Boston University

OpenBU **http://open.bu.edu**

BU Open Access Articles BU Open Access Articles

2019-01-24

Computational periscopy with an ordinary digital camera

Charles Saunders, John Murray-Bruce, Vivek K Goyal. 2019. "Computational periscopy with an ordinary digital camera." NATURE, Volume 565, Issue 7740, pp. 472 - 475 (6). https://doi.org/10.1038/s41586-018-0868
https://hdl.handle.net/2144/39239
"Downloaded from OpenBU. Boston University's institutional repository."

Computational Periscopy with an Ordinary Digital Camera

Charles Saunders[1,2], John Murray-Bruce[1,2], and Vivek K Goyal[1*]

[1]Department of Electrical and Computer Engineering, Boston University, 8 St. Mary's Street, Boston, MA 02215 USA.
[2]Contributed equally.
[*]To whom correspondence should be addressed; e-mail: goyal@bu.edu.

Computing the proportions of light arriving from different directions allows a diffusely reflecting surface to play the role of a mirror in a periscope. Such computational periscopy has previously depended on light travel distances being proportional to times of flight and thus has mostly been achieved with expensive, specialized ultrafast optical systems[1–12]. We introduce a 2D computational periscopy technique that requires only a single photograph captured with an ordinary digital camera. Our technique recovers the position of an opaque object and the scene behind (but not completely obscured by) the object, where both the object and scene are outside the line of sight of the camera, without requiring controlled or time-varying illumination. Non-line-of-sight imaging with only inexpensive, ubiquitous equipment may have considerable value in monitoring of hazardous environments, navigation, and detecting hidden adversaries.

The ability to accurately image scenes or detect objects hidden from direct view has many potential applications. Active optical methods for this non-line-of-sight (NLOS) imaging problem have been developed recently, most depending on *transient imaging* (TI). In a typical TI configuration, an imaging device consisting of a light source and a light detector lacks direct view of the scene but does have direct view of a diffusely reflecting surface that itself has direct view of the scene. Illumination of a small patch on the diffuse surface with a short light pulse creates transient illumination of the NLOS scene, which is observed indirectly through light that reaches the detector after reflection from the diffuse surface. Kirmani *et al.*[1] first demonstrated TI-based NLOS scene geometry recovery through multilateration, and the modeling was extended to include variable reflectivity of scene elements and non-impulsive illumination[13]. Subsequent early works used TI to infer shapes of objects with non-specular surfaces hidden from direct view of the observer[2,3]. The pioneering demonstrations used femtosecond laser illumination and picosecond-resolution streak cameras[1–3]. Cost of TI acquisition can be reduced dramatically with homodyne time-of-flight sensors[14–16], and the increasing availability of single-photon avalanche diode (SPAD) detectors and detector arrays with time-correlated single photon counting (TCSPC) modules has led to their use in TI-based NLOS imaging[4–12]. SPADs with TCSPC are common in many LIDAR applications and were recently used for very long-range three-dimensional imaging[17] and to capture photometric and geometric information from as few as one detected photon per pixel[18–21]. In addition to lowering the cost of NLOS imaging systems, SPAD-based systems have facilitated the extension of previous round-trip distances of around 1 m to a few meters for NLOS hidden-object estimation[5] and over 50 m for human localization at long range by coupling a telescope to a single-element SPAD[9]. Furthermore, room geometry reconstruction by probing a single visible wall using a picosecond laser and SPAD with TCSPC has been demonstrated[22]. Other demonstrated applications of TI include NLOS estimation of object motion and size[23] and single-viewpoint estimation of angular reflectance properties[24].

To address the high cost and impracticality outside of laboratory scenarios of existing methods, we developed an approach to computational periscopy using only an ordinary digital camera. The imaging method is passive, with radiosity of the NLOS scene due to sources hidden from view and uncontrolled. The NLOS resolution is based on computational inversion of the influence of the scene of interest on the penumbra of an occluding object of known size and shape that is in an *a priori* unknown position. Previous exploitations of penumbrae required precise knowledge of occluder positions and used laser illumination and SPAD-based detection[25,26], required occluder motion[27], or had the more limited objective of producing a one-dimensional projection of the moving portion of a scene[28]. A very recent work uses calibration measurements of a complex occluder in a light field reconstruction[29]. NLOS tracking of a moving object using laser illumination—without image formation—was also demonstrated by Klein et al.[30]. We do not require calibration, controlled illumination, time-resolved light detection, or scene motion, and we obtain a full-color 2D image.

We demonstrated computational periscopy using an experimental setup consisting of a 4 megapixel digital camera, a 20-inch liquid crystal display (LCD) color monitor with 4:3 aspect ratio, and a black rectangular occluding object of size 7.7 cm-by-7.5 cm supported by a black-colored flat 7 mm-wide stand (Fig. 1). This particular shape of occluder was chosen for computational convenience; any known occluder shape and size would be incorporated similarly. Additional experiments using a 3D, non-black occluder are included in the Supplementary Materials[31]. Light from the LCD monitor, due to the unknown displayed scene and the monitor's background light, illuminates a visible white Lambertian surface placed fronto-parallel with the monitor at a distance of 1.03 m. A monitor was used to allow convenient testing of multiple scenes; additional results using both 2D and 3D diffuse reflecting scenes are presented in the Supplementary Materials[31]. The camera measurement, which includes shadows and penumbrae cast by the occluder, is a raw 14-bit, 2016-by-2016 pixel image with the color channels interleaved according to a Bayer filter RGBG pattern. After averaging of the two green channels and averaging over 16-by-16 blocks, three 126-by-126 images, one per color channel, are extracted and passed to a computer algorithm for occluder position and scene image recovery.

With the occluder positioned at $\boldsymbol{p}_o$ between the monitor and visible wall and the monitor at distance D, the irradiance of a wall patch at $\boldsymbol{p}_w$ is given by

$$I(\boldsymbol{p}_w) = \int_{\boldsymbol{x}\in\mathcal{S}} \frac{\cos(\measuredangle(\boldsymbol{p}_w-\boldsymbol{x},\boldsymbol{n}_x))\cos(\measuredangle(\boldsymbol{x}-\boldsymbol{p}_w,\boldsymbol{n}_w))}{\|\boldsymbol{p}_w-\boldsymbol{x}\|_2^2} V(\boldsymbol{x},\boldsymbol{p}_w;\boldsymbol{p}_o)\mu(\boldsymbol{x},\boldsymbol{p}_w)f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + b(\boldsymbol{p}_w), \qquad (1)$$

where $f(\boldsymbol{x})$ is the monitor scene radiosity, and integrating over $\boldsymbol{x} \in \mathcal{S} = \{(x, D, z): x, z \in \mathbb{R}\}$ represents the combining of contributions from the entire scene at $\boldsymbol{p}_w$. The first weighting factor in the integrand models the radial falloff of flux density and two foreshortening effects: the wall patch relative to the direction of the incident light and the monitor pixel relative to the viewing angle of that pixel, with $\boldsymbol{n}_x$ and $\boldsymbol{n}_w$ the monitor and wall surface normals, respectively, and $\measuredangle(\cdot,\cdot)$ denoting the angle between its vector arguments. The second weighting factor, $V(\boldsymbol{x}; \boldsymbol{p}_w; \boldsymbol{p}_o)$, is a Boolean-valued *visibility function* that equals 1 when the path from $\boldsymbol{x}$ to $\boldsymbol{p}_w$ is unoccluded and equals 0 otherwise. The factor $\mu(\boldsymbol{x}, \boldsymbol{p}_w)$ describes the radiometric model for the monitor's variation with view angle[31]. The final term, $b(\boldsymbol{p}_w)$, represents the contribution at the visible wall from sources outside the modeled scene area $\mathcal{S}$. Eq. 1 is an instance of the *rendering equation* from computer graphics specialized to our setting[32].

With reflection from the visible wall assumed to be Lambertian, the reduced-resolution digital photograph of the visible wall is modeled with a discretization of Eq. 1 with $\boldsymbol{p}_\mathrm{w}$ taking $(126)^2 = 15876$ values in the camera's field of view (FOV). Namely, for each color channel we obtain a simple affine model $\mathbf{y} = \mathbf{A}(\boldsymbol{p}_\mathrm{o})\mathbf{f} + \mathbf{b}$, where the digital photograph is vectorized into a column vector, and the *light transport matrix* $\mathbf{A}(\boldsymbol{p}_\mathrm{o})$ has 15876 rows and a number of columns that depends on the attempted reconstruction resolution[31]. Forming an image of the hidden scene amounts to inverting the resulting linear system for each color channel.

The visibility function—equivalently, the presence of the occluder—is central to the conditioning of the inversion. Without an occluder, the weighting factors in Eq. 1 depend too weakly on $\boldsymbol{x}$ for well-conditioned recovery of $f(\boldsymbol{x})$[2,13,25,31]. The presence of an occluder introduces shadows and penumbrae that make it plausible that some image formation is possible, but everyday experience suggests this is extremely limited. In discretized form, without an occluder, the rows of $\mathbf{A}$ are too similar to enable well-conditioned inversion. Variations in visibility function $V(\boldsymbol{x}; \boldsymbol{p}_\mathrm{w}; \boldsymbol{p}_\mathrm{o})$ caused by the presence of an occluder improve the conditioning of $\mathbf{A}(\boldsymbol{p}_\mathrm{o})$ for inversion because its columns become more different. By treating a portion of the scene plane as resolvable if and only if it is visible in at least one camera measurement and invisible in at least one camera measurement, we define a *computational FOV* (Fig. 3)[31].
Recovering $\boldsymbol{p}_\mathrm{o}$ and $\mathbf{f}$ from the single snapshot camera measurement $\mathbf{y}$ is a nonlinear problem. Since the number of measurements (rows of $\mathbf{A}(\boldsymbol{p}_\mathrm{o})$) is large relative to the recoverable resolution of the hidden scene, the measurements $\mathbf{y}$ reside close to low-dimensional affine subspace that is dependent on the occluder position $\boldsymbol{p}_\mathrm{o}$ and background $\mathbf{b}$. The occluder position is estimated from $\mathbf{y}$ through

$$\hat{\boldsymbol{p}}_\mathrm{o} = \arg\max_{\boldsymbol{p}_\mathrm{o}} \|\mathbf{A}(\boldsymbol{p}_\mathrm{o})(\mathbf{A}(\boldsymbol{p}_\mathrm{o})^\mathrm{T}\mathbf{A}(\boldsymbol{p}_\mathrm{o}))^{-1}\mathbf{A}(\boldsymbol{p}_\mathrm{o})^\mathrm{T}\mathbf{y}\|_2^2, \qquad (2)$$

where $\mathbf{A}(\boldsymbol{p}_\mathrm{o})$ is the computed light transport matrix for an occluder position $\boldsymbol{p}_\mathrm{o}$; the omission of the unknown $\mathbf{b}$ does not greatly degrade the estimate[31]. The three estimates obtained by solving this maximization for each color channel are averaged to obtain a single $\hat{\boldsymbol{p}}_\mathrm{o}$.

Given the estimated occluder position $\hat{\boldsymbol{p}}_\mathrm{o}$, an estimate $\hat{\mathbf{A}} = \mathbf{A}(\hat{\boldsymbol{p}}_\mathrm{o})$ of the true light transport matrix $\mathbf{A}(\boldsymbol{p}_\mathrm{o})$ is computed. If the estimated occluder position were exactly correct and model mismatch and background contributions were inconsequential, pre-multiplying the vectorized measurements $\mathbf{y}$ (for each color channel) by the pseudoinverse $\hat{\mathbf{A}}^\dagger = (\hat{\mathbf{A}}^\mathrm{T}\,\hat{\mathbf{A}})^{-1}\,\hat{\mathbf{A}}^\mathrm{T}$ would yield the least-squares estimate of the hidden scene's RGB content. To improve robustness to noise and model mismatch, we exploit transverse spatial correlations prevalent in real-world scenes by promoting sparsity in the scene's gradient via total variation (TV) regularization[33]:

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \left\|\hat{\mathbf{A}}\mathbf{f} - \mathbf{y}\right\|_2^2 + \lambda\,\|\mathbf{f}\|_\mathrm{TV}, \qquad (3)$$

where the operator $\|\cdot\|_\mathrm{TV}$ denotes the TV semi-norm and λ is the TV-regularization parameter.

To further improve image quality, we take the differences of measurements $\mathbf{y}$ for (vertically) neighboring blocks of 16-by-16 pixels and also of corresponding rows of $\hat{\mathbf{A}}$ before solving an optimization problem analogous to Eq. 3[31]. Since any light originating from outside the

computational FOV has slow spatial variation, the background contribution **b** is approximately constant, i.e. $\mathbf{b}_{i+1} \approx \mathbf{b}_i \approx b$, for neighboring pixels and is thus approximately cancelled:

$$\mathbf{y}_{i+1} - \mathbf{y}_i \approx \left(\mathbf{a}_{i+1}^{\mathrm{T}}\mathbf{f} + b\right) - \left(\mathbf{a}_i^{\mathrm{T}}\mathbf{f} + b\right) \approx (\mathbf{a}_{i+1} - \mathbf{a}_i)^{\mathrm{T}}\mathbf{f}. \qquad (4)$$

This also provides some robustness towards ambient light, as verified by additional experiments included in the Supplementary Materials (31). Ultimately, one is able to produce an image of the computational FOV because all appreciable variations in y are due to this portion of the scene.

We propose also an alternative reconstruction method whereby we exploit a property of the light transport matrices as $\boldsymbol{p}_o$ is varied. Inaccurate occluder transverse position or depth leads to shifted or magnified (or minified) estimates of the displayed scene[31]. This observation is exploited to produce a multiplicity of additional reconstructions, which are nonlinearly combined to produce a single noise-reduced image[31].

Each hidden-scene patch was a block of 35-by-35 monitor pixels, and since the monitor has 1280-by-1024 resolution, $\lfloor 1280/35 \rfloor = 36$ by $\lfloor 1024/35 \rfloor = 29$ rectangular scene pixels were obtained, each of size 1.12 cm-by-1.05 cm. The scene was aligned to the screen's top and left edges, occupying 30.5 cm vertically and 40.3 cm horizontally. The distance from the camera to the visible wall was approximately 1.5 m, such that its FOV was 43.7 cm by 43.7 cm, centered at (74.1, 26.4) cm. With this configuration, several 36-by-29 scene-pixel test images were used to evaluate our computational periscope. One experimental scene was an anthropomorphic mushroom image with approximate dimensions 26 cm-by-19 cm (Fig. 4(a), top row). An exposure of 175 ms maximized signal strength while avoiding saturation, and 20 such exposures were taken and averaged, yielding effective exposure of 3.5 s. The snapshot was fed to a computer algorithm to produce a reconstruction (Fig. 4(c), top row) by estimation of occluder position, vertical differencing of data and light transport matrix, and image estimation. This snapshot (Fig. 4(b), top row) was also fed into another computer algorithm to produce a reconstruction (Fig. 4(d), top row) by estimation of occluder position, formation of 49 image estimates for a 7-by-7 array of postulated occluder positions, and nonlinear combining of the 49 estimates. For comparison, a reconstruction (Fig. 4(e), top row) was formed from the snapshot along with the actual occluder position. With unoptimized code on a desktop computer, initial occluder estimation took 18 minutes and subsequent hidden scene recovery (using the approximate background cancellation method) took an additional 48 seconds. Most of that computation time is for forming $\mathbf{A}(\boldsymbol{p}_o)$ matrices[31].

Reducing exposure time under 1 s enabled the capture of a 1 frame per second movie (Supplementary Video).

Results are provided for three additional scenes (Fig. 4, last three rows). Exposure time to maximize signal strength while avoiding saturation varied between 175 ms and 425 ms, and the average of 20 such exposures gave the inputs to the computational method (Fig. 4(b)). Estimated occluder positions are reported in supplementary Table S1.

Results for the initial scene (Fig. 4, first row) show that with our computational imaging method clearly resolves moderately-sized features like the white and red patches along with larger features like the head and yellow face; smaller features, such as eyes and unibrow, are visible with worse

accuracy. Similarly, for the second scene (Fig. 4, second row), even the white teeth and blue + on the hat are present in the reconstructions along with larger features like the face and hat. These two scenes demonstrate that measurements that are difficult to distinguish visually (Fig. 4(b)) may yield distinct and clearly identifiable reconstructions.

Occluder position estimates have roughly centimeter accuracy (Table S1). Reconstructions based on estimated occluder position (Fig. 4(c) and (d)) have similar quality to those based on known occluder position (Fig. 4(e)), demonstrating robustness to the lack of knowledge of occluder position.

The results show that the penumbra cast by an object may contain enough information to both estimate the position of the object and computationally construct an image of the computational FOV created by the object. In such a setting, we demonstrate 2D color NLOS imaging is possible with an ordinary digital camera, without requiring time-varying illumination and high-speed sensing.

References

1. Kirmani, A., Hutchison, T., Davis, J. & Raskar, R. Looking around the corner using transient imaging. In *Proc. IEEE 12th Int. Conf. Computer Vision*, 159–166 (2009).
2. Velten, A. *et al.* Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature Commun.* **3**, 745 (2012).
3. Gupta, O., Willwacher, T., Velten, A., Veeraraghavan, A. & Raskar, R. Reconstruction of hidden 3D shapes using diffuse reflections. *Opt. Express* **20**, 19096–19108 (2012).
4. Xu, K. *et al.* Image contrast model of non-line-of-sight imaging based on laser range-gated imaging. *Opt. Eng.* **53**, 061610 (2013).
5. Laurenzis, M. & Velten, A. Nonline-of-sight laser gated viewing of scattered photons. *Opt. Eng.* **53**, 023102 (2014).
6. Buttafava, M., Zeman, J., Tosi, A., Eliceiri, K. & Velten, A. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Opt. Express* **23**, 20997–21011 (2015).
7. Gariepy, G., Tonolini, F., Henderson, R., Leach, J. & Faccio, D. Detection and tracking of moving objects hidden from view. *Nature Photonics* **10**, 23–27 (2016).
8. Klein, J., Laurenzis, M. & Hullin, M. Transient imaging for real-time tracking around a corner. In *SPIE Proc. Electro-Optical Remote Sensing X*, 998802 (2016).
9. Chan, S., Warburton, R. E., Gariepy, G., Leach, J. & Faccio, D. Non-line-of-sight tracking of people at long range. *Opt. Express* **25**, 10109–10117 (2017).
10. Tsai, C.-y., Kutulakos, K. N., Narasimhan, S. G. & Sankaranarayanan, A. C. The geometry of first-returning photons for non-line-of-sight imaging. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 7216–7224 (2017).
11. Heide, F. *et al.* Non-line-of-sight imaging with partial occluders and surface normals. arXiv:1711.07134v3 [cs.CV] (2018).
12. O'Toole, M., Lindell, D. B. & Wetzstein, G. Confocal non-line-of-sight imaging based on the light-cone transform. *Nature* **555**, 338–341 (2018).

13. Kirmani, A., Jeelani, H., Montazerhodjat, V. & Goyal, V. K. Diffuse imaging: Creating optical images with unfocused time-resolved illumination and sensing. *IEEE Signal Process. Lett.* **19**, 31–34 (2012).
14. Heide, F., Hullin, M. B., Gregson, J. & Heidrich, W. Low-budget transient imaging using photonic mixer devices. *ACM Trans. Graphics* **32**, 1 (2013).
15. Heide, F., Xiao, L., Heidrich,W. & Hullin, M. B. Diffuse mirrors: 3D reconstruction from diffuse indirect illumination using inexpensive time-of-flight sensors. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 3222–3229 (2014).
16. Kadambi, A., Zhao, H., Shi, B. & Raskar, R. Occluded imaging with time-of-flight sensors. *ACM Trans. Graphics* **35**, 1–12 (2016).
17. Pawlikowska, A. M., Halimi, A., Lamb, R. A. & Buller, G. S. Single-photon three-dimensional imaging at up to 10 kilometers range. *Opt. Express* **25**, 11919–11931 (2017).
18. Kirmani, A. *et al.* First-photon imaging. *Science* **343**, 58–62 (2014).
19. Shin, D., Kirmani, A., Goyal, V. K. & Shapiro, J. H. Photon-efficient computational 3D and reflectivity imaging with single-photon detectors. *IEEE Trans. Computational Imaging* **1**, 112–125 (2015).
20. Altmann, Y., Ren, X., McCarthy, A., Buller, G. S. & McLaughlin, S. Lidar waveform-based analysis of depth images constructed using sparse single-photon data. *IEEE Trans. Image Process.* **25**, 1935–1946 (2016).
21. Rapp, J. & Goyal, V. K. A few photons among many: Unmixing signal and noise for photon-efficient active imaging. *IEEE Trans. Computational Imaging* **3**, 445–459 (2017).
22. Pediredla, A. K., Buttafava, M., Tosi, A., Cossairt, O. & Veeraraghavan, A. Reconstructing rooms using photon echoes: A plane based model and reconstruction algorithm for looking around the corner. In *Proc. IEEE Int. Conf. Computational Photography*, 1–12 (2017).
23. Pandharkar, R. *et al.* Estimating motion and size of moving non-line-of-sight objects in cluttered environments. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 265–272 (2011).
24. Naik, N., Zhao, S., Velten, A., Raskar, R. & Bala, K. Single view reflectance capture using multiplexed scattering and time-of-flight imaging. *ACM Trans. Graphics* **30**, 171:1–171:10 (2011).
25. Thrampoulidis, C. et al. Exploiting occlusion in non-line-of-sight active imaging. *IEEE Trans. Computational Imaging* **4**, 419–431 (2018).
26. Xu, F. *et al.* Revealing hidden scenes by photon-efficient occlusion-based opportunistic active imaging. *Opt. Express* **26**, 9945–9962 (2018).
27. Torralba, A. & Freeman, W. T. Accidental pinhole and pinspeck cameras: Revealing the scene outside the picture. *Int. J. Computer Vision* **110**, 92–112 (2014).
28. Bouman, K. L. *et al.* Turning corners into cameras: Principles and methods. In *Proc. 23rd IEEE Int. Conf. Computer Vision*, 2270–2278 (2017).
29. Baradad, M. *et al.* Inferring light fields from shadows. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 6267–6275 (2018).
30. Klein, J., Peters, C., Martín, J., Laurenzis, M. & Hullin, M. B. Tracking objects outside the line of sight using 2D intensity images. *Scientific Reports* **6**, 32491 (2016).
31. Supplementary materials.
32. Kajiya, J. T. The rendering equation. In *Proc. Conf. on Computer Graphics and Interactive Techniques (SIGGRAPH '86)*, 143–150 (1986).

33. Beck, A. & Teboulle, M. A fast iterative shrinkage-thresholding algorithm. *SIAM J. Imaging Sciences* **2**, 183–202 (2009).

Acknowledgments We thank F. Durand, W. T. Freeman, Y. Ma, J. Rapp, J. H. Shapiro, A. Torralba, F. N. C. Wong, and G. W. Wornell for discussions. This work was supported by the Defense Advanced Research Projects Agency (DARPA) REVEAL Program Contract No. HR0011-16-C-0030.

Author Contributions V.K.G. conceptualized the project, obtained funding, and supervised the research. C.S. and J.M.-B. developed methodology, performed experiments, wrote software, and validated results. C.S. produced visualizations. J.M.-B. wrote the original draft. C.S., J.M.-B., and V.K.G. reviewed and edited the paper.

Reprints and permissions information is available at www.nature.com/reprints.
Competing interests The authors declare no competing interests.
Computer codes used to generate the results in this manuscript are available on GitHub at https://github.com/Computational-Periscopy/Ordinary-Camera. Documentation on how to use the codes to reproduce the results are also included therein.
Correspondence should be addressed to V.K.G (email: goyal@bu.edu).
Supplementary Information including supplementary materials and methods, supplementary text, additional references, Figs. S1 to S19, Tables S1 and S2 and Supplementary Video is available in the online version of this paper.

Captions for Figures

Figure 1: **Experimental setup for computational periscopy.** Controlled by a laptop PC, the standard digital camera obtains a snapshot of the irradiance distribution on a visible imaging wall, which is induced by the penumbra of an occluding object due to light emanating from a scene of interest. The scene of interest is implemented with an LCD monitor for ease of performing experiments with many scenes. The snapshot is fed through a computer algorithm to recover an image of the scene of interest and an estimate of the hidden occluder position.

Figure 2: **Reconstruction procedure. a,** Camera measurements are de-interleaved to give RGB data from Bayer pattern measurements, with the green channels averaged. **b,** Estimated occluder position compared to true position. **c,** Hidden scene reconstruction from camera measurements and estimated occluder position.

Figure 3: **Computational field of view.** Portions of the scene plane that are visible to some part of the camera field of view (FOV) (region **a**) and occluded from some part of the camera FOV form the computational FOV (region **b**). Within the computational FOV, areas that influence the penumbrae of multiple occluder edges (region **c**) are better conditioned than those that influence only one.

Figure 4: **Reconstructions for different hidden scenes. a,** Four ground-truth scenes are displayed on the monitor; **b,** Camera measurement obtained for corresponding scene; **c,** A single reconstruction using the differential framework and the estimated occluder position; **d,** Reduced-

noise final reconstructions obtained by combining the reconstructions obtained from each of 49 postulated occluder positions, around the estimated value; **e,** Reconstructions obtained assuming the true occluder position is a priori known and using the same algorithm as in **d**.

Methods

Equipment details. The scenes were displayed on a Dell LCD monitor model 2001FP, which has 4:3 aspect ratio and 1280-by-1024 resolution. With no line-of-sight from monitor to camera, visibility was via a white Elmer's Foam Board, which is visually diffuse; any specular component that is present is not modeled and thus not directly exploited. The camera was a FLIR Grasshopper3 model GS3-U3-41S4C-C, which has 2016-by-2016 resolution (4.1 megapixels). It was used with a Tamron M118FM16 lens with 16 mm focal length and f/1.4 aperture. Control of scene and camera was through a Lenovo ThinkPad P51s laptop computer.

Data acquisition from camera. A Python script was used to control the data acquisition. It performed the following steps. First, a test scene was displayed on the scene monitor. Then, to form a snapshot with suppressed noise, multiple camera measurements were taken in succession and summed. A pre-calibrated shutter speed was used that approximately utilizes the full dynamic range of the camera per exposure. The final measurement for each block is a raw, 14-bit, 2016-by-2016 pixel image with the three color channels interleaved due to the RGBG Bayer filter pattern. In forming the three 1008-by-1008 color channel images, the two green channels were averaged. Each color channel was further averaged over 16-by-16 blocks to produce a 126-by-126 data matrix that was reshaped to column vector **y** of length 15876.

Computing a light transport matrix $\mathbf{A}(\boldsymbol{p}_o)$. Recall that the element $[\mathbf{A}(\boldsymbol{p}_o)]_{i,j}$ of the light transport matrix $\mathbf{A}(\boldsymbol{p}_o)$ represents the weighting of the contribution of light from hidden-scene pixel j™$\{1, 2, \ldots, N\}$ to camera FOV pixel j™$\{1, 2, \ldots, M\}$, where $M = 15876$ and N is the resolution at which recovery of the hidden scene is attempted (e.g., $N = 1044$ for producing a 29 by-36 reconstruction). For any calculation of $\mathbf{A}(\boldsymbol{p}_o)$, including the many computations for finding an estimate $\hat{\boldsymbol{p}}_o$, we performed the calculation in Eq. S11 (see Supplementary text) with $L = 64$.

Computing estimate $\hat{\boldsymbol{p}}_o$ of occluder position $\boldsymbol{p}_o$. Estimation of occluder position $\boldsymbol{p}_o$ was through a grid search approach outlined in Algorithm 1 (see Supplementary text). The algorithm is based on the camera measurements **y** being made to reside near the range of $\mathbf{A}(\hat{\boldsymbol{p}}_o)$, which is a low-dimensional subspace of the 15876-dimensional space of downsampled measurements. The desired estimate of the hidden occluder's position, $\hat{\boldsymbol{p}}_o$, is the one that minimizes the Euclidean distance between **y** and the range space of $\mathbf{A}(\hat{\boldsymbol{p}}_o)$ or, equivalently, maximizes the Euclidean norm of the orthogonal projection of **y** onto the range space of $\mathbf{A}(\hat{\boldsymbol{p}}_o)$[34]. In practice, poor conditioning of $\mathbf{A}(\boldsymbol{p}_o)$ for certain candidate occluder positions $\boldsymbol{p}_o$ makes it more robust to orthogonally project to the smaller subspace spanned by the left singular vectors of $\mathbf{A}(\hat{\boldsymbol{p}}_o)$ that are associated with the "significant" singular values, which are the ones that are within some factor κ™$(0,1)$ of the largest singular value.[1] Then if $\mathbf{A}(\boldsymbol{p}_o)$ is approximated by the truncated singular value decomposition (SVD) $\mathbf{U\Sigma V}^T$ using only significant singular values, Eq. 2 can be written using

[1] For instance, when $\boldsymbol{p}_o$ is such that the occluder does not cast a shadow in the camera's FOV, $\mathbf{A}(\boldsymbol{p}_o)$ is very poorly conditioned for inversion. Only some number $N_0 < N$ of the singular values

$$
\begin{aligned}
\|\mathbf{A}(\boldsymbol{p}_\mathrm{o})(\mathbf{A}(\boldsymbol{p}_\mathrm{o})^\mathrm{T}\mathbf{A}(\boldsymbol{p}_\mathrm{o}))^{-1}\mathbf{A}(\boldsymbol{p}_\mathrm{o})^\mathrm{T}\mathbf{y}\|_2^2 &= \left\|\mathbf{U\Sigma V}^\mathrm{T}\left((\mathbf{U\Sigma V}^\mathrm{T})^\mathrm{T}(\mathbf{U\Sigma V}^\mathrm{T})\right)^{-1}(\mathbf{U\Sigma V}^\mathrm{T})^\mathrm{T}\mathbf{y}\right\|_2^2 \\
&= \|\mathbf{U\Sigma V}^\mathrm{T}(\mathbf{V\Sigma U}^\mathrm{T}\mathbf{U\Sigma V}^\mathrm{T})^{-1}\mathbf{V\Sigma U}^\mathrm{T}\mathbf{y}\|_2^2 \\
&= \|\mathbf{U\Sigma V}^\mathrm{T}\mathbf{V\Sigma}^{-2}\mathbf{V}^\mathrm{T}\mathbf{V\Sigma U}^\mathrm{T}\mathbf{y}\|_2^2 \\
&= \|\mathbf{UU}^\mathrm{T}\mathbf{y}\|_2^2 = (\mathbf{UU}^\mathrm{T}\mathbf{y})^\mathrm{T}\mathbf{UU}^\mathrm{T}\mathbf{y} \\
&= (\mathbf{U}^\mathrm{T}\mathbf{y})^\mathrm{T}\mathbf{U}^\mathrm{T}\mathbf{U}(\mathbf{U}^\mathrm{T}\mathbf{y}) = \|\mathbf{U}^\mathrm{T}\mathbf{y}\|_2^2,
\end{aligned}
$$

as in Algorithm 1.

For a given discretization in Eq. S11, the cost of computing $\mathbf{A}(\boldsymbol{p}_\mathrm{o})$ for a single occluder position is $\Theta(LMN)$ and the cost of its SVD is $\Theta(N^2M)$ for $N < M$ [35]. Hence the cost of computing the occluder position estimate using Algorithm 1 is $\Theta((LMN + N^2M)n^3)$, where n is the total number of possible occluder positions considered.

While this approach is effective (Table S1), it can be computationally prohibitive for large n. With minimal loss in performance, dramatic cost reduction can be achieved by first searching a coarse grid of, say, $m \ll n$ points in each coordinate to obtain an initial estimate $\hat{\boldsymbol{p}}_\mathrm{o}$, and then refining that estimate by searching within its neighborhood. An initial coarse search is more accurate with projections to low-dimensional subspaces, and the finer searches are more accurate with projections to higher-dimensional subspaces. We found three searches, with κ first 0.75, then 0.5, and finally 0.05, to be effective. In addition, a coordinate ascent-based search provided further improvements in terms of computational complexity.

Computing scene estimate $\hat{\hat{\mathbf{f}}}$ using single occluder position estimate $\hat{\boldsymbol{p}}_\mathrm{o}$. Each time a scene estimate $\hat{\mathbf{f}}$ was to be computed from a single occluder position estimate, it was found by solving the TV-regularized optimization problem

$$
\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \left\|\hat{\mathbf{A}}\mathbf{f} - \mathbf{y}\right\|_2^2 + \lambda \|\mathbf{f}\|_\mathrm{TV}, \qquad \text{(S1)}
$$

which repeats Eq. 3 for convenience, for each color channel independently. Specifically, the isotropic total-variation semi-norm

$$
\|\mathbf{f}\|_{TV} = \sum_{i,j} \sqrt{(\mathbf{F}_{i,j} - \mathbf{F}_{i+1,j})^2 + (\mathbf{F}_{i,j} - \mathbf{F}_{i,j+1})^2}, \qquad \text{(S2)}
$$

was used, where $\mathbf{F} \in \mathbb{R}^{N_1 \times N_2}$ is $\mathbf{f}$ reshaped to the dimensions of the N_1-by-N_2 image we are reconstructing.

The optimization in Eq. S1 was solved using the fast iterative shrinkage-thresholding algorithm (FISTA) of Beck and Teboulle[32]. The algorithm requires an initial estimate $\mathbf{f}^{(0)}$, which was taken

---

will be significantly larger than zero. Hence, orthogonally projecting to the range of $\mathbf{A}(\boldsymbol{p}_\mathrm{o})$ will retain $N - N_0$ dimensions of $\mathbf{y}$ that depend deterministically but highly erratically on $\boldsymbol{p}_\mathrm{o}$; it is as if those directions are chosen uniformly at random, reducing the reliability of estimating the correct $\boldsymbol{p}_\mathrm{o}$.

as the least-squares estimate $(\widehat{\mathbf{A}}^{\mathrm{T}}\widehat{\mathbf{A}})^{-1}\widehat{\mathbf{A}}^{\mathrm{T}}\boldsymbol{y}$. The regularization parameter λ was decided upon empirically for each test scene.

Computing final scene estimate with spatial differencing. By noting that un-modelled, multi-bounce light and ambient background light will tend to be spatially slowly varying or close to constant in the camera measurements, we can augment the model with an approximately constant background term. Specifically, recall the model in Eq. S8:

$$\mathbf{y} = \mathbf{A}(\boldsymbol{p}_{\mathrm{o}})\mathbf{f} + \mathbf{b},$$

where $\mathbf{b} = [b_1, b_2, \ldots, b_i, b_{i+1}, \ldots, b_M]$ models the unknown background and M = 15876 is the number of camera FOV pixels. Taking the difference between two neighboring camera measurements, i.e. $\mathbf{y}_{i+1} - \mathbf{y}_i$, gives:

$$\mathbf{y}_{i+1} - \mathbf{y}_i \approx \mathbf{a}_{i+1}^{\mathrm{T}}\mathbf{f} + b_{i+1} - \mathbf{a}_i^{\mathrm{T}}\mathbf{f} - b_i \approx (\mathbf{a}_{i+1} - \mathbf{a}_i)^{\mathrm{T}}\mathbf{f} + (b_{i+1} - b_i).$$

Further imposing the slowly varying background assumption, $b_{i+1} \approx b_i$, implies that

$$\mathbf{y}_{i+1} - \mathbf{y}_i \approx (\mathbf{a}_{i+1} - \mathbf{a}_i)^{\mathrm{T}}\mathbf{f}. \qquad \text{(S3)}$$

Eq. S3 can therefore be rewritten in matrix-vector form as follows:

$$\mathbf{Dy} = \mathbf{DAf},$$

where **D**, is the so-called difference matrix. Similarly to Eq. S1 (equivalently, Eq. 3 in the main manuscript), we formulate and solve the optimization problem

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \left\|\mathbf{D}\widehat{\mathbf{A}}\mathbf{f} - \mathbf{Dy}\right\|_2^2 + \lambda \|\mathbf{f}\|_{\mathrm{TV}} \qquad \text{(S4)}$$

obtained by combining the new linear forward model with the usual TV prior. This new approach empirically exhibits increased robustness to model mismatch. As such, only slight improvements can be gained by combining multiple hidden-scene reconstructions using different postulated occluder locations. This slight improvement is significantly outweighed by the reduction in computational complexity of not having to compute a multiplicity of reconstructions.

Computing postulated occluder positions $\left\{\hat{\boldsymbol{p}}_{\mathrm{o},k}\right\}_{k=1}^{48}$ from occluder position estimate $\hat{\boldsymbol{p}}_{\mathrm{o},0}$. We wished to generate a set of postulated occluder positions such that the scene reconstructions using these positions exhibit predetermined horizontal or vertical shifts. Shifts in the *x* or *z* components of $\hat{\boldsymbol{p}}_{\mathrm{o}}$ lead to shifts of the entire scene reconstruction by an amount proportional to $D/(\hat{\boldsymbol{p}}_{\mathrm{o},0})_y$, following from an application of the similar triangles property.
Let $\boldsymbol{p}_{\mathrm{o},h,v} = ((\boldsymbol{p}_{\mathrm{o}})_x + hW(\boldsymbol{p}_{\mathrm{o}})_y/D, (\boldsymbol{p}_{\mathrm{o}})_y, (\boldsymbol{p}_{\mathrm{o}})_z + vH(\boldsymbol{p}_{\mathrm{o}})_y/D)$ denote an occluder position that results in an *h*-pixel horizontal shift and *v*-pixel vertical shift in the reconstructed scene, where *W* is the width and *H* the height of a scene pixel. Then a lexicographic ordering of the set

$$\left\{\boldsymbol{p}_{\mathrm{o},h,v} : (h,v) \in \{-6,-4,\dots,4,6\} \times \{-6,-4,\dots,4,6\}, (h,v) \neq (0,0)\right\}$$

gives $\left\{\hat{\boldsymbol{p}}_{\mathrm{o},k}\right\}_{k=1}^{48}$ as required. Note that $\hat{\boldsymbol{p}}_{\mathrm{o},0}$ is precisely $\boldsymbol{p}_{\mathrm{o},0,0}$.

Computing final scene estimate from scene estimates $\{\hat{\mathbf{f}}_k\}_{k=0}^{48}$. Once a set of scene estimates $\{\hat{\mathbf{f}}_k\}_{k=0}^{48}$ was computed, they were registered and combined. The scene estimates were generated using postulated occluder positions that result in intentional, integer-pixel shifts in the reconstruction. Thus, to align each of the scene estimates, the reverse shifts were applied with zero-padding to form a registered ensemble of estimates $\{\hat{\mathbf{f}}_k^{\mathrm{reg}}\}_{k=0}^{48}$. Examples are shown in Figs. S1 and S2.

To form the final estimate $\hat{\mathbf{f}}$, we combined the 49 registered estimates with a nonlinear procedure employed independently for each pixel. Consider the set of registered estimates $\{\hat{\mathbf{f}}_k^{\mathrm{reg}}\}_{k=0}^{48}$ for one pixel $\hat{\mathbf{f}}_i$. To balance the outlier rejecting property of the median with the variance reduction property of the sample mean, we select a parameter θ (empirically set to 0.25) and use the sample mean of the samples that are within θ of the median. More explicitly, let m_i denote the median of $\{\hat{\mathbf{f}}_k^{\mathrm{reg}}\}_{k=0}^{48}$. The estimate is

$$\hat{\mathbf{f}}_i = \mathrm{mean}\left(\left\{\hat{\mathbf{f}}_{i,k}^{\mathrm{reg}}, k \in \{0,1,\dots,48\} : \left|\hat{\mathbf{f}}_{i,k}^{\mathrm{reg}} - m_i\right| < \theta\right\}\right), \qquad \text{(S5)}$$

as illustrated in Fig. S3. This method of combining the ensemble is inspired by the alpha-trimmed mean[36].

Data availability

Raw data captured with our digital camera during experiments for this work are available on GitHub at https://github.com/Computational-Periscopy/Ordinary-Camera.

34. Vetterli, M., Kovačević, J. & Goyal, V. K. *Foundations of Signal Processing* (Cambridge Univ. Press, 2014).
35. Golub, G. H. & Van Loan, C. F. *Matrix Computations* (Johns Hopkins Univ. Press, Baltimore, MD, 1989), third edn.
36. Bednar, J. B. &Watt, T. L. Alpha-trimmed means and their relationship to median filters. *IEEE Trans. Acoust., Speech, Signal Process.* **ASSP-32**, 145–153 (1984).