FAST COMPUTATION OF DENSE STEREO CORRESPONDENCES BY STOCHASTIC SAMPLING OF MATCH QUALITY[1]

Thayne R. Coffman†, Alan C. Bovik**

* The University of Texas at Austin, Austin, TX, 78712, USA
† 21st Century Technologies, Austin, TX, 78759, USA

ABSTRACT

We present a method for computing dense stereo correspondences in calibrated monocular video by iteratively and stochastically sampling match quality values in the disparity search space. Most existing methods exhaustively compute local correspondence quality before searching for a globally optimal solution. Instead, we iteratively refine a correspondence estimate by perturbing it with random noise and formulating an *influence* at each sample based on the perturbation and its effect on correspondence match quality. Local influence is aggregated to recover consistent trends in match quality caused by the piecewise-continuous structure of the scene. Correspondence estimates for a given frame pair are seeded with the estimates from the previous frame pair, allowing convergence to occur across multiple frame pairs.

Index Terms—Stereo vision, computational geometry, stochastic approximation, recursive estimation, simulated annealing

1. INTRODUCTION

Computational stereo techniques estimate the 3D structure of a scene by analyzing two or more 2D images captured from different viewpoints. Sparse stereo systems estimate depth to the scene at a small number of specific points. Dense stereo systems instead generate depth estimates at all pixels in a region of the imagery. The core problem in dense computational stereo is computing the correspondence between all the pixels in the two (or more) images being analyzed [5]. Finding efficient and robust solutions to this problem remains an active research field.

This paper presents new techniques for computing dense stereo correspondence. In contrast with most methods, we explore the search space stochastically and non-exhaustively. We present a formulation of the *influence* that search space samples should exert on their local neighbors and methods for aggregating those influences such that the stochastic search converges towards the true solution. We also describe a new approach for seeding our stochastic search with previous search results, to reduce convergence time and improve accuracy.

The work presented here is part of a larger system that is directed towards real-time dense reconstruction of urban scenes containing both static objects and moving vehicles. By modeling both static and moving objects and operating in real time, the final system will significantly improve the situational awareness of its users and will enable more independent autonomous unmanned vehicle operation. The scene is imaged by a single calibrated visible-light camera mounted on a low-flying unmanned aerial vehicle or other surveillance platform. Figure 1 shows a frame of characteristic input imagery alongside a set of elevation estimates. Our approach assumes full knowledge of camera position and orientation, but without control over either.

Figure 1: Example imagery (left) and elevation estimates (right).

Relevant surveys of this area can be found in [3] and [8]. The most relevant approaches are *cooperative techniques*, which provided both very early and very recent advances in the field [7][10], and approaches based on *simulated annealing* [4] and *microcanonical annealing* [1][2]. With the exception of annealing-based approaches, the vast majority of approaches exhaustively compute match quality at all possible correspondences within the confines of the epipolar constraint, before attempting to find a globally consistent solution.

2. APPROACH

Our approach is driven by two opinions. First, we believe that dense stereo correspondence solutions can be computed faster by stochastically sampling the match quality instead of computing it for every possible pixel pairing. Second, we believe that the piecewise continuity constraint and continuity of matching likelihood constraint [6] can be used to skip exploration of portions of the disparity search space.

[1] This work was supported by USAF contracts FA8651-04-C-0233 and FA8651-05-C-0117. This paper has been cleared for release by AFRL; PA Clearance Number: AAC/PA 01-18-07-027.

For each iteration of our cooperative approach, we introduce random noise into each pixel's disparity estimate, compute the match quality at the perturbed location, and compute the influence that the new sample should exert on the local solution region towards or away from the perturbed disparity estimate. Influences are then aggregated under the argument that perturbations towards the correct solution will generate more consistent and larger influences than superfluous improvements in match quality.

We first pair frames from the video stream in order to maintain a constant ratio of stereo baseline to minimum depth to scene, following [9]. We then rectify the images to align the principal camera axes. (This is not a full projective rectification into a standard stereo geometry.) The dense correspondence approach computes disparity magnitude at each pixel, which is assumed to be along the known epipolar direction. Following correspondence matching, we triangulate to estimate range to the scene at each pixel, and then convert the dense point clouds into abstract surface models for the final output of the static modeling system. Static modeling byproducts are used to detect and track any moving vehicles in the scene, which are then identified and modeled by different means.

2.1. Stochastic search of disparity space

In our target application, we know the direction of each pixel's epipolar line because our camera positions and orientations are known. As a result, we can pre-compute unit vectors along the epipolar directions and represent the estimated correspondences by a scalar-valued disparity magnitude field *D(r,c)*. Infinite depth results in zero disparity as a result of our rectification scheme.

We bound our search by conservative assumptions on the minimum and maximum elevation of elements in the scene. Given our typical viewing angles, this gives tighter bounds on disparity at each pixel than bounds on depth. As a result, however, each pixel has different disparity bounds. We normalize *D(r,c)* over the allowed bounds at each pixel, yielding a normalized disparity on the range [0.0,1.0]. This normalized disparity is also equal to a "percent elevation" over the assumed minimum and maximum elevation range.

Most approaches would compute the pixel-wise match quality *q(r,c,D(r,c))* for every row, column, and every allowable disparity magnitude *D(r,c)* before optimizing to find a global solution. This requires *O(RCL)* computations for an *R* by *C* image with *L* possible disparity values. Instead, we iteratively refine an estimate $D_i(r,c)$ by randomly perturbing it with an additive noise $\Delta_{di}(r,c) \in [-\delta_{\max}, \delta_{\max}]$, computing the match quality of the perturbed disparities, and computing an *influence* $I^*_i(r,c)$ to be exerted by each new sample $\tilde{q}_i(r,c,\tilde{D}_i(r,c))$. The sign and magnitude of the influence are based on the sign and magnitude of the random perturbation, as well as its resulting effect on the match quality at that pixel. These pixel-wise influences are then aggregated over a local region *A* and applied to $D_i(r,c)$ to move it towards the correct solution.

We thus avoid exhaustively computing *q(r,c,D(r,c))* by instead stochastically sampling it and using local aggregation to extract consistent trends in match quality caused by the piecewise-continuous structure of the scene. Consistent trends are captured and inconsistent noise is rejected by our aggregated influence $I_i(r,c)$. As described in Section 2.4, the approach also lets us refine our disparity estimate over multiple frame pairs in a manner similar to recursive estimation approaches.

Table 1: Iterative stochastic correspondence search

1. For each new frame, precompute epipolar directions and disparity bounds, and initialize the disparity magnitude estimate $D_0(r,c)$ (Section 2.4).
2. For each stage, perform *N* iterations with noise magnitude $\delta_{\max}$ and aggregation neighborhood size *A* (Section 2.3):
   a. Compute match quality $q_i(r,c,D_i(r,c))$.
   b. Generate independent uniformly distributed random perturbations $\Delta_{di}(r,c) \in [-\delta_{\max}, \delta_{\max}] \cap [-D_i(r,c), 1-D_i(r,c)]$ at each pixel and set $\tilde{D}_i(r,c) = D_i(r,c) + \Delta_{di}(r,c)$.
   c. Compute perturbed match quality $\tilde{q}_i(r,c,\tilde{D}_i(r,c))$.
   d. Compute pixel-wise influence (Section 2.2): $I^*_i(r,c) = f(q_i(r,c), \tilde{q}_i(r,c), \Delta_{di}(r,c))$
   e. Aggregate influence over a local region (Section 2.2): $I_i = g(I^*_i, A)$
   f. $D_{i+1}(r,c) = D_i(r,c) + I_i(r,c)$
   g. Smooth depth estimates $D_{i+1}(r,c)$ by averaging over aggregation neighborhood size *A*.

Our approach is explained by the pseudocode given in Table 1. It is centered around the concept of an influence, $I_i(r,c)$, which we define in Section 2.2 below. For now, we ask the reader to postulate that we can define influence such that (when aggregated in a local neighborhood) influence tends towards zero when the correct (optimal) solution is reached, and when the current estimate $D_i(r,c)$ is suboptimal the influence tends to have sign and magnitude such that adding influence to $D_i(r,c)$ will move it closer to the true solution.

2.2. Influence formulation and aggregation

Consider Figure 2, which shows a region of random disparity perturbations, resulting changes in match quality, resulting pixel-wise influences, and finally aggregated influences. The pre-perturbation disparity estimates $D_i(r,c)$ are typically smooth, with smoothly-varying match qualities q_i. Randomly perturbing the disparity estimates introduces "noise" into $\tilde{q}_i$. The job of the pixel-wise influence is to extract structure from $\Delta_{qi} = \tilde{q}_i - q_i$. At a first level, pixel-wise influence should be positive for perturbations that increase both disparity and match quality, and negative for perturbations that decrease both disparity and match quality. When the perturbation *decreases* match quality, pixel-wise influence should either be zero or directed away from the perturbation. Randomness ensures that some perturbations will increase disparity and some will decrease disparity, and likewise we expect that some match quality values will increase and some will decrease. By selectively giving influence to only those that improve the solution and aggregating over a local neighborhood, we can iteratively refine our estimates towards a better solution.

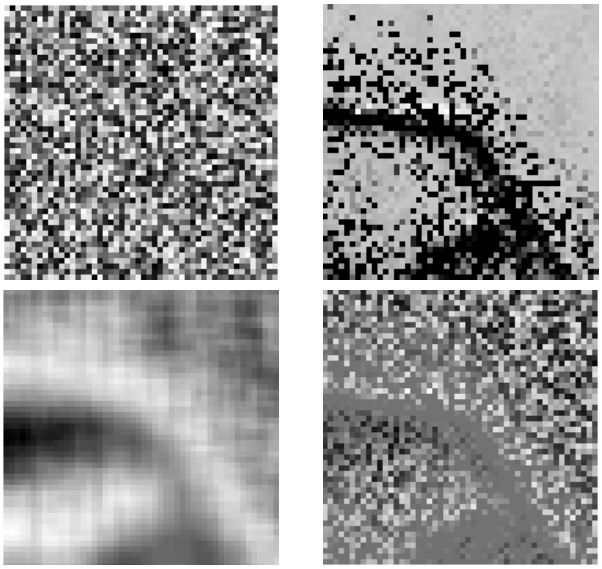

Figure 2: Influence formulation and aggregation. (Clockwise from top left) Disparity perturbations, match quality perturbations, pixel-wise influences, and aggregated influences.

We define influence in the following way:

$$I^* = \Delta_{Di}\left(w_1 + w_2\right)/2 \quad (1)$$

$$w_1 = \max(C_1\tilde{q}_i - (C_1 - 1), 0), \text{ with } C_1 \approx 10 \quad (2)$$

$$w_2 = \Delta_{qi} / \max_{r,c} \Delta_{qi} \quad (3)$$

The definition of I^* in (1) attempts to combine the two general approaches we have explored to date. The max operation in (2) ensures that only high perturbed quality values are given influence. The normalization in (3) ensures that w_2 is bound from above by 1.0 and will achieve that maximum for at least one pixel in each iteration.

Aggregated influence should be aligned with perturbations that move the disparity estimate towards the correct solution and be nearly zero for perturbations away from the correct solution. Influence magnitudes should also be balanced such that they do not shrink too quickly as match qualities near 1.0, but large influences do not "overshoot" the optimal solution and cause undesirable oscillations or instabilities in the estimates.

A set of desirable characteristics can also be defined for aggregation. Generally, aggregation is used to remove noise from pixel-wise influences while preserving consistent trends. Our typical aggregation mechanism has been simply to average pixel-wise influences over a local region of size A x A. We have also explored non-isotropic filters that prevent smearing near occlusion boundaries, with promising initial results but at a significant runtime cost. Averaging and non-isotropic smoothing impose smoothness or piecewise-smoothness constraints (respectively) on the solutions.

2.3. Search schedule

The iterations in the stochastic search are performed according to a schedule analogous to an annealing schedule in simulated annealing, or a training schedule in a self-organizing feature map. These schedules, which define values for the parameters $\delta_{\max}$, A, and N for each stage, are modified and tuned often to explore alternatives for improved performance.

Independent of fine-tuning, these search schedules have some global characteristics that are motivated by the need for them to be as robust as possible, and to recover details in the scene. Different input imagery resolutions may require different parameters (specifically values for the aggregation neighborhood size, A). As a result, A is specified as a fraction of the row and column resolution of the input, typically in the range of 1/30 to 1/120 for NTSC imagery. Maximum perturbation magnitude, $\delta_{\max}$, is given as a percentage of the full disparity search range (defined by assumptions on minimum and maximum scene elevation), which will have different absolute sizes at each pixel. Typical values for $\delta_{\max}$ range from 50% to 10%.

Search schedules have few stages, with larger perturbations and aggregation neighborhoods in early stages and smaller values in the later stages. This is common for searches which seek to both minimize convergence time and retain detail in the solution. We use a unique schedule for the first frame pair that emphasizes larger perturbations and aggregations. Searches in subsequent frame pairs are seeded with an estimated solution (Section 2.4). We perform a fixed number of iterations in each stage, which gives better control over the tradeoff between accuracy and runtime.

2.4. Seeding the computation

Since we are generating dense depth estimates for each frame and we know the camera's motion from one frame to the next, we can initialize our search with the results from the previous frame pair, instead of estimating the solution from scratch. Searches are seeded by projecting the previous estimates to where they would appear in the new frame of reference. In order to properly treat occlusions introduced by the camera motion, this process must include Z-buffering of the depth estimates.

This seeding introduces aspects of recursive estimation, although our approach is not explicitly formulated in those terms. As a result of carrying over past estimates to new frame pairs, we do not need to compute a perfect solution as soon as a new scene element is visible – we can instead converge to the solution over time. This lets us reduce the iterations for each frame pair and use search schedules with smaller perturbations and aggregation neighborhoods, to retain detail in the solution.

3. RESULTS

Figure 3 shows an example image frame from a stereo pair, and the resulting reconstructed point cloud. The point cloud is shown colored according to the original imagery but viewed from a slightly different angle than either of the two original images. The approach is recovering the gross structure of the scene well, although work remains to capture small details.

Performance is measured on an NTSC dataset provided by the Air Force Research Laboratory. Extrinsic camera parameter information is available in Global Positioning System (GPS) coordinates and Euler angles. Positions are known for a sparse set of building corners, fiducials, and stationary vehicles in the scene. Finding characteristic test imagery with dense ground truth is an ongoing challenge.

Figure 4 shows sparse reconstruction accuracy vs. stereo baseline ratio (the ratio of stereo baseline to minimum depth [9]).

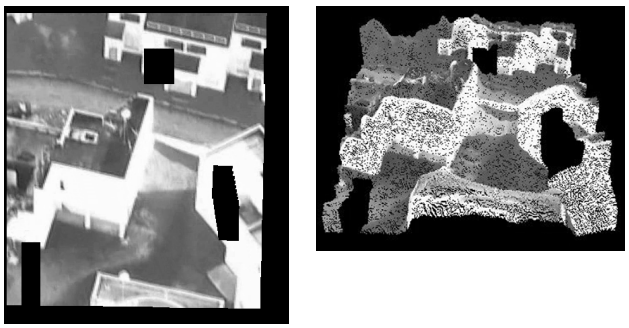

Figure 3: Example input imagery (left) and reconstructed dense point cloud (right), viewed from a slightly different angle.

A family of curves is given. Because our camera parameter knowledge is imperfect, a scene element may appear at a pixel other than the expected projection of its ground truth location. To compensate, we measure distance to the closest reconstructed point within a small radius in pixels. We believe that a radius of 5 pixels (the middle curve) is reasonable. Given this, we estimate our reconstruction accuracy against sparse ground truth to be approximately ±2m, at stereo baseline ratios of 0.04 and 0.17-0.20. Our evaluation sequences have depths of 150-225m, so this corresponds to slightly over 1% error. We have also tested our approach on half-resolution imagery with no reduction in accuracy, and on 10fps video with minimal reduction in accuracy. As a result, we believe that adequate end-system performance could be achievable on less-than-NTSC video resolutions and rates.

4. DISCUSSION

This paper presents a new approach for computing dense stereo correspondences by iteratively and stochastically sampling match quality values in the disparity search space, as well as a formulation of the *influence* each new match quality sample exerts on its local region of the solution. We believe that dense correspondence solutions can be computed faster if pixel-wise match quality is sampled instead of exhaustively computed and the trends in match quality values caused by the piecewise continuity constraint and continuity of matching likelihood constraint are used to guide a stochastic search of disparities.

Annealing approaches evaluate perturbed estimates against an explicit objective function that accounts for both local and global characteristics. In contrast, our approach evaluates the effects of perturbations at each pixel independently and decides whether to accept or reject the perturbations based entirely on single-pixel effects. We then rely on aggregation over local neighborhoods to extract consistent trends in the effects of those perturbations, reject spurious results, and move the solution towards the global optimum.

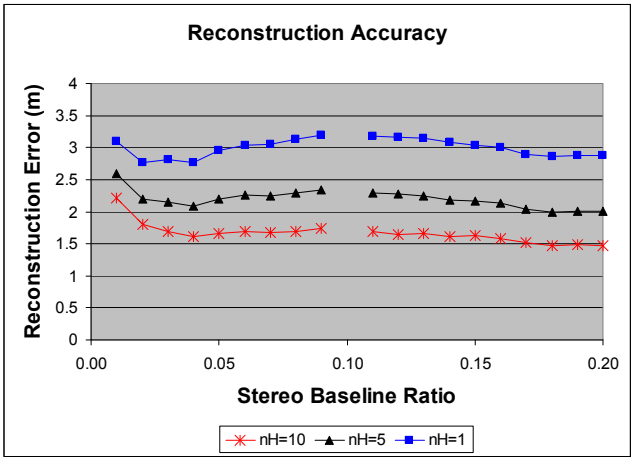

Figure 4: Average reconstruction accuracy measured against sparse ground truth, vs. stereo baseline ratio. A family of curves is given to account for imperfect camera parameter knowledge.

5. REFERENCES

[1] S. Barnard, "Stochastic stereo matching over scale," *Int. J. of Computer Vision*, Vol. 3, No. 1, pp. 17-32, 1989.

[2] M. Creutz, "Microcanonical Monte Carlo Simulation," *Phys. Review Letters*, Vol. 50, No. 9, pp. 1411-1414, 1983.

[3] U.R. Dhond and J.K. Aggarwal, "Structure from stereo – a review," *IEEE Trans. on Systems, Man, and Cybernetics*, Vol. 19, No. 6, pp. 1489-1510, 1989.

[4] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. PAMI*, Vol. 6, No. 6, pp. 721-741, 1984.

[5] H. Hirschmuller, "Improvements in real-time correlation-based stereo vision," *Proc. IEEE Workshop on Stereo and Multi-Baseline Vision*, 2001.

[6] Y. Kim and J.K. Aggarwal, "Positioning 3-D objects using stereo images," *IEEE Journal of Robotics and Automation*, Vol. 3, No. 4, pp. 361-373, 1987.

[7] D.C. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, Vol. 194, pp. 283-287, 1976.

[8] D. Scharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms," *Int. J. of Computer Vision*, Vol. 47, pp. 7-42, 2002.

[9] R. Vidal and J. Oliensis, "Structure from Planar Motions with Small Baselines," *Proc. ECCV*, pp. 383-398, 2002.

[10] C.L. Zitnick and T. Kanade, "A Cooperative Algorithm for Stereo Matching and Occlusion Detection," *IEEE Trans. PAMI*, Vol. 22, No. 7, pp. 675-684, 2000.