Efficient Stereoscopic Ranging via Stochastic Sampling of Match Quality

Thayne Richard Coffman, *Senior Member, IEEE*, and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*—We present an efficient method that computes dense stereo correspondences by stochastically sampling match quality values. Nonexhaustive sampling facilitates the use of quality metrics that take unique values at noninteger disparities. Depth estimates are iteratively refined with a stochastic cooperative search by perturbing the estimates, sampling match quality, and reweighting and aggregating the perturbations. The approach gains significant efficiencies when applied to video, where initial estimates are seeded using information from the previous pair in a novel application of the Z-buffering algorithm. This significantly reduces the number of search iterations required. We present a quantitative accuracy evaluation wherein the proposed method outperforms a microcanonical annealing approach by Barnard [2] and a cooperative approach by Zitnick and Kanade [27], while using fewer match quality evaluations than either. The approach is shown to have more attractive memory usage and scaling than alternatives based on exhaustive sampling.**

***Index Terms*—Computational geometry, cooperative stereo, recursive estimation, simulated annealing, stereo vision, stochastic approximation.**

I. Introduction

After more than 30 years of research, intensive effort is still being applied to improve computational stereo techniques that reconstruct dense scene structure estimates from stereo or monocular imagery. The core problem is to determine the correspondences between all the pixels in two (or more) images being analyzed. This computation, which at its root is based on a measure of *local match quality*, remains a challenge, and it accounts for the majority of complexity and runtime in computational stereo approaches.

We present a new method, named quality-efficient stochastic sampling (QUESS), which reduces the number of match quality computations required to accurately estimate dense stereo correspondences from calibrated monocular video. Most approaches exhaustively compute the match qualities of all potential correspondences. Instead, we apply a stochastic and cooperative search in the solution space. This approach reduces the number of match quality evaluations and facilitates the use of more complex quality metrics, as well as metrics defined on noninteger depth or disparity domains (for which exhaustive search is impossible). QUESS iteratively applies a simple formulation of local and aggregated *influences* which, together with techniques for seeding depth estimates from the previous frame, enables an efficient stochastic and cooperative search for dense stereo correspondences in calibrated video. It is motivated by passive aerial modeling applications, although it can be applied to other related problems.

Following a brief background discussion in Section II, Section III describes the approach as it is applied to standard-geometry stereo pairs. Section IV describes extensions that enable use on calibrated monocular video. Section V presents analyses of accuracy, number of match quality evaluations, scalability, runtime, and memory usage. Comparisons against an early stochastic approach by Barnard [2] and a cooperative approach by Zitnick and Kanade [27] are also presented. We end with conclusions in Section VI.

II. Background

The topic of automated stereo reconstruction still lacks a robust and deployable general solution. A number of open research problems remain. Runtime and efficiency continue to be challenges, as well as finding match quality metrics that are robust to image quality, lighting, and perspective changes. Robustness to camera path (in single-camera stereo) and scene orientation are also issues.

Aerial modeling from calibrated monocular video has received somewhat less attention than other stereo applications and lacks a generally applicable solution. A single camera follows an aerial platform's known but independently controlled path, with position and orientation changing incrementally between frames. The stereo geometry is nonstandard and constantly changing, and stereo frame pairings must be selected from a set of buffered frames. Intrinsic and extrinsic camera parameter values are available. Expected characteristics include large absolute ranges (hundreds or thousands of meters), large absolute disparities (tens or hundreds of pixels), and large disparity ranges. Approaches must address complex and uncontrolled outdoor scenes with moving objects, and be robust to uncontrolled lighting and other imaging artifacts. A reliable solution to this challenging problem would enable a wide variety of applications in the commercial, government, and military domains.

A substantial body of related work exists. Relevant surveys of sparse stereo approaches can be found in [7], dense stereo in [4]

Manuscript received March 21, 2009; revised September 22, 2009. First published October 20, 2009; current version published January 15, 2010. This work was supported in part by the U.S. Air Force under contracts FA8651-04-C-0233 and FA8651-05-C-0117. This paper has been approved for public release by AFRL, PA Case Number 96ABW-2009-0122. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Lina J. Karam.

T. R. Coffman is with 21st Century Technologies, Austin, TX 78759 USA, and also with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: tcoffman@21technologies.com).

A. C. Bovik is with the Laboratory for Image and Video Engineering, Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: bovik@ece.utexas.edu).

Digital Object Identifier 10.1109/TIP.2009.2035002

and [22], and multiview stereo in [23]. A few techniques are of particular relevance. *Cooperative techniques* [17], [27] repeatedly apply local (and/or nonlocal) effects to iterate towards a solution. Approaches using *simulated annealing* [9] or *microcanonical annealing* [2], [5] apply stochastic sampling and stochastic optimization. Beyond these, however, the use of either stochastic sampling or stochastic optimization is very rare in the literature. *Multiview stereo* systems simultaneously process a sequence of many images in nonstandard stereo geometries. Also relevant are systems targeting real-time operation (see [4]). The system in [14] is unique in its use of an active, foveating and vergent stereo geometry.

These approaches provide inspiration, but they each have disadvantages that require recombination and extension. Cooperative techniques typically compute local match quality exhaustively at integer-valued disparities. As a result, the match quality metric is evaluated many times, requiring considerable amounts of memory and runtimes which change with the camera and scene geometries. Stochastic search approaches avoid exhaustive sampling of all possible local solutions, but they can be slow to converge and still typically quantize disparity values. Real-time systems [13], [26] often require unattractive assumptions, such as a fixed geometry (yielding an insufficient stereo baseline for aerial modeling) or specialized hardware.

QUESS combines existing techniques with new methods yielding particular advantages. Iterative cooperative processing allows straightforward control of runtimes and facilitates initialization of estimates using results from previous frame pairs. Directly estimating real-valued depths allows the use of quality metrics that are not constrained to integer disparity values [24]. Stochastic sampling exploits piecewise continuity and continuity of matching likelihood constraints [15] to greatly reduce the search space, allowing the use of more complex local match quality metrics (see [11]) while maintaining acceptable runtimes. QUESS is cooperative and stochastic, combining the advantages of both.

III. Basic Two-Frame Approach

Here, we discuss the representation of the estimated quantities, the core stochastic cooperative search and the local and aggregated *influences* on which it is based.

A. Definitions and Representations

Consider two 2-D images $I_A(i,j)$ and $I_B(i,j)$, with image I_A defined as the reference image. The images are assumed to be in a standard stereo geometry with known (or assumed) stereo baseline and intrinsic parameters. A scalar floating-point depth $\hat{D}(i,j)$ can be estimated at each pixel based on the *local match quality function* $Q(i,j,\hat{d};I_A,I_B)$, which varies with depth estimate $\hat{d}$. Depth estimates can be converted to and from equivalent disparities as required (e.g., to compute Q or for evaluation in disparity-based frameworks).

TABLE I
Overview of the QUESS Approach

For each search stage
1. Initialize depth estimates $\hat{D}_0(i,j)$
2. For each iteration n=1..N
   a. Add noise to get perturbed estimate $\tilde{D}_n(i,j)$
   b. Sample Q at perturbed depth estimate
   c. Compute local influence $J_n^*(i,j)$ from Q samples
   d. Filter $J_n^*(i,j)$ to get aggregated influence $J_n(i,j)$
   e. $\hat{D}_n(i,j) = \hat{D}_{n-1}(i,j) + J_n(i,j)$
   f. Smooth depth estimates $\hat{D}_n(i,j)$

Direct depth estimation contrasts with estimating integer disparity values and postprocessing to recover sub-pixel disparities or depths. Estimating floating-point depth directly is preferable in situations with large depth ranges and significant spatial variations. It also avoids quantization and supports match quality metrics defined on continuous domains.

B. Stochastic Cooperative Search

An overview of QUESS is given in Table I. Depth estimates are iteratively refined with a stochastic cooperative search. QUESS perturbs the depth estimates, reweights perturbations using *local influence* computed from their effects on Q, and adds *aggregated influence* to the estimates to incrementally move them towards a better solution. The search is guided by a schedule analogous to those used in simulated annealing.

Depth estimates at each pixel are initialized from the previous search stage, previous frame, or from a uniform distribution over the bounds $\hat{D}_{\min}(i,j)$ and $\hat{D}_{\max}(i,j)$. Depth bounds vary pixel by pixel and are defined using any prior knowledge about the scene (including disparity bounds).

In each iteration $n = 1 \ldots N$, random noise $\Delta_{D_n}(i,j)$ is added to the previous depth estimate $\hat{D}_{n-1}(i,j)$ to form a candidate depth estimate $\tilde{D}_n(i,j) = \hat{D}_{n-1}(i,j) + \Delta_{D_n}(i,j)$. Q is evaluated at the candidate to compute a new sample, $\tilde{q}_n(i,j) = Q(i,j,\tilde{D}_n(i,j);I_A,I_B)$.

The noise $\Delta_{D_n}(i,j)$ added to each depth estimate is uniformly distributed, subject to two constraints. The first constraint is a maximum depth perturbation magnitude $\delta_{\max} \in [0,1]$ relative to the pixel-specific depth bounds

$$\frac{||\Delta_{D_n}(i,j)||}{\hat{D}_{\max}(i,j) - \hat{D}_{\min}(i,j)} \leq \delta_{\max}. \tag{1}$$

By gradually reducing $\delta_{\max}$ Q samples in later iterations are forced to be closer to current estimates. The second constraint limits $\Delta_{D_n}(i,j)$ so the perturbed estimate $\tilde{D}_n(i,j)$ remains within bounds; see (2), shown at the bottom of the page.

Constraint (2) is necessary because even if $\delta_{\max}$ is small, (1) may not prevent $\tilde{D}_n(i,j)$ from falling outside the allowable

$$\Delta_{D_n}(i,j) \in \left[\hat{D}_{\min}(i,j) - \hat{D}_{n-1}(i,j), \hat{D}_{\max}(i,j) - \hat{D}_{n-1}(i,j)\right] \tag{2}$$

range if $\hat{D}_{n-1}(i,j)$ is close to $\hat{D}_{\min}(i,j)$ or $\hat{D}_{\max}(i,j)$. The constraints are introduced before sampling, instead of sampling and then clipping, to avoid biases caused by over-sampling at the extremes of the depth range.

A *local influence* $J_n^*(i,j)$ is computed at each pixel by preferentially weighting depth perturbations that improve Q. Local influence is aggregated over a support region W to produce *aggregated influence* $J_n(i,j)$. Aggregated influence is added to the last iteration's depth estimate to incrementally improve the estimates, $\hat{D}_n(i,j) = \hat{D}_{n-1}(i,j) + J_n(i,j)$. Details of local and aggregated influence appear in Section III-C.

Finally, depth estimates are smoothed at the end of each iteration, modeling the piecewise continuity constraint [18], and helping the effects of the influence function to propagate.

Search parameters vary by stages. The search schedule defines the maximum depth perturbation magnitude $\delta_{\max}$, aggregation neighborhood W, and number of iterations N for each stage. W is a square region with side length specified as a fraction of the average of the row and column resolutions, μ_{res}, to insulate search parameters from changes in image resolution. W and $\delta_{\max}$ are large in early stages to capture gross scene structure. They both shrink in later stages to capture detail and force convergence of the estimates. This is analogous to the cooling process of simulated annealing or the organization process of self-organizing maps.

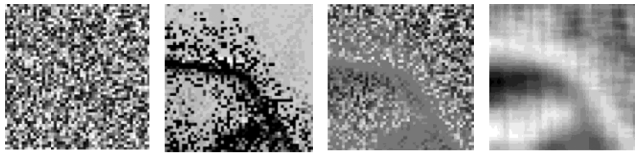
The QUESS approach is heuristic. While good results are achieved and convergence is enforced by the search schedule, the estimates are not guaranteed to be optimal.

C. Match Quality, Local Influence, and Aggregated Influence

Local influence is derived from the stochastic samples of Q. Many alternative Q metrics were explored, including weighted sum of absolute differences, squared error, and normalized cross-correlation (hereafter, XCORR). Although QUESS enables the use of more complex Q, excellent results are achieved even with simple definitions. XCORR is used since it provides superior performance in our simulations owing to the divisive normalization.

Fig. 1 shows example depth perturbations $\Delta_{D_n}(i,j)$ (expressed in grayscale), resulting changes in match quality $\Delta_{q_n}(i,j) = \tilde{q}_n(i,j) - q_{n-1}(i,j)$, local influence $J_n^*(i,j)$, and aggregated influence $J_n(i,j)$.

Local influence selectively weights depth perturbations that improve the depth estimate $\hat{D}_n(i,j)$ as inferred by improvements in Q. Random depth perturbations result in "noisy" $\tilde{q}_n(i,j)$ and $\Delta_{q_n}(i,j)$. Some perturbations increase depth while others decrease depth. Some increase quality while many decrease quality. Local influence should be positive where perturbations that increase depth also increase quality, and negative



(a) (b) (c) (d)

Fig. 1. Influence formulation and aggregation. (a) Depth perturbations $\Delta_{D_n}(i,j)$. (b) Changes in $Q\Delta_{q_n}(i,j)$. (c) Local influence $J_n^*(i,j)$. (d) Aggregated influence $J_n(i,j)$.

where perturbations that decrease depth increase match quality. Where perturbations *decrease* match quality, local influence should be either zero or oriented away from the perturbation.

Results are improved by categorizing pixels as either *contributing* or *noncontributing*. For contributing pixels, $J_n^*(i,j)$ is the depth perturbation realizing the maximum historical sample of Q in the current search stage. For noncontributing pixels, $J_n^*(i,j) = 0$.

A pixel is contributing if it passes two tests.

1. A minimum on the standard deviation of local pixels.
2. A minimum on the range of Q samples at that pixel.

These tests inhibit local influence from pixels where Q is unreliable due to insufficient texture or other features that may cause Q values to be similar (e.g., a dominant gradient along the epipolar direction).

Local influence is defined as (3)–(5), shown at the bottom of the page, where $std_{9\times9}(i,j;I_A)$ is the standard deviation in a local 9×9 square region. The mask $M_n(i,j)$ is defined relative to all historical Q samples, but the local influence of contributing pixels is defined relative to Q samples in the current search stage only (via $d_n^*(i,j)$). This exploits all knowledge of Q to identify reliable samples, but still forces the estimates to converge and capture detail in later search stages. Computing local influence requires maintaining only $d_n^*(i,j)$, $M_n(i,j)$, and two minima/maxima of Q.

Like local influence, there is flexibility in the definition of aggregated influence. It should capture consistent trends in local influence that reflect scene structure, reject spurious local influences caused by artifacts, and tend towards zero when an acceptable solution is reached.

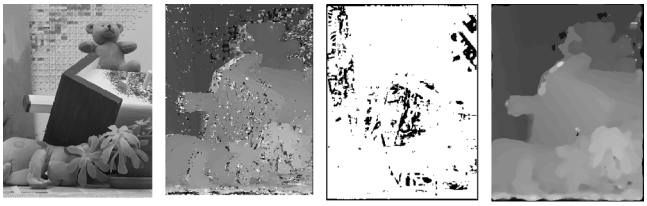
Averaging $J_n^*(i,j)$ over W is efficient and can be effective on some scenes. However, this enforces smoothness where piecewise-smoothness is instead desired. Anisotropic smoothing prevents loss of detail along boundaries [1], [20], with promising accuracy but at a significant cost. Other robust aggregation approaches, including order-statistic filtering (e.g., [3] and [16]) or bilinear filtering can be applied.

We found a selective median filter to be particularly effective. Median filters that combine partial histograms for each column

$$J_n^*(i,j) = \begin{cases} d_n^*(i,j) - \hat{D}_{n-1}(i,j), & \text{if } M(i,j) = \text{true} \\ 0, & \text{if } M(i,j) = \text{false} \end{cases} \tag{3}$$

$$d_n^*(i,j) = \hat{D}_k(i,j) \text{ such that } \tilde{q}_k(i,j) = \max \tilde{q}_{1\ldots n}(i,j) \tag{4}$$

$$M_n(i,j) = [std_{9\times9}(i,j;I_A) > \alpha] \cap \{[\max \tilde{q}(i,j) - \min \tilde{q}(i,j)] > \beta\} \tag{5}$$



(a) (b) (c) (d)

Fig. 2. Selective median filtering. (a) Reference image. (b) Disparity maximizing XCORR. (c) Contributing pixel mask $M_n(i,j)$. (d) aggregated influence $J_n(i,j)$.

are $O(n)$ in pixel number n and $O(1)$ in filter kernel size when applied to integer images [21]. Efficiency is improved by incrementally updating bin indices [12]. We implemented two novel extensions. The first only includes a value in the histograms if it passes a mask. The second applies the filter to floating point images by returning the histogram bin center containing the median—an approximation with bounded error to the true median.

Fig. 2 illustrates the effects of the filter on a representative disparity image. Depicted are the reference image, the disparity at which XCORR is maximized, an example contribution mask, and the results of selective median filtering.

IV. Extensions to Calibrated Video

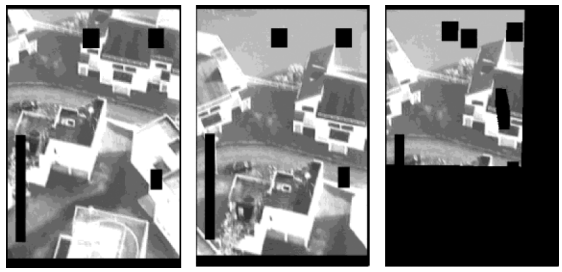
Modifications of the search process combined with pre- and postprocessing provide additional efficiencies when QUESS is applied to calibrated video input.

A. Preprocessing

The processing necessary to operate on calibrated aerial video comprises frame pairing, rectification, and input masking.

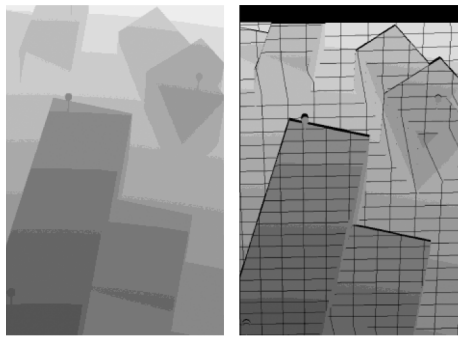
When applied to calibrated video, QUESS generates depth estimates for each frame in the video stream. The most recent frame in the stream is defined as the reference frame I_A. Following [25], stereo pairs are formed by selecting a nonadjacent frame I_B (see Fig. 3) to maintain a target ratio τ_0 between the stereo baseline T and the minimum depth to the scene. To simplify computation, T is defined simply as the Euclidean distance between the camera origins. In addition to providing a stereo baseline sufficient to generate accurate depths (which pairing adjacent frames would not do), this provides robustness to changes in platform speed and direction, making it possible to tune other parameters to a more consistent geometry.

QUESS gains advantages from rectifying, but rectification is not strictly necessary. Image I_B is reprojected to a plane that is parallel to (but not necessarily coplanar with) the image plane of I_A, thereby creating I_{BR}. This can be described as a partial planar rectification. After rectification, scene elements at infinite depth exhibit zero disparity, and unit vectors in the epipolar direction at each pixel in I_A can be computed and stored. Epipolar lines are not collinear or parallel as in a standard stereo geometry. However, the rectification allows quick conversion between estimated depth and equivalent disparity for computing match quality. This process also enforces correspondences to lie on epipolar lines.



(a) (b) (c)

Fig. 3. (a) Reference image I_A masked for artifacts. (b) Paired image I_B masked for artifacts. (c) Rectified paired image I_{BR} masked for artifacts, rectification, and scene assumptions.



(a) (b)

Fig. 4. (a) Depths (pseudocolored) in frame I_A. (b) Depths re-projected to frame I_{A+10} by Z-buffering.

Following frame pairing and rectification, a mask is computed to identify pixels satisfying various constraints on correspondences.

1. That I_{BR} has the same domain as I_B.
2. That corresponding pixels lie within the images.
3. That pixels are not coincident with known artifacts.
4. That depth estimates respect known scene boundaries.

Pixels failing these constraints are black in Fig. 3(c).

B. Modifications to Stochastic Cooperative Search

QUESS stochastic search is modified in three ways. First, estimates and intermediate values are initialized using results from the previous frame pair when available. This lets estimates converge over multiple frames—existing estimates are refined instead of generating entirely new estimates. As shown in Fig. 4, depth estimates can be seeded aggressively since camera position and orientation differ little between adjacent frames. Estimates from the last frame are re-projected to the new reference frame and adjusted for changes in camera origin, using Z-buffering [8] to address occlusion. A similar application of Z-buffering is used to initialize key quantities such as $d_n^*(i,j)$, $M_n(i,j)$, and statistics of Q, letting the search leverage results from the previous frame pair. Any small gaps can be filled by nearest-neighbor interpolation.

Second, the search schedule is modified to exploit the redundancy between frame pairs. The search schedules still require successive stages to decrease the perturbation magnitudes and neighborhood sizes, thus capturing both gross scene structure

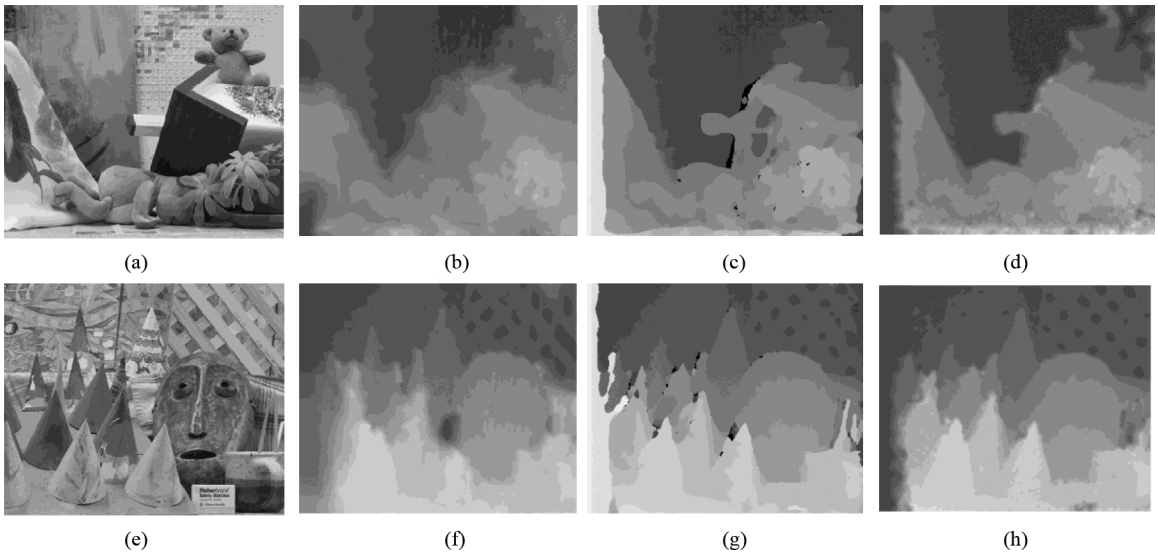


Fig. 5. Results on Middlebury Teddy (a)–(d) and Cones (e)–(h) data. (a), (e) Reference image. (b), (f) Micro-Canonical Annealing (MCA) [2] disparities. (c), (g) Zitnick–Kanade (ZK) [27] disparities. (d), (h) QUESS disparities.

TABLE II
QUESS SEARCH SCHEDULE PARAMETERS

Parameter	Search Stage			
	1	2	3	4
N, Middlebury data	30	30	30	30
N, first video frame	20	20	20	20
N, later video frames	4	6	8	8
W side length	$\frac{\mu_{res}}{30}$	$\frac{\mu_{res}}{40}$	$\frac{\mu_{res}}{60}$	$\frac{\mu_{res}}{120}$
$\delta_{\max}$	0.50	0.25	0.15	0.03

and detail. A unique schedule is used for the first frame that performs more iterations and emphasizes larger perturbations. All subsequent frames use schedules with significantly fewer iterations. These emphasize smaller neighborhoods and smaller depth perturbation in order to refine the existing solution and capture detail, although they must also capture larger structures in newly-visible regions. Combined with aggressive seeding, the modified search schedule lets estimates converge over many frame pairs with very few iterations (and Q evaluations) per pair.

Third, conservative assumptions constrain the depth estimates at each pixel. Application-specific assumptions can considerably improve speed and accuracy. For example, aerial modeling benefits more from bounds on elevation than bounds on disparity, and those bounds are easier to estimate reliably (e.g., from existing low-resolution elevation data).

C. Postprocessing

After computing depth estimates with respect to I_A, intrinsic and extrinsic camera parameters are used to compute equivalent 3-D positions in an absolute reference frame. The final output is a 3-D point cloud for each frame. These clouds can be fused and converted to surface models for further analysis using tools and techniques such as [6] and [19].

V. PERFORMANCE ANALYSIS

A. Middlebury Stereo Pairs

Evaluation on the Middlebury data lets us compare QUESS against many leading approaches, although the benefits of QUESS are stronger when processing aerial video data.

A variety of parameter combinations were explored and the best results are presented. QUESS used XCORR over a 5×5 window for Q, and influence thresholds of $\alpha = 0$ and $\beta = 0.10$. Its search parameters are given in Table II. Relatively many Q samples are required for a single stereo pair but far fewer samples can used for video data. MCA parameters were set following [2] (minimum temperature 30, maximum temperature 300, 500 iterations per scale, and 85% of iterations for cooling). The parameter λ, which weights smoothness against match quality, was set empirically to $\lambda = 80$. Following [27], ZK used 15 iterations, occlusion threshold 0.005, and a $5 \times 5 \times 3$ support region. We empirically set inhibition exponent $\alpha = 1.25$ and used 1×1 absolute differences (AD) for Q. Alternative quality metrics such as 5×5 sum of absolute differences, 1×1 squared differences, and 5×5 sum of squared differences did not improve results.

Reference images and computed disparity are shown in Fig. 5 for the Teddy and Cones datasets, and performance metric values are given for all Middlebury datasets in Table III.

QUESS is not competitive with leading approaches on the Tsukuba or Venus scenes. On the Teddy and Cones scenes it is within the range of results posted for other approaches, although it is not a top performer. This is not unexpected since many algorithms leading the Middlebury evaluation emphasize single stereo pairs of indoor scenes taken at short range, or scenes that have large planar regions, large areas of low contrast, or relatively simple geometries. These scenes allow assumptions and techniques that may be less attractive for outdoor aerial modeling or other data. QUESS performs best on the two scenes that are most representative of outdoor scenes in their complexity, disparity ranges, nonplanar geometry, and higher texture. These

TABLE III
MIDDLEBURY EVALUATION RESULTS FOR MICRO-CANONICAL ANNEALING (MCA), ZITNICK-KANADE (ZK), AND QUESS (Q). NON, ALL, AND DISC STAND FOR NONOCCLUDED, ALL, AND NEAR DISCONTINUITIES

	Tsukuba			Venus			Teddy			Cones			Avg.
	Non	All	Disc	Non	All	Disc	Non	All	Disc	Non	All	Disc	
MCA	20.0	**21.8**	48.6	22.8	24.1	53.3	34.0	39.6	51.3	26.8	32.9	49.1	35.4
ZK	27.0	28.5	**30.4**	27.4	28.1	**41.7**	20.4	27.9	**33.0**	13.8	22.8	**24.6**	27.1
Q	**23.4**	25.1	44.7	**16.2**	**17.7**	44.0	**16.0**	**24.6**	36.6	**11.0**	**19.3**	27.3	**25.5**

results show that QUESS is viable even on types of data it does not emphasize

Our ZK results did not repeat those achieved in [27], where the authors obtained nonoccluded error rates of 1.5%–3.0% on the Tsukuba scene. Results for other scenes were not given. Our observed performance was significantly tdifferent on Tsukuba (27.0%–28.5%) and was slightly worse than QUESS on all metrics except those exclusively measuring results near disparity discontinuities. ZK provides a novel method for explicitly identifying occlusions, so superior performance near discontinuities is expected. Because we used our own implementation of ZK, the differences between our results and those of [27] imply that an important subtlety of the approach may have been missed in either the published description or in our implementation.

MCA performance was generally poor, both qualitatively and quantitatively. Appealing qualitative results were shown in [2] on other datasets, but we are not aware of prior MCA results posted for Middlebury data.

Approaches that stochastically sample Q are rare in the literature and none appear among the over 60 approaches with posted Middlebury results at the time of writing. While many approaches use stochastic models of the disparity field, they sample Q exhaustively and apply deterministic optimization algorithms. By combining stochastic and cooperative techniques, QUESS outperforms approaches from each category. It is also the only approach we are aware of using nonexhaustive stochastic sampling and optimization that is competitive with the top 60 performers on any of the Middlebury datasets. Nonexhaustive sampling of Q provides complexity, runtime, and memory benefits as discussed below.

B. Aerial Video

Performance was evaluated on a calibrated monocular aerial video dataset provided by the Air Force Research Laboratory. The dataset contains 32 videos of a suburban scene captured at 60 frames per second (interlaced) and 720×480 resolution. The scene spans 220×225 m in the horizontal and 17 m in the vertical. Sparse ground truth positions are known for 301 locations, including building corners, fiducial markers, and ground locations. The platform traveled at 35 mph at elevations around 110 m, with camera declinations of -45 to -50 degrees, yielding true depths in the range of 150 to 220 m. Platform position and orientation is known for each frame. Field of view and nontrivial offsets in position and orientation between the platform and camera were estimated by minimizing the re-projection error of ground truth positions. Analysis was performed on a representative 200-frame de-interlaced sequence.

Calibration inaccuracies in intrinsic and extrinsic parameters shape the definition of the accuracy metric. Sparse ground truth is projected to a 2-D pixel location. Depth estimates for all pixels within a radius r are considered and absolute error (AE) is defined as the minimum Euclidean distance between the ground truth position and the estimated 3-D positions. Mean absolute error (MAE) averages AE over all visible ground truth points and all frames. This defines a family of MAE metrics E_r with error values decreasing monotonically with r.

Results are presented for E_5, which was selected by inspection based on residual re-projection error. Between 1800 and 2200 ground truth comparisons contributed to each MAE value. Results are presented for downsampled 360×240 video due to the memory limitations of the ZK approach, discussed in Section V-E.

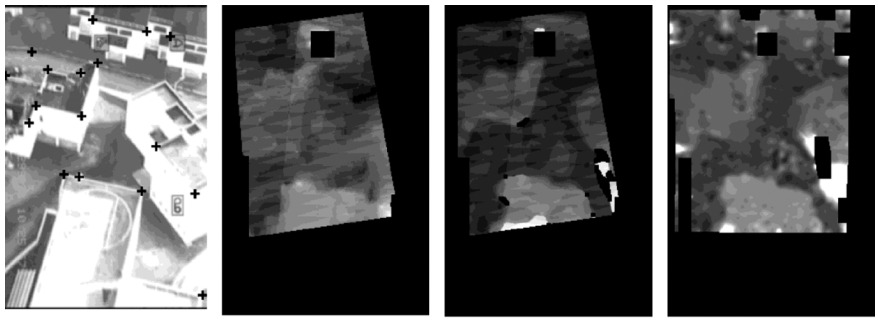
A variety of parameter combinations were explored and the best results are presented. QUESS used XCORR over a 5×5 window for the match quality metric, and influence thresholds of $\alpha = 1.5$ and $\beta = 0.15$. Its search parameters are given in Table II, which require significantly fewer Q evaluations per frame than when processing a single stereo pair. MCA parameters were identical to the evaluation on Middlebury data. ZK used ten iterations, occlusion threshold 0.02, a $5 \times 5 \times 3$ support region, $\alpha = 1.5$ and 5×5 sum of absolute differences (SAD) for the match quality metric.

The target stereo baseline was varied over $0.01 \leq \tau_0 \leq 0.20$ for MCA and QUESS, and up to 0.25 for ZK to capture all important trends. Intentionally loose elevation assumptions simulated imprecise *a priori* scene knowledge (45 m vertical range versus the actual 17 m range) and defined the disparity ranges for each approach. MCA and ZK require a standard stereo geometry, so a planar rectification following [10] was applied. Disparity estimates were converted to depth estimates and transformed to the reference frame for evaluation.

Fig. 6 shows a reference image and example reconstructions for MCA, ZK, and QUESS. Results are shown for $\tau_0 = 0.07$, at which QUESS achieves its best accuracy. Fig. 7 plots reconstruction error against target stereo baseline ratio τ_0 for the three approaches.

As seen in Fig. 7, QUESS outperforms both MCA and ZK at the sparse evaluation positions. QUESS achieves $E_5 = 1.15$ m at $\tau_0 = 0.07$. This equates to 0.62% estimate error relative to absolute depth, which is 0.29 pixels average disparity magnitude error at that baseline. ZK performance is somewhat competitive, and achieves $E_5 = 1.88$ m at $\tau_0 = 0.18$, resulting in 1.01% depth estimate error and 1.50 pixels average disparity magnitude error. MCA performance is not competitive, achieving a minimum $E_5 = 3.39$ at $\tau_0 = 0.08$. These relative results are consistent with evaluations on the Middlebury data.

A few trends are evident. The accuracy of all approaches degrades for small τ_0, where the reduced disparity range makes



(a) (b) (c) (d)

Fig. 6. Example reconstructions from aerial video. (a) Reference images with sparse evaluation positions marked. (b) MCA estimated elevations. (c) ZK estimated elevations. (d) QUESS estimated elevations.

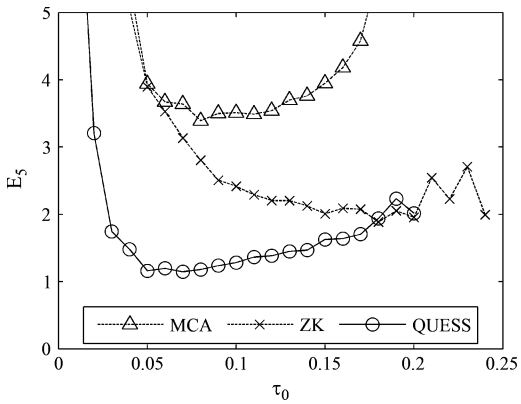


Fig. 7. Reconstruction error versus stereo baseline.

disparity estimates sharply quantized, and depth estimates are simultaneously more sensitive to disparity error. Metrics defined at sub-pixel disparities ([24]) would help, but increased sensitivity will remain. QUESS and MCA accuracy degrade at higher τ_0 where viewpoint changes make correspondence matching more difficult. ZK accuracy gradually improves as τ_0 grows and the effects of disparity quantization are reduced. ZK accuracy becomes unstable as τ_0 grows further, likely because few frames are successfully paired and few depth estimates are generated per frame. QUESS outperforms ZK and MCA, but it can produce inaccurate results in large regions of low texture or contrast (as do most other approaches).

These results demonstrate that by combining stochastic and cooperative techniques, QUESS outperforms both a stochastic approach and an exhaustive cooperative approach on a realistic and complex dataset. Accurate range estimates are generated from high-range calibrated aerial video. QUESS has a variety of additional advantages which are discussed next.

C. Number and Scalability of Match Quality Evaluations

The primary advantages of QUESS are the generation of depth estimates using fewer evaluations of Q, and its attractive scaling properties with respect to video resolution, stereo baseline ratio, and scene bound assumptions.

Analysis of Q evaluations is focused on comparing to ZK because ZK can be used as a representative for other approaches. For example, the current MCA implementation is not competitive with respect to Q evaluations because Q is recomputed on each iteration which results in 500 Q samples per pixel *per scale*. We could reimplement MCA to exhaustively sample Q at each scale and then use a lookup table, but then MCA would still be no better than ZK in number of Q samples. Any approach that exhaustively samples Q will encounter the same scalability issues as ZK.

QUESS requires $K_Q = 2RC\alpha_Q(\vec{p},\tau_0)N_{\text{tot}}$ evaluations of Q for R by C images, where $\alpha_Q(\vec{p},\tau_0)$ is the fraction of overlapping pixels in the stereo pair (a function of τ_0 and camera path $\vec{p}$), and N_{tot} is the total number of iterations in the schedule. The value $\alpha_Q \in [0,1]$, so $K_Q \leq 2RCN_{\text{tot}}$. QUESS generates $RC\alpha_Q(\vec{p},\tau_0)$ depth estimates per frame pair.

Exhaustive approaches are more difficult to characterize. ZK requires

$$K_{ZK} = R^*(\vec{p},\tau_0)C^*(\vec{p},\tau_0)\alpha_{ZK}(\vec{p},\tau_0)D(\vec{p},\tau_0;R,C,B)$$

evaluations, where R^* and C^* are the number of rows and columns after projective rectification, α_{ZK} is analogous to α_Q, and D is the *range* of potential disparities for resolution RxC and scene bounds B. D is a complex function of camera path and stereo baseline that grows with increasing baseline. An upper bound on D is not easily determined. ZK generates $RC\alpha_{ZK}(\vec{p},\tau_0)$ depth estimates per frame pair.

Fig. 8 plots the average Q evaluations per frame for the aerial video test data, as a function of τ_0. QUESS requires fewer evaluations per frame only in some ranges. QUESS shows a steady decline in K_Q as α_Q shrinks with increasing τ_0 and other factors remain constant. For ZK, the number of evaluations grows with τ_0 for low τ_0 until a decrease in α_{ZK} dominates and K_{ZK} follows. The values of R^*, C^*, D, and α_{ZK} are all complicated functions of camera path and baseline. QUESS achieves its best accuracy at 2.6×10^6 Q evaluations per frame and ZK achieves its best accuracy at 2.2×10^6 per frame. On the surface, ZK may appear superior in number of Q evaluations (although its accuracy is worse), but this is not the complete story.

Fig. 9 plots the number of depth estimates generated per frame. Increasing the stereo baseline decreases the number of estimates per frame as overlap decreases. Qualitatively different behavior is seen in the number of evaluations of Q *per depth estimate* in Fig. 10. QUESS uses a constant number of match quality evaluations per depth estimate, independent

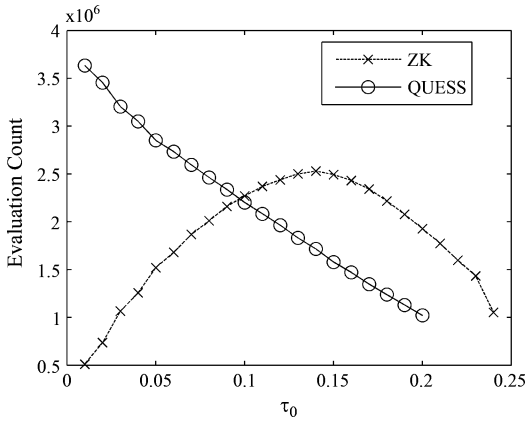


Fig. 8. Average Q evaluations per frame.

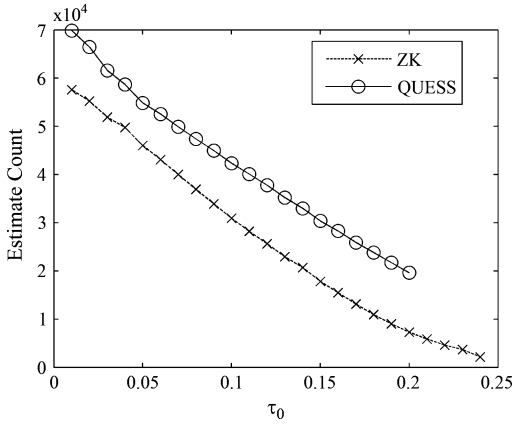


Fig. 9. Average depth estimates per frame.

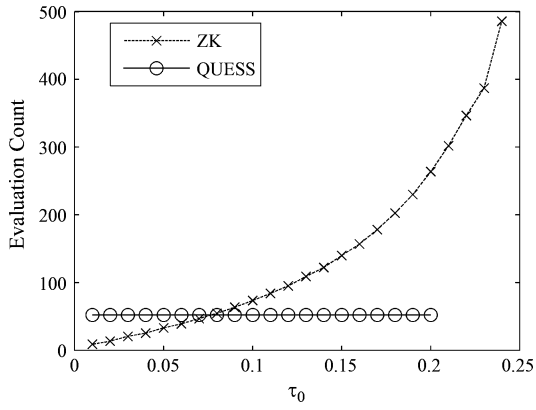


Fig. 10. Average Q evaluations per depth estimate.

of stereo baseline. The number of ZK evaluations per estimate is dominated by the growth of D with increasing τ_0. QUESS achieves its best performance at 52 evaluations per estimate, but ZK requires an average of 203 per estimate. QUESS generates more accurate results using 75% fewer Q evaluations per estimate.

The specifics of the analysis will differ between exhaustive approaches, but the themes generalize. For roughly linear camera paths, all factors α_* will decrease with increasing τ_0 (here, $\alpha_Q > \alpha_{ZK}$ by coincidence only). However, α_*, R^*, and C^* are determined by frame pairing and rectification. Any exhaustive algorithm that needs a standard stereo geometry and uses the same projective rectification will share these values. The value D is also shared. For any exhaustive approach, α_* will decrease with τ_0 and D will increase, resulting in more evaluations of Q per estimate. By contrast, QUESS can freely optimize τ_0 without changing the number of Q evaluations.

QUESS also has more attractive scaling properties with respect to resolution and other factors. For QUESS, Q evaluations scale directly with $O(n)$ ($n =$ number of pixels). For ZK, doubling R and C also doubles D, so K_{ZK} scales with $O(n^{3/2})$. This applies to any method that exhaustively computes Q at integer disparity magnitudes in a search range based on camera and scene geometry—doubling resolution causes an unavoidable $O(n^{3/2})$ scaling in evaluations of Q.

ZK inhibition computations scale with $O(n^2)$ in the number of pixels, as opposed to the $O(n^{3/2})$ scaling described in [27]. For each row, column, and disparity, inhibition is summed over a second disparity index whose range is also linear in resolution. For simple Q metrics, computing ZK inhibition may dominate computing Q, but that is not necessarily true for complex Q metrics or for other approaches.

Scene geometry and camera path have complex and significant effects on D for any exhaustive approach, further complicating their use under uncontrolled scene and camera geometries. By contrast, the number of Q evaluations in QUESS is independent of scene geometry and camera path once the number of iterations is chosen.

D. Runtime

Runtimes are given in Table IV for each dataset at three different resolutions. Runtime was measured with a single 2.5-GHz dual-core CPU with 3.5-GB RAM. While QUESS was not the fastest of the three approaches, a direct comparison does not reflect all the relevant issues. Algorithms were implemented in Matlab and were vectorized, but with no effort to optimize the implementations. As a result, QUESS did not capitalize on opportunities to reduce the number of match quality evaluations by exploiting increases in τ_0. MCA and ZK, however, do benefit because lower frame overlap shrinks the size of the data cube created for rectified stereo pairs.

QUESS runtimes were significantly lower on video data than on the single stereo pairs because the initialization techniques described in Section IV-B allow its search schedules to be shortened. Runtime scales with number of pixels, increasing by a factor of about four for every doubling in resolution. QUESS runtimes are independent of τ_0, and for an optimized implementation would actually decrease with increasing τ_0. This insulation of runtimes also applies to stereo geometry and relative scene orientation. Neither characteristic holds for approaches that exhaustively sample Q.

ZK runtimes follow the number of Q evaluations shown in Fig. 8. Runtime ranges are given because ZK runtime varies significantly with τ_0. As expected, each doubling of resolution creates an 8-fold increase *per frame* in Q evaluations, an approximate 8-fold increase in total runtime, and doubling of Q evaluations and runtime *per depth estimate*. ZK thus scales with $O(n^{3/2})$ in the number of pixels.

TABLE IV
RUNTIME (SECONDS/FRAME) FOR MICRO-CANONICAL ANNEALING (MCA), ZITNICK-KANADE (ZK), AND QUESS (Q). RESULTS ARE GIVEN FOR FULL-RESOLUTION (FR), 1/2 RESOLUTION, 1/4 RESOLUTION, AND 1/8 RESOLUTION

	Middlebury (average)			Aerial Video		
Resolution	FR	1/2	1/4	1/2	1/4	1/8
MCA	95.7	21.4	5.7	10.9-27.5	3.8-8.1	1.6-2.5
ZK	95.9	12.3	1.3	15.6-23.0	1.8-2.8	0.3-0.4
Q	359.1	72.5	19.0	41.1	9.5	2.9

MCA runtimes also follow the number of Q evaluations, and also vary significantly with τ_0. Q evaluations scale linearly in pixels in the current implementation because Q is recomputed each iteration. If we instead quantize the disparity space and sample it exhaustively, MCA will scale with $O(n^{3/2})$ like ZK but with a much lower hidden constant that it currently has.

The use of simple Q functions actually minimizes the runtime advantages of QUESS over other approaches. As more complex metrics are used, evaluation of Q becomes a larger percentage of runtime and the advantages of quality-efficient stochastic sampling become more pronounced.

E. Memory Usage

QUESS has memory advantages over approaches that exhaustively sample Q and retain all samples in memory, and it is thus more attractive than exhaustive approaches on memory-constrained devices or platforms. QUESS requires storing approximately 20 floating-point values for each pixel, independent of τ_0. All aspects of memory usage scale with $O(n)$ and are independent of stereo baseline, frame overlap, camera motion, and scene structure.

ZK memory usage follows directly from the number of Q evaluations analyzed in Section V-C, since all Q samples are stored in a 3-D data cube. As a result, ZK memory also scales with $O(n^{3/2})$. ZK requires two data cubes of this size. At its best accuracy, ZK requires the simultaneous storage of over 400 floating point values per depth estimate.

As currently implemented, MCA memory requirements are similar to QUESS because Q is recomputed in each iteration. If we instead compute Q values once at all disparities and store them for later retrieval, MCA memory usage becomes nearly identical to ZK. Both alternatives have significant disadvantages for MCA.

Similar scaling properties are shared by other exhaustive approaches. The size of any exhaustive cube of Q samples is affected by stereo baseline, frame overlap, camera motion, scene orientation, and scene structure. Few of these factors are easily controlled so exhaustive approaches can cause problems on limited memory devices. Our aerial video comparisons were performed at 360×240 resolution because even on a modern machine with significant memory, ZK generated Matlab out-of-memory errors on 720×480 imagery.

VI. CONCLUSION

This paper presents quality-efficient stochastic sampling (QUESS), a new stochastic and cooperative sampling approach for generating dense stereo correspondence estimates using fewer local match quality metric evaluations than exhaustive approaches. It is based on a set of general techniques that are easily applied to a variety of applications. Its strengths are maximized when operating on calibrated monocular video, but in more common applications such as robotic navigation the approach suffers no loss. It exploits the continuity of matching likelihood constraint to skip portions of the disparity search space. Estimates are initialized from the previous frame pair's results to allow convergence across multiple frame pairs. A relatively simple formulation of *local influence* selectively re-weights random perturbations injected into the solution, and *aggregated influence* extracts consistent trends from the stochastic sampling of match quality. QUESS is both stochastic and cooperative, with advantages from both. It was shown to outperform both Barnard's stochastic approach [2] and Zitnick and Kanade's cooperative approach [27] on a complex and representative dataset, while requiring fewer match quality evaluations.

QUESS has a number of advantages over exhaustive approaches. It requires fewer match quality metric evaluations per depth estimate, with corresponding gains in efficiency. It reduces memory requirements and provides better scaling in both runtime and memory. Runtime and memory usage are insulated from a variety of factors that cannot be easily controlled, including stereo baseline, camera path, and scene orientation and structure. Advantages are demonstrated using simple quality metrics, but become more pronounced as metric complexity increases. It facilitates the use of complex and robust metrics, and metrics defined on noninteger disparities.

Potential future work includes further tuning of search schedules, exploring influence formulations based on gradient ascent search, and exploring efficient anisotropic filtering [1] for influence aggregation and depth estimate smoothing. Further work with more complex quality metrics is also of interest, as is developing optimized and parallelized implementations.

REFERENCES

[1] S. T. Acton, A. C. Bovik, and M. M. Crawford, "Anisotropic diffusion pyramids for image segmentation," presented at the IEEE Int. Conf. Image Processing., 1994.
[2] S. Barnard, "Stochastic stereo matching over scale," *Int. J. Comput. Vis.*, vol. 3, no. 1, pp. 17–32, 1989.
[3] A. C. Bovik, T. S. Huang, and D. C. Munson, "A generalization of median filtering using linear combinations of order statistics," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-31, no. 6, pp. 1342–1350, Dec. 1983.
[4] 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.
[5] M. Creutz, "Microcanonical Monte Carlo simulation," *Phys. Rev. Lett.*, vol. 50, no. 9, pp. 1411–1414, 1983.
[6] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," presented at the ACM Int. Conf. Computer Graphics and Interactive Techniques, 1996.
[7] U. R. Dhond and J. K. Aggarwal, "Structure from stereo—A review," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, pp. 1489–1510, 1989.

[8] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practice*, 2nd ed. Boston, MA: Addison-Wesley, 1996.
[9] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721–741, Jun. 1984.
[10] R. Hartley, "Theory and practice of projective rectification," *Int. J. Comput. Vis.*, vol. 35, no. 2, pp. 115–127, 1998.
[11] H. Hirschmüller and D. Scharstein, "Evaluation of cost functions for stereo matching," presented at the IEEE CVPR, 2007.
[12] T. Huang, G. Yang, and G. Tang, "A fast two-dimensional median filtering algorithm," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 13–18, 1979.
[13] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka, "A stereo machine for video-rate dense depth mapping and its new applications," presented at the IEEE CVPR, 1996.
[14] W. N. Klarquist and A. C. Bovik, "FOVEA: A foveated, multi-fixation, vergent active stereo system for dynamic three-dimensional scene recovery," *IEEE Trans. Robot. Autom.*, vol. 14, no. 5, pp. 755–770, May 1998.
[15] Y. Kim and J. K. Aggarwal, "Positioning 3-D objects using stereo images," *IEEE J. Robot. Autom.*, vol. 3, no. 4, pp. 361–373, Apr. 1987.
[16] H. G. Longbotham and A. C. Bovik, "Theory of order statistic filters and their relationship to linear FIR filters," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 2, pp. 275–287, Feb. 1989.
[17] D. C. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, pp. 283–287, 1976.
[18] D. C. Marr and T. Poggio, "A computational theory of human stereo vision," in *Proc. Roy. Soc. Lond. B*, 1979, vol. 204, pp. 301–328.
[19] P. Merrell *et al.*, "Real-time visibility-based fusion of depth maps," presented at the IEEE ICCV, Oct. 2007.
[20] P. Perona and J. Malik, "Scale-space and edge detection using anisotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, Jul. 1990.
[21] S. Perreault and P. Hebert, "Median filtering in constant time," *IEEE Trans. Image Process.*, vol. 18, no. 9, pp. 2389–2394, Sep. 2007.
[22] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, 2002.
[23] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," presented at the IEEE CVPR, 2006.
[24] R. Szeliski and D. Scharstein, "Sampling the disparity space image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 3, pp. 419–425, Mar. 2004.
[25] R. Vidal and J. Oliensis, "Structure from planar motions with small baselines," in *Proc. ECCV*, 2002, pp. 383–398.
[26] R. Yang and M. Pollefeys, "Multi-resolution real-time stereo on commodity graphics hardware," presented at the IEEE CVPR, 2003.
[27] 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.

Thayne Richard Coffman (M'08–SM'09) received the B.S. and M.Eng. degrees in computer science and electrical engineering from the Massachusetts Institute of Technology, Cambridge, in 1996. He is currently pursuing the Ph.D. degree in electrical and computer engineering at The University of Texas at Austin.

He is also employed full-time as a Technical Fellow and member of the executive team at 21st Century Technologies, Austin, TX, where he identifies, directs, and performs research that addresses critical needs in the military and intelligence communities. His research interests include computational stereo, image processing, image and video understanding, autonomous systems, statistical pattern classification, graph-theoretic pattern classification, social network analysis, and network intrusion detection. He has worked in the past for Trilogy Software, Hughes Network Systems, and the National Institute of Standards and Technology. He has authored or coauthored 16 technical articles and book chapters, and has five U.S. and international patent applications pending.

Alan Conrad Bovik (S'80–M'81–SM'89–F'96) received the B.S., M.S., and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana-Champaign, Urbana, in 1980, 1982, and 1984, respectively.

He is currently the Curry/Cullen Trust Endowed Professor at The University of Texas at Austin, where he is the Director of the Laboratory for Image and Video Engineering (LIVE) in the Center for Perceptual Systems. His research interests include image and video processing, computational vision, digital microscopy, and modeling of biological visual perception. He has published over 450 technical articles in these areas and holds two U.S. patents. He is also the author of *The Handbook of Image and Video Processing* (Elsevier, 2005, 2nd ed.) and *Modern Image Quality Assessment* (Morgan & Claypool, 2006).

Dr. Bovik has received a number of major awards from the IEEE Signal Processing Society, including: the Education Award (2007); the Technical Achievement Award (2005); the Distinguished Lecturer Award (2000); and the Meritorious Service Award (1998). He is also a recipient of the Distinguished Alumni Award from the University of Illinois at Urbana-Champaign (2008), the IEEE Third Millennium Medal (2000), and two journal paper awards from the International Pattern Recognition Society (1988 and 1993). He is a Fellow of the Optical Society of America and the Society of Photo-Optical and Instrumentation Engineers. He has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996–1998; Editor-in-Chief, IEEE TRANSACTIONS ON IMAGE PROCESSING, 1996–2002; Editorial Board, PROCEEDINGS OF THE IEEE, 1998–2004; Series Editor for Image, Video, and Multimedia Processing, Morgan and Claypool Publishing Company, 2003–present; and Founding General Chairman, First IEEE International Conference on Image Processing, Austin, TX, November 1994. He is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial, and academic institutions.