 82(397):76–89.

Roe, A. W., Parker, A. J., Born, R. T., and DeAngelis, G. C. (2007). Disparity channels in early vision. *Journal of Neuroscience*, 27(44):11820–31.

Roy, S. (1999). Stereo without epipolar lines: A maximum flow formulation. *International Journal of Computer Vision*, 34(2/3):147–161.

Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42.