CoLux: Multi-Object 3D Micro-Motion Analysis Using Speckle Imaging

BRANDON M. SMITH, PRATHAM DESAI, VISHAL AGARWAL, and MOHIT GUPTA,
University of Wisconsin–Madison

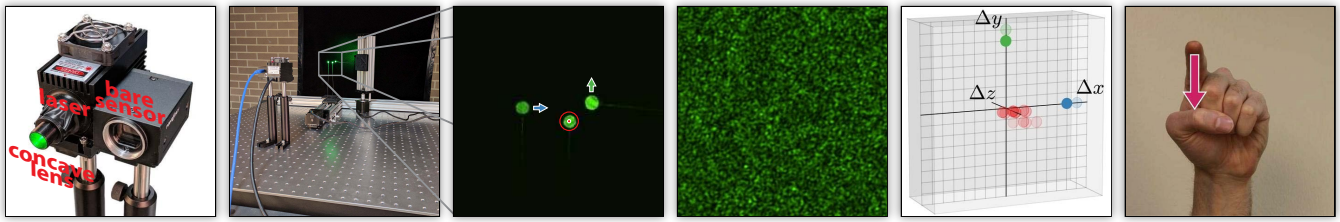


Fig. 1. From left to right: hardware prototype of CoLux consisting of a bare sensor and a laser diode, experimental setup for micro-motion measurement of three small objects at macroscopic distances from the sensor, close-up image of the objects captured by a conventional camera (not shown) with motion arrows ($+x$, $+y$, and $+z$) overlaid, speckle image of the same objects, 3D motion histogram showing microscopic motion modes computed from the speckle image sequence, example application of CoLux: subtle finger gesture recognition.

We present CoLux, a novel system for measuring micro 3D motion of multiple independently moving objects at macroscopic standoff distances. CoLux is based on speckle imaging, where the scene is illuminated with a coherent light source and imaged with a camera. Coherent light, on interacting with optically rough surfaces, creates a high-frequency speckle pattern in the captured images. The motion of objects results in movement of speckle, which can be measured to estimate the object motion. Speckle imaging is widely used for micro-motion estimation in several applications, including industrial inspection, scientific imaging, and user interfaces (e.g., optical mice). However, current speckle imaging methods are largely limited to measuring 2D motion (parallel to the sensor image plane) of a single rigid object. We develop a novel theoretical model for speckle movement due to multi-object motion, and present a simple technique based on global scale-space speckle motion analysis for measuring small (5-50 microns) compound motion of multiple objects, along all three axes. Using these tools, we develop a method for measuring 3D micro-motion histograms of multiple independently moving objects, without tracking the individual motion trajectories. In order to demonstrate the capabilities of CoLux, we develop a hardware prototype and a proof-of-concept subtle hand gesture recognition system with a broad range of potential applications in user interfaces and interactive computer graphics.

CCS Concepts: • **Computing methodologies** → **Computational photography**; • **Human-centered computing** → *Interaction devices*;

Additional Key Words and Phrases: Computational imaging; micro motion measurement; user interfaces; gesture recognition

ACM Reference format:
Brandon M. Smith, Pratham Desai, Vishal Agarwal, and Mohit Gupta. 2017. CoLux: Multi-Object 3D Micro-Motion Analysis Using Speckle Imaging. *ACM Trans. Graph.* 36, 4, Article 34 (July 2017), 12 pages.
DOI: http://dx.doi.org/10.1145/3072959.3073607

Authors' email: {bmsmith,prathamd,vishala,mohitg}@cs.wisc.edu.

DOI: http://dx.doi.org/10.1145/3072959.3073607

1 INTRODUCTION

The ability to measure extremely small non-rigid or multi-object motion has a wide range of applications, from biological cell imaging [Godinez and Rohr 2015], to hand-gesture recognition [Lien et al. 2016] for user interfaces. Measuring such micro-motions at macroscopic stand-off distances is generally not possible with conventional cameras and vision systems without using sophisticated optics. Furthermore, measuring multi-object or non-rigid motion is fundamentally more challenging than tracking a single object due to the considerably higher number of degrees of freedom, especially if the objects are devoid of texture.

We propose CoLux, a novel micro-motion measurement technique based on speckle imaging. Coherent light, on interacting with optically rough objects, creates a high-frequency speckle pattern. An example is shown in Fig. 1. Motion of the objects results in movement of the speckle pattern, and, given only the moving speckle pattern, it is possible to estimate the object motions. Previous speckle-based motion estimation techniques have been limited to *single*-body rigid motion estimation [Gregory 1978; Jacquot and Rastogi 1979; Jakobsen et al. 2012; Zalevsky et al. 2009], or sensor ego-motion estimation [Jo et al. 2015; Zizka et al. 2011]. Furthermore, most previous techniques are largely restricted to measuring 2D lateral motion (parallel to the sensor plane). Although some techniques [García et al. 2008; Jakobsen et al. 2012; Jo et al. 2015; Zizka et al. 2011] can measure axial motion, the motion sensitivity along the axial direction is 1-2 order of magnitude lower than that of lateral motion.

Multi-object motion analysis via bare sensor speckle imaging: In designing CoLux, our goal is to measure the motions of *multiple independently moving objects* with high <50 μm (microns) sensitivity along *all three spatial dimensions* (we assume the dominant motion of the objects to be translation, e.g., small cells moving in a medium, or finger tips performing a subtle gesture). The motion sensitivity of a speckle imaging system is directly proportional to the amount of sensor defocus. Our goal is to measure micro-motions. Therefore,

we use a bare (lens-less) sensor in CoLux. In addition to reducing cost and hardware complexity, the extreme defocus of a bare sensor results in extreme motion sensitivity. However, such large defocus also leads to overlap of multiple speckle patterns from different objects, and also cross-speckle due to interference of light from different objects. The resulting speckle pattern does not obey laws of speckle movement due to rigid object motion, and thus individual object trajectories cannot be recovered.

Our key observation is that, by using light sources with high temporal, but low spatial coherence, the cross-speckle terms vanish. By exploiting the statistical randomness of speckle patterns [Goodman 2000], we present a simple and efficient technique for recovering individual object motions, without needing to separate the individual speckle patterns. The proposed technique is based on a global scale-space analysis of the sequence of captured speckle images, which enables measuring multiple object motions simultaneously; previous motion measurement techniques based on local optical flow [Jo et al. 2015] cannot recover multiple motions. As an additional benefit, the global method can measure complex 3D compound motions with considerably higher axial motion sensitivity (<50 μm) as compared to previous approaches [García et al. 2008; Jakobsen et al. 2012; Jo et al. 2015; Zizka et al. 2011], while retaining high lateral motion sensitivity (<5 μm). Although CoLux achieves high motion sensitivity along all three axes, it cannot track multiple individual objects. Instead, it measures aggregate motion statistics of the scene, represented as a 3D motion histogram, which can be analyzed to recover the dynamic configuration of the scene, e.g., to recognize subtle hand gestures.

Hand gesture recognition system: To demonstrate the capabilities and potential applicability of CoLux, we develop a proof-of-concept hand gesture recognition system that can differentiate subtle finger movements. We are motivated by the fact that micro-motion gestures are difficult, if not impossible, to capture with conventional vision sensors alone. Our intention is not to compete directly with gesture recognition techniques based on conventional computer vision techniques that use RGB(D) cameras. Instead, CoLux offers complementary benefits: extreme motion sensitivity, but no spatial specificity, whereas conventional cameras offer spatial specificity but lower motion sensitivity. We take inspiration from Soli [Lien et al. 2016], a recent gesture recognition method based on millimeter-wave radar sensing, and measure aggregate motion of the scene (hand) at high temporal resolution, without tracking individual objects (fingers). While the overall goal of CoLux and Soli are similar, they differ significantly in the underlying imaging modality (speckle imaging vs. radar). While Soli has required significant custom hardware development, our prototype consists of off-the-shelf components, as shown in Fig. 1. Such components are widely used in low-cost and low-power devices such as optical laser mice, and are commercially available in compact form factors. We have trained a light-weight random forest classifier that can run in real-time on commodity hardware, and show recognition results on several single- and multi-finger gestures involving 3D motions. With our prototype system, we achieve an overall gesture-level accuracy of 83% for 5 different subjects over 7 gestures.

Contributions: This paper makes two main contributions:

(1) A novel theoretical model and practical method for measuring *multiple rigid or nonrigid motions*. We present a technique based on global scale-space analysis of the speckle sequence to recover small (5-50 μm) compound motions of multiple objects along all three axes. As a secondary benefit, this global technique achieves higher axial motion sensitivity (<50 μm) as compared to previous speckle based motion estimation approaches.
(2) Conceptual design, hardware prototyping, and a novel proof-of-concept system for high-speed *subtle gesture recognition*, with a broad range of potential applications in interactive computer graphics, gaming, and virtual reality.

2 RELATED WORK

Video motion magnification: Recently, techniques for magnifying small periodic motion in videos have been proposed [Davis et al. 2014; Wadhwa et al. 2013; Wu et al. 2012]. These can be considered *digital* motion magnification techniques because they digitally amplify motion in videos by suppressing noise in the desired temporal frequency bands. In contrast, CoLux amplifies motion sensitivity using *optical* imaging techniques *during capture* to measure micro-motions. Such subtle motions cannot be captured by a conventional camera, and thus cannot be magnified via digital post-processing. Video motion magnification techniques serve as visualization tools. On the other hand, the speckle patterns captured by the CoLux sensor are not suitable for human visualization, but can instead be used for quantitative motion measurement and analysis.

Non-rigid and multi-object motion analysis: Multi-object motion analysis techniques can be broadly classified into two categories. The first category encompasses techniques that track locations of individual objects over time. For example, most camera-based hand-tracking and gesture recognition systems [Weichert et al. 2013; Xu et al. 2015] explicitly estimate a hand's pose and skeletal structure. Tracking spatio-temporal trajectories of individual objects can provide highly detailed motion information but is not always possible if objects lack texture or if the motions are small. CoLux belongs to the second category of techniques that do not explicitly compute the 3D structure of the scene or track individual points. For instance, recent techniques based on alternative sensing modalities such as millimeter-wave radar [Lien et al. 2016] and radio waves [Zhao et al. 2014] recognize hand gestures by performing aggregate motion analysis of the entire scene over time. While previous techniques [Lien et al. 2016; Zhao et al. 2014] achieve limited motion sensitivity or require dedicated hardware, CoLux can achieve high motion sensitivity leading to fine-grained motion classification with off-the-shelf, typically inexpensive hardware technology.

3 SPECKLE FORMATION MODEL

This section is intended to introduce potentially unfamiliar concepts, and to establish notation. Speckle phenomena are well-studied, and more detailed introductions can be found in, e.g., [Goodman 2007].

Consider a temporally coherent light source (e.g., a laser diode) illuminating an optically rough surface Ψ, as shown in Fig. 2 (a). The light emitted by a coherent source is characterized by the underlying

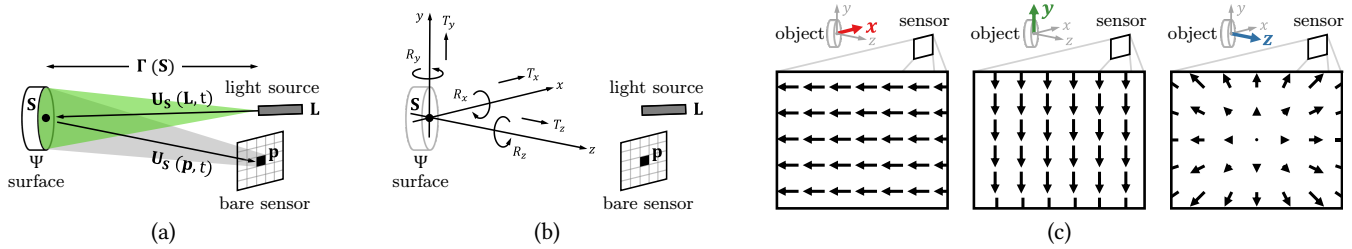


Fig. 2. (a) Speckle formation model. A bare sensor images a surface Ψ illuminated by a coherent light source located at **L**, which is approximately co-located with the sensor. (b) Speckle motion model for rigid object motion. As the object moves in 3D, the speckle pattern recorded by the sensor also moves. (c) Qualitative depiction of speckle motion. Lateral object motion results in speckle pattern shift. Axial motion results in speckle image contraction or expansion.

electric field **U**, which varies sinusoidally over time t:

$$\mathbf{U_S(L, t)} = U_S \cos(\omega t + \phi_S(t)), \tag{1}$$

where **L** is the spatial location of the light source, U_S is the amplitude ($U_S = \sqrt{L_S}$, L_S is the radiant intensity of the source), $\phi_S(t)$ is the phase of the light emitted by the source towards a scene point **S**, and $\omega = \frac{2\pi c}{\lambda}$ is the modulation frequency, where c is the speed of light, and λ is the wavelength of the coherent source. In practice, a coherent source emits a narrow band of wavelengths $[\lambda_{min}, \lambda_{max}]$. λ is the mean wavelength, i.e., $\lambda = \frac{\lambda_{min}+\lambda_{max}}{2}$.

Suppose the surface is imaged by a bare sensor, as shown in Fig. 2 (a). The electric field at pixel **p** due to the light reflected from the point **S** is given as:

$$\mathbf{U_S(p}, t) = \alpha(\mathbf{S})\, U_S \cos\left(\omega t + \phi_S(t) - 2\pi \frac{2\Gamma(\mathbf{S})}{\lambda} + \phi_S^r\right), \tag{2}$$

where $\alpha(\mathbf{S})$ encodes the light attenuation due to reflection at **S** and the intensity fall-off due to propagation. The phase of the emitted electric field is shifted by $2\pi \frac{2\Gamma(\mathbf{S})}{\lambda}$ during propagation along the path $\mathbf{L} \rightarrow \mathbf{S} \rightarrow \mathbf{p}$, where $\Gamma(\mathbf{S})$ is the distance of **S** from the source (we assume that the sensor and the source are co-located). ϕ_S^r is the change in phase due to reflection at point **S**.

Since the sensor is bare, we assume that each pixel collects light from every point on the surface. The total electric field $\mathbf{U(p)}$ at pixel **p** is then given by integrating the fields $\mathbf{U_S(p}, t)$ from all scene points over the surface Ψ:

$$\mathbf{U(p}, t) = \int_{\Psi} \mathbf{U_S(p}, t)\, d\mathbf{S} = \int_{\Psi} \beta(\mathbf{S}) \cos\left(\omega t + \widehat{\phi}_S(t)\right) d\mathbf{S}, \tag{3}$$

where $\beta(\mathbf{S}) = \alpha(\mathbf{S})\, U_S$ and $\widehat{\phi}_S(t) = \phi_S(t) - 2\pi \frac{2\Gamma(\mathbf{S})}{\lambda} + \phi_S^r$. The *speckle image* **I**, which is the measured image brightness due to this electric field, is given as:

$$\mathbf{I(p)} = \kappa \int_0^{\tau} \left(\mathbf{U(p}, t)\right)^2 dt, \tag{4}$$

where τ is the sensor integration time, and κ is a proportionality factor incorporating sensor gain. An example speckle pattern observed by illuminating small pieces of chalk is shown in Fig. 1.

Properties of Speckle

Statistical randomness: A speckle pattern due to reflection of coherent light from an optically rough surface is statistically random [Goodman 2000]. Intuitively, this is because each point on the illuminated surface acts as a secondary light source that emits spherical wavefronts. The total light received at a camera pixel is the superposition of all the wavefronts. The phase of each of these wavefronts varies rapidly as the path lengths (from scene point to sensor) change due to surface roughness, resulting in a statistically random speckle intensity distribution. This statistical randomness manifests as the following two properties of speckle images:

$$\underbrace{\boxed{(\mathbf{I} * \mathbf{I})(u, v) = \Lambda(u, v)}}_{\textbf{Auto-correlation property}}, \tag{5}$$

where $\mathbf{I}(u, v)$ is a speckle image, $[u, v]$ are image coordinates, and $*$ is the 2D correlation operator. $\Lambda(u, v) = \kappa\delta(u, v)$ is a scaled dirac-delta function $\delta(u, v)$, as shown in Fig. 3 (a), where $\kappa = \sum_{u,v} \left(\mathbf{I}(u, v)\right)^2$ is the square-norm of the speckle image; and

$$\underbrace{\boxed{(\mathbf{I}_1 * \mathbf{I}_2)(u, v) = 0}}_{\textbf{Cross-correlation property}}, \tag{6}$$

where $\mathbf{I}_1(u, v)$ and $\mathbf{I}_2(u, v)$ are speckle images due to reflection from two different rough surfaces Ψ_1 and Ψ_2, respectively. Intuitively, these two properties state that speckle images can be treated as mutually orthogonal random functions, i.e., with high probability, a speckle pattern is uncorrelated with anything but itself.

High spatial frequency: The mean 'size' χ of an individual speckle in a speckle image is proportional to the wavelength of light, and is given as $\chi \approx \frac{\lambda\Gamma}{\Omega}$ [Dainty 1975; Goodman 2000, 2007], where λ is the wavelength of light, Γ is the distance of the object from the sensor, and Ω is the area of the illuminated pattern. Speckle size depends on several other factors, including imaging geometry, surface properties including roughness and BRDF, and sensor properties including pixel size, aperture and focal length. For visible or NIR wavelengths ($\sim 380 - 800$ nm), the speckle size is limited only by the sensor pixel size, resulting in extremely high spatial frequencies.

4 SPECKLE MOTION MODEL: RIGID MOTION

In this section, in order to make the paper as self-contained as possible, we review the basics of the speckle motion model due to single object rigid motion. For detailed derivations, the interested reader is referred to previous work [Jacquot and Rastogi 1979; Jakobsen et al. 2012; Jo et al. 2015] and the supplementary technical report.

A well known result in optics is that small object motion does not change the speckle pattern, but only translates or scales it by a small

amount. This result, known as the *homology condition* [Gregory 1976, 1978; Tiziani 1972, 1978], or more recently as the *memory effect* [Judkewitz et al. 2015; Katz et al. 2012], has found extensive applications in speckle based metrology, including deformation measurement of large structures such as aircraft wings and submarine walls [Gregory 1978], imaging through scattering media [Bertolotti et al. 2012; Judkewitz et al. 2015; Katz et al. 2012], and camera-based ego-motion estimation [Jo et al. 2015].

Consider the imaging configuration shown in Fig. 2 (b). An approximately planar surface patch Ψ is illuminated by a coherent light source. Let the origin of the coordinate system be at a point $\mathbf{S}$ on Ψ. The z axis is perpendicular to the plane containing Ψ. Suppose the patch undergoes a small 6 DOF rigid motion given by a translation vector $T = [T_x, T_y, T_z]^\mathsf{T}$, and a rotation vector $T = [R_x, R_y, R_z]^\mathsf{T}$.

Let $\mathbf{I}(u, v)$ and $\mathbf{I'}(u, v)$ be the two speckle images captured by the sensor, before and after the motion, respectively. From the homology conditions stated above, it follows that the speckle pattern does not change, but only locally displaces (shifts) between the two images. Thus, the intensity at a pixel $\mathbf{I'}(u', v')$ in the image captured after motion is the same as the intensity at a different pixel $\mathbf{I}(u, v)$ in the image captured before motion. Assuming a paraxial sensor, the relationship between the speckle image displacement vector $[\Delta u, \Delta v] = [u' - u, v' - v]$ and the object motion is given by a linear system of equations [Jacquot and Rastogi 1979]:

$$\begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} = \mathbf{M_{trans}} T + \mathbf{M_{rot}} R\,, \tag{7}$$

where $\mathbf{M_{trans}}$ and $\mathbf{M_{rot}}$ are 2×3 matrices, whose entries depend on the geometric configuration (relative locations of the patch, sensor, and light source), as well as the radiometric characteristics of the imaging system (e.g., sensor pixel size, wavelength of light). In this paper, we assume that the scene is composed of infinitesimally small surface patches, and that the dominant motion of every patch can be approximated as a translation. We further assume that the sensor is bare (lens-less), and that the source and the principal point of the bare sensor are co-located along the z axis.[1] Under these simplifying assumptions, the relationship between speckle motion $[\Delta u, \Delta v]$ in the image and the object translation in 3D space is given by:

$$\underbrace{\boxed{\begin{pmatrix} \Delta u \\ \Delta v \end{pmatrix} = \begin{pmatrix} \frac{2}{p} & 0 & -\frac{u}{d} \\ 0 & \frac{2}{p} & -\frac{v}{d} \end{pmatrix} \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix},}}_{\textbf{Speckle Motion Model}} \tag{8}$$

where p is the side length of the sensor pixels (assuming square pixels), and d is the distance between scene point S and the light source. Eq. (8) was derived from Eq. (40) in [Jacquot and Rastogi 1979] (see our supplementary report for details), and is similar to Eq. (16) in [Jakobsen et al. 2012].

A qualitative depiction of speckle motion is shown in Fig. 2 (c). As can be seen from Eq. 8, and as observed in prior work (e.g., [Jakobsen et al. 2012; Jo et al. 2015]), lateral object motion (T_x or T_y) results in translation of the speckle image and axial object motion (T_z) results in radial expansion or contraction of the speckle image around the principal point.

[1]The results and techniques in the paper are valid for general geometric configurations. This assumption is made only for ease of exposition.

Increasing Motion Sensitivity using Speckle Imaging

Consider a conventional pin-hole sensor imaging a small planar surface patch located at a distance d along its optical axis. Then, under perspective projection, the image motion $\left[\Delta u_{\text{persp}}, \Delta v_{\text{persp}}\right]$ due to small object translation $T = [T_x, T_y, T_z]^\mathsf{T}$ is given as:

$$\begin{pmatrix} \Delta u_{\text{persp}} \\ \Delta v_{\text{persp}} \end{pmatrix} = \begin{pmatrix} \frac{f}{dp} & 0 & -\frac{u}{d} \\ 0 & \frac{f}{dp} & -\frac{v}{d} \end{pmatrix} \begin{pmatrix} T_x \\ T_y \\ T_z \end{pmatrix}, \tag{9}$$

where f is the focal length of the sensor. Comparing Eqs. 8 and 9, we define the motion sensitivity ratios as the ratio of the image-space motions between speckle and conventional imaging for a given motion along the x, y and z axes. Intuitively, the motion sensitivity ratio is the ability of a bare sensor speckle imaging system to magnify motion, as compared to a conventional perspective imaging system. Suppose the patch moves along the x axis by a unit distance, i.e., $T = [1, 0, 0]^\mathsf{T}$. The motion sensitivity ratio is then:

$$\mathcal{R}_x^{\text{sensitivity}} \quad = \quad \frac{\Delta u}{\Delta u_{\text{persp}}} = \frac{\frac{2}{p}}{\frac{f}{dp}} = \frac{2d}{f}\,. \tag{10}$$

The motion sensitivity ratio along the y axis is the same, and is derived similarly. Consider, for example, a sensor with $p = 6$ μm. Then, an object motion of 3 μm along the x and y axes will create a single pixel speckle shift. In comparison, for a perspective sensor with focal length $f = 20$ mm, and scene distance $d = 0.5$ meters, a motion of 0.15 mm will create a single pixel motion, resulting in a motion sensitivity ratio of 50:1.

Axial motion sensitivity: For unit motion along z axis, i.e., $T = [0, 0, 1]^\mathsf{T}$, the theoretical motion sensitivity ratio is:

$$\mathcal{R}_z^{\text{sensitivity}} \quad = \quad \frac{\Delta u}{\Delta u_{\text{persp}}} = \frac{\frac{-u}{d}}{\frac{-u}{d}} = 1\,, \tag{11}$$

which is considerably lower than the motion sensitivity ratio along x and y. For this reason, previous works on speckle-based motion analysis tend to focus on non-axial motion (e.g., [Tiziani 1978; Zalevsky et al. 2009]), and the few that estimate axial motion [García et al. 2008; Jakobsen et al. 2012; Jo et al. 2015; Zizka et al. 2011] achieve considerably less resolution.

5 3D (3-AXIS) COMPOUND MICRO-MOTION MEASUREMENT

In this section, we present a practical computational technique for measuring movement in speckle images, based on performing a simple, global scale-space analysis of the speckle image sequence. As we show in Section 6, such a *global* analysis technique is critical for measuring multiple object motions simultaneously; previous methods based on local optical flow [Jo et al. 2015] cannot recover multiple motions (see Figure 4 in the technical report). As a secondary benefit, this scale-space technique enables measuring 3D compound motions with higher axial motion sensitivity (<50 μm) as compared to local methods [Jo et al. 2015], while retaining high

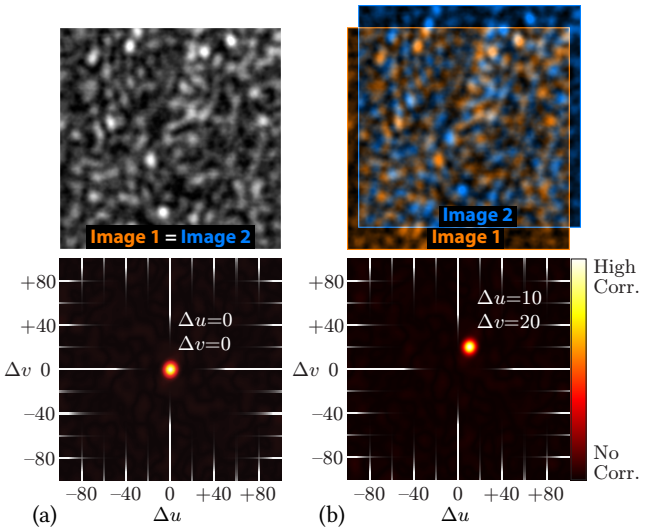


Fig. 3. Lateral motion measurement. (a) The 2D autocorrelation of a speckle image $\mathbf{I}_1$ is approximately a delta function centered at the origin (bottom). (b) $\mathbf{I}_2$ is a spatially translated version of $\mathbf{I}_1$, where the shift is given by $(\Delta u, \Delta v)$. As a result, the correlation $\mathbf{I}_1 * \mathbf{I}_2$ is a delta function centered at $(\Delta u, \Delta v)$.

lateral motion sensitivity (<5 μm), similar to what has been reported in previous work, e.g. [Synnergren 1997].

5.1 Measuring Lateral Object Translation

Consider a single object translating laterally (i.e., parallel to the sensor, or x-y, plane), so that the object motion between two successive speckle images $\mathbf{I}$ and $\mathbf{I}'$ is given by the translation vector $T = [T_x, T_y, 0]^\mathsf{T}$. The resulting speckle motion (Eq. 8) is given by the speckle flow vector $[\Delta u, \Delta v] = \frac{2}{p}[T_x, T_y]$. Since the speckle motion is constant over the entire image (not a function of u and v), it follows that $\mathbf{I}'$ is a spatially shifted (translated) version of $\mathbf{I}$. Consequently, due to the auto-correlation property of speckle images (Eq. 5), the 2D cross-correlation image $\mathbf{I}^{\mathbf{corr}} = \mathbf{I} * \mathbf{I}'$ is a shifted delta function, centered at $[\Delta u, \Delta v]$ [Synnergren 1997]:

$$\mathbf{I}^{\mathbf{corr}}(u, v) = \mathbf{I} * \mathbf{I}' = \Lambda(u - \Delta u, v - \Delta v)\,, \tag{12}$$

as shown in Fig. 3 (b). For example, given a pixel size $p = 6$ μm, a small object motion of 3 μm will result in the peak location getting shifted by 1 pixel. Thus, given pixel size p, we can estimate the scene motion (T_x, T_y) by finding the peak location $(\Delta u, \Delta v)$ in the cross-correlation image $\mathbf{I}^{\mathbf{corr}}(u, v)$.

5.2 Measuring Axial Object Translation

Consider an object translating axially (i.e., parallel to the z axis), so that the motion between two successive speckle images $\mathbf{I}$ and $\mathbf{I}'$ is given by the vector $T = [0, 0, T_z]^\mathsf{T}$. The resulting speckle motion (Eq. 8) is given by the vector $[\Delta u, \Delta v] = [\frac{-u}{d}T_z, \frac{-v}{d}T_z]$, where d is the distance of the object from the sensor. This speckle motion vector specifies a radial scaling (expansion/contraction) of the speckle image. Let $\mathbf{I}_\chi^{\mathbf{scale}}(u, v)$ be the scaled version of an image $\mathbf{I}(u, v)$, around its principal point $[c_u, c_v]$:

$$\mathbf{I}_\chi^{\mathbf{scale}}(u, v) = \mathbf{I}\left(u + \chi(u - c_u), v + \chi(v - c_v)\right)\,, \tag{13}$$

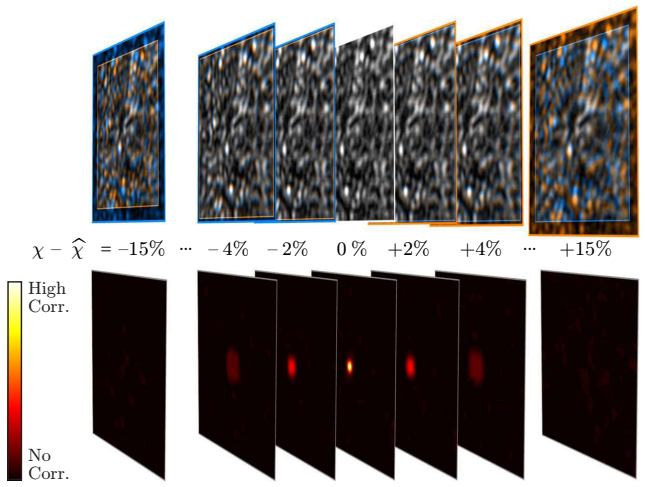


Fig. 4. Axial motion measurement. $\mathbf{I}_\chi^{\mathbf{scale}}$ (orange speckles) is a scaled version of $\mathbf{I}$ (blue speckles), which means $\mathbf{I}_\chi^{\mathbf{scale}} * \mathbf{I}$ is most like a delta function when χ equals the correct scale factor $\widehat{\chi}$ such that $\mathbf{I}_{\widehat{\chi}}^{\mathbf{scale}} = \mathbf{I}$.

where χ is the scale. Then, the speckle image $\mathbf{I}'$ after axial object motion is given as a scaled version of the original speckle image $\mathbf{I}$:

$$\mathbf{I}' = \mathbf{I}_\chi^{\mathbf{scale}}\,, \tag{14}$$

where the scale factor χ is given as $\chi = \frac{T_z}{d}$ (since the speckle motion vector $[\Delta u, \Delta v] = \frac{T_z}{d}[-u, -v]$, as discussed above). Thus, given scene depth d, we can estimate axial motion T_z by measuring the scale factor χ between $\mathbf{I}'$ and $\mathbf{I}$. If scene depths d are unknown, but the range of depths over which we measure T_z is small relative to d, i.e., $d \gg d_{\max} - d_{\min}$, then, the $\frac{1}{d}$ factor can be considered approximately constant, and the axial motion can be recovered up to a constant multiplicative factor.

Estimating χ using scale-space analysis of speckle images: The scale factor can be estimated by comparing $\mathbf{I}'$ with differently scaled versions of $\mathbf{I}$.[2] Let the *correct* scale be $\widehat{\chi}$, so that $\mathbf{I}' = \mathbf{I}_{\widehat{\chi}}^{\mathbf{scale}}$. In order to determine $\widehat{\chi}$, we perform a 1D search over a range of χ, e.g., $\chi = -0.20, -0.19, \ldots, 0.20$. For each candidate χ, we compute the 2D cross-correlation of $\mathbf{I}'$ with the scaled version $\mathbf{I}_\chi^{\mathbf{scale}}$:

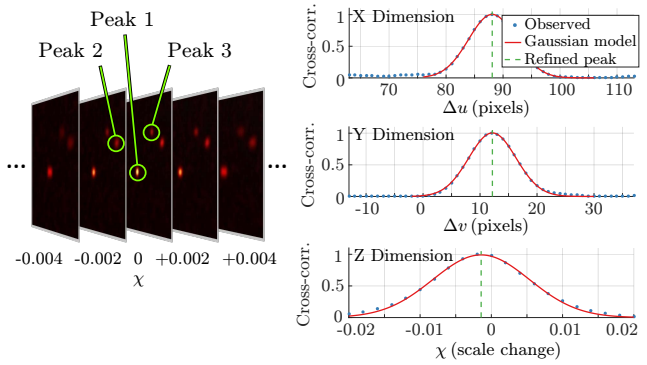
$$\mathbf{I}_\chi^{\mathbf{corr}} = \mathbf{I}_\chi^{\mathbf{scale}} * \mathbf{I}'\,. \tag{15}$$

Due to the auto-correlation property of speckle (Eq. 5), the correlation image $\mathbf{I}_{\widehat{\chi}}^{\mathbf{corr}}$ corresponding to the correct scale will be the most similar to a delta function, i.e., it will have the highest peak, as shown in Fig. 4. Thus, we can estimate $\widehat{\chi}$ by creating a stack of $\mathbf{I}_\chi^{\mathbf{corr}}$ images, and finding the image that has the highest peak:

$$\widehat{\chi} = \arg\max_{\chi}\ \text{peakVal}\left(\mathbf{I}_\chi^{\mathbf{corr}}\right)\,, \tag{16}$$

where peakVal($\mathbf{I}_\chi^{\mathbf{corr}}$) operator returns the height of the peak in image $\mathbf{I}_\chi^{\mathbf{corr}}$. In Section 7, we demonstrate that this global scale space analysis can measure axial motions with precision <50 μm.

[2] Scaling with bicubic interpolation is necessary in practice, as bilinear interpolation over-smooths the speckle pattern.



(a) Cross-correlation peaks (b) Refinement of Peak 2

Fig. 5. Local peak finding and refinement. (a) Local maxima are selected in the stack of cross-correlation images. (b) To increase precision, we observe that, in general, peaks exhibit a Gaussian-like profile. We therefore fit a Gaussian model to each peak, the centroid of which gives a precise sub-pixel and sub-scale location.

5.3 Measuring Compound Object Translation

The motion of an object undergoing compound translation (simultaneous lateral and axial motion), given by the translation vector $T = [T_x, T_y, T_z]^\mathsf{T}$, can be recovered by building upon the lateral and axial motion estimation methods discussed above. First, we create a stack of 2D cross-correlation images $\mathbf{I}_\chi^{\mathbf{corr}}$. From this stack, we find the image that has the highest peak, corresponding to the correct image scale $\widehat{\chi}$ (as discussed in Section 5.2), and then determine the highest peak location $(\Delta u, \Delta v)$, as discussed in Section 5.1. Then, the object translation vector can be recovered as:

$$T_x = \frac{p}{2}\Delta u, \quad T_y = \frac{p}{2}\Delta v, \quad T_z = d\widehat{\chi}\,. \tag{17}$$

5.4 Achieving Sub-Pixel and Sub-Scale Accuracy

The precision of the motion measurement method discussed above depends on how accurately we can locate the local maxima in the stack of scale-space cross-correlation images. One simple approach is to apply a maximum filter over the 3-dimensional stack and select the values that match the maximum filter output. This produces a set of 3D pixel coordinates that correspond to local maxima, as shown in Figure 5(a). The resolution of this simple approach is limited to one pixel for lateral motion, and to $\triangle\chi$, the difference between consecutive scale values, for axial motion.

In order to increase the precision, we observe that, in general, peaks exhibit a Gaussian-like profile, as shown in Figure 5 (b). We therefore fit a Gaussian model to each peak, the centroid of which gives a refined sub-pixel and sub-scale location. In practice this refinement step is not crucial for determining the Δu and Δv coordinates due to the large lateral motion sensitivity ratio. However, such sub-scale refinement is critical for axial motion due to low sensitivity, and computing several cross-correlation images across fine scale increments is expensive. Therefore, we fit a 1D Gaussian model across the scale dimension of each initial peak to refine the χ coordinate. This allows us to reduce the size of the scale-space search by increasing the scale increment, while avoiding fitting a more complicated 3D Gaussian model.

Computational complexity: To further improve the computational efficiency, we use the cross-correlation theorem [Smith 2002]:

$$f * g = \mathcal{F}^{-1}\left(\mathrm{conj}(\mathcal{F}(f)) \cdot \mathcal{F}(g)\right), \tag{18}$$

where f and g are functions (e.g., $f = \mathbf{I}_1$ and $g = \mathbf{I}_2$), $\mathcal{F}$ is the Fourier transform, $\mathcal{F}^{-1}$ is the inverse Fourier transform, conj is the complex conjugate, and $\cdot$ denotes element-wise multiplication. $\mathcal{F}$ and $\mathcal{F}^{-1}$ can be computed efficiently on a GPU [Kapinchev et al. 2015] via the fast Fourier transform (FFT) algorithm.

6 SPECKLE MOTION MODEL: MULTI-OBJECT MOTION

In this section, we present a speckle motion model for scenes with *non-rigid motion*. We model the scene as a collection of multiple independently moving objects, such that the inter-object distance is large as compared to the size of the objects. Each individual object is assumed to be moving rigidly. Such *discrete, dynamic, scattering center* model of the scene has been used recently in the context of finger gesture recognition [Lien et al. 2016], and is applicable in several settings, including material science [Rollie and Sundmacher 2010], and particle image velocimetry [Sinha 1988].

One obvious way to recover the motion of multiple independently moving objects is to separate them spatially in the captured image by using a lens-based imaging system [Zalevsky et al. 2009]. In such a system, the amount of lens defocus must be lower than the inter-object distance in order to ensure that images of different objects are spatially separated. However, the motion sensitivity of a speckle imaging system is directly proportional to the amount of defocus [Archbold and Ennos 1972; Gregory 1976; Jo et al. 2015]. This results in a tradeoff between spatial resolution and motion sensitivity. On one extreme, if the sensor is focused on the scene, different objects can be trivially separated, but the motion sensitivity is low. On the other extreme, using a bare sensor (extreme defocus) leads to high motion sensitivity, but the light reflected from all scene objects overlaps, rendering inapplicable the rigid body motion estimation techniques discussed in the previous section. *How can we overcome this fundamental tradeoff? Can we measure motion of multiple objects, while maintaining the high motion sensitivity and low system complexity of a bare sensor?*

Speckle formation for multiple objects: Consider two optically rough objects Ψ_1 and Ψ_2 being illuminated by a coherent source and imaged by a bare sensor. Then, from the speckle formation model (Eqs. 3 and 4), the total speckle image $\mathbf{I_{tot}}$ due to light reflected from both the objects is given as:

$$\mathbf{I_{tot}}(\mathbf{p}) = \kappa \int_0^\tau \left(\int_{\Psi_1,\Psi_2} \beta(\mathbf{S}) \cos\left(\omega t + \widehat{\phi}_S(t)\right) d\mathbf{S} \right)^2 dt\,. \tag{19}$$

Note that the inner integral is over scene points in both objects Ψ_1 and Ψ_2. By expanding the inner integral, and re-arranging the terms, the above equation can be written as:

$$\mathbf{I_{tot}}(\mathbf{p}) = \underbrace{\mathbf{I_1}(\mathbf{p})}_{\textbf{Speckle due to } \Psi_1} + \underbrace{\mathbf{I_2}(\mathbf{p})}_{\textbf{Speckle due to } \Psi_2} + \underbrace{\mathbf{I_{cross}}(\mathbf{p})}_{\textbf{Cross speckle term}}, \tag{20}$$

where $\mathbf{I_i}(\mathbf{p}) = \kappa \int_0^\tau \left(\int_{\Psi_i} \beta(\mathbf{S}) \cos(\omega t + \widehat{\phi}_S(t))\, d\mathbf{S}\right)^2 dt$ is the speckle image that the sensor would capture if it observed only the patch

$\Psi_i,\ i \in [1,2]$. The cross term $\mathbf{I_{cross}}(\mathbf{p})$ is given as:

$$\mathbf{I_{cross}}(\mathbf{p}) = 2\kappa \int_0^{\tau} \left(\int_{\Psi_1} \int_{\Psi_2} \beta_1 \beta_2\, \mathbf{c_1}\, \mathbf{c_2}\, d\mathbf{S}_1\, d\mathbf{S}_2 \right)^2 dt\,, \qquad (21)$$

where, for brevity, $\beta_i = \beta(\mathbf{S}_i)$, and $\mathbf{c_i} = \cos(\omega t + \widehat{\phi}_{S_i}(t))$. Intuitively, $\mathbf{I_{cross}}$ is the component of the total speckle image $\mathbf{I_{tot}}(\mathbf{p})$ due to interference between light reflected from Ψ_1 and Ψ_2. $\mathbf{I_{cross}}$ depends not only on the absolute motion of the individual objects, but also their relative motion and location. Consequently, $\mathbf{I_{cross}}$ *does not follow the homology conditions, and we cannot apply the laws of speckle movement due to rigid motion (Eq. (8)) to* $\mathbf{I_{tot}}$.

6.1 Eliminating the Cross Speckle Term

Our key observation is that the cross-term $\mathbf{I_{cross}}$ *vanishes if the light source has high temporal, but low spatial coherence.* The degree of spatial coherence of a light source is specified in terms of its coherence area A_C [Goodman 2000], which is defined as the area of a surface perpendicular to the direction of propagation (at a given distance from the source), over which the emitted light remains coherent with itself. Consider two scene points $\mathbf{S}_1$ and $\mathbf{S}_2$. Let $\phi_{S_1}(t)$ and $\phi_{S_2}(t)$ be the phases of light emitted towards them (Eq. 1). If $\mathbf{S}_1$ and $\mathbf{S}_2$ lie within the coherence area of the light source, then the relative phase is fixed over time, i.e., $\phi_{S_1}(t) - \phi_{S_2}(t) = \phi_{12}$. As a result, light reflected from these two points interferes, creating a speckle pattern. However, if the distance between the points is larger than the coherence area, the phases $\phi_{S_1}(t)$ and $\phi_{S_2}(t)$ fluctuate randomly with respect to each other. As a result, the cross term $\mathbf{I_{cross}}$, which contains a time integral of the product of cosines of the two phases, vanishes over time. The light reflected from these two points does not interfere; only the intensities are added together, similar to incoherent light. In this case, $\mathbf{I_{tot}}(\mathbf{p}) = \mathbf{I_1}(\mathbf{p}) + \mathbf{I_2}(\mathbf{p})$.

Light sources with low spatial coherence: If a mode-locked laser with high spatial coherence illuminates the scene, the cross term may not vanish even for distant objects. On the other hand, the coherence area of white-light sources may be too small to create speckle. For our approach, we use sources for which the coherence area is sufficiently large so that individual objects create a speckle pattern, but sufficiently small so that light reflected from different objects does not interfere. Such sources have been used recently for incoherent holography [Cossairt et al. 2014] and imaging through scattering media [Katz et al. 2014]. These can be implemented at low cost by, e.g., placing a narrow-band filter in front of a halogen lamp or via a laser diode [Cossairt et al. 2014].

6.2 Measuring Multiple Simultaneous Motions

Consider the speckle pattern due to two independently moving sobjects Ψ_1 and Ψ_2. As discussed above, we can eliminate the cross term by using light source with low spatial coherence. However, the total speckle image still consists of two speckle components, each moving independently, as illustrated in Figures 6 (a-c). **Please see the supplementary video for a visualization**. Let the total speckle image due to the individual patches $\Psi_i,\ i \in [1,2]$ before and after the motion of objects be:

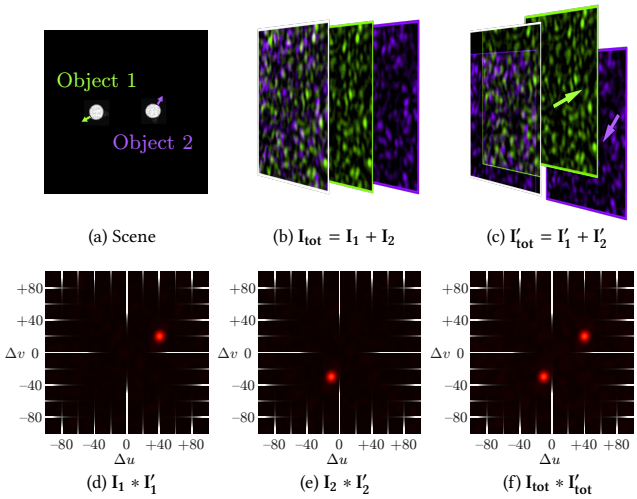


Fig. 6. Multi-object motion. (a) Microscopic movement of two objects. (b) The observed speckle image $\mathbf{I_{tot}}$ is the summation of the speckle images $\mathbf{I_1}$ and $\mathbf{I_2}$ due to Object 1 and Object 2, respectively. (c) Observed speckle image $\mathbf{I'_{tot}}$ after object motion. The **microscopic** movements of Object 1 and Object 2 result in **macroscopically** shifted speckle images $\mathbf{I'_1}$ and $\mathbf{I'_2}$. The individual speckle images are artificially colored here for illustration purposes; in reality, their color is uniform and reflects the wavelength of the light source. (d) The 2D cross-correlation result of $\mathbf{I_1}$ and $\mathbf{I'_1}$ due to Object 1 motion. (e) The 2D cross-correlation result of $\mathbf{I_2}$ and $\mathbf{I'_2}$ due to Object 2 motion. (f) The 2D cross-correlation result of $\mathbf{I_{tot}} * \mathbf{I'_{tot}}$ due to both motions. $\mathbf{I_1}$ and $\mathbf{I'_1}$ are uncorrelated with $\mathbf{I_2}$ and $\mathbf{I'_2}$, which is crucial for measuring multiple motions simultaneously.

$$\begin{aligned} \text{Before:}\ \ \mathbf{I_{tot}}(\mathbf{p}) &= \mathbf{I_1}(\mathbf{p}) + \mathbf{I_2}(\mathbf{p}) \\ \text{After:}\ \ \mathbf{I'_{tot}}(\mathbf{p}) &= \mathbf{I'_1}(\mathbf{p}) + \mathbf{I'_2}(\mathbf{p})\,. \end{aligned}$$

We cannot use the speckle motion model derived in Section 4 directly on $\mathbf{I_{tot}}$ and $\mathbf{I'_{tot}}$ because multiple speckle patterns are super-imposed in these images. Due to the random nature of speckle patterns, it is extremely challenging to separate $\mathbf{I_{tot}}$ and $\mathbf{I'_{tot}}$ into individual speckle patterns.

Our main insight is that *although the speckle patterns cannot be easily separated, their motion can be separated* by exploiting the cross-correlation property of speckle patterns (Eq. 6). Specifically, we compute the correlation of the speckle images $\mathbf{I_{tot}}$ and $\mathbf{I'_{tot}}$:

$$\begin{aligned} \mathbf{I^{corr}_{tot}} &= \mathbf{I_{tot}} * \mathbf{I'_{tot}} = (\mathbf{I_1} + \mathbf{I_2}) * (\mathbf{I'_1} + \mathbf{I'_2}) \\ &= \mathbf{I_1} * \mathbf{I'_1} + \mathbf{I_2} * \mathbf{I'_2} + \mathbf{I_1} * \mathbf{I'_2} + \mathbf{I_2} * \mathbf{I'_1}\,, \end{aligned}$$

where we have dropped the image indices u and v. According to the cross-correlation property of speckle patterns, the correlation between speckle patterns from two different optically rough surfaces is zero. As a result, $\mathbf{I_1} * \mathbf{I'_2} = \mathbf{I_2} * \mathbf{I'_1} = 0$, and we get:

$$\mathbf{I^{corr}_{tot}} = \mathbf{I^{corr}_1} + \mathbf{I^{corr}_2}\,, \qquad (22)$$

where $\mathbf{I^{corr}_i} = \mathbf{I_i} * \mathbf{I'_i}$ is the correlation image due to the motion of object Ψ_i, imaged individually. In general, if K independently

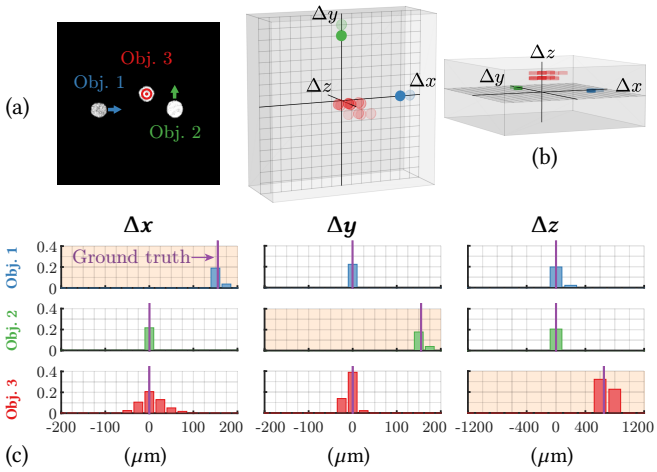


Fig. 7. 3D motion histogram. (a) The scene: three small objects moving along x, y, and z axes. (b) Two views of the 3D motion histogram showing the three different motion modes. (c) 1D motion histograms along each axis for each object. CoLux can recover multi-object motion with extremely high accuracy. Here the bins are set to $25\mu m \times 25\mu m \times 200\mu m$ for illustration purposes; average error is <5 μm laterally and <50 μm axially.

moving objects are imaged simultaneously, we get:

$$\underbrace{\mathbf{I}_{\mathbf{tot}}^{\mathbf{corr}} = \sum_{\mathbf{i}=1}^{K} \mathbf{I}_{\mathbf{i}}^{\mathbf{corr}}}_{\textbf{Multi-Object Speckle Correlation}} \tag{23}$$

This is an important equation. It states that under the given assumptions (small objects moving independently, illuminated by a light source with low spatial coherence), the *speckle correlation image due to multiple objects moving simultaneously is the sum of the correlation images due to the motion of objects imaged individually.* Since each individual correlation image is a shifted delta function (the shift corresponding to the motion of that object, as shown in Section 5), the total correlation image is a sum of shifted delta functions, as illustrated in Figures 6 (d-f). Each peak corresponds to the motion of a single object, and can be easily isolated. The 3D object motion estimation methods discussed in the previous section can then be applied to each peak, to create a *3D motion histogram* of the scene, where a non-zero bin value corresponds to the 3D motion of an object, [3] as shown in Figure 7. While the 3D motion histogram represents only the aggregate scene motion and cannot be used to track individual objects, it is an intuitive signature for multi-object 3D motion analysis, and is indicative of the dynamic scene configuration. For example, as we show in Section 8, the motion histogram from a moving hand can be used for accurate hand gesture recognition.

A note on the small, distant objects model: Our model of the scene consisting of small and distant objects is an idealization. In general, objects may have a finite spatial extent, and inter-object distances may be small. This will result in reduced contrast, and a non-zero cross term. In practice, if a light source with a narrow bandwidth and a small coherence area is used, the cross term remains negligible and the speckle contrast is sufficiently high. In the next section, we perform a detailed analysis of the motion sensitivity of our technique for single and multiple objects.

[3]Multiple objects moving with the exact same velocity lie in the same bin of the motion histogram, and cannot be measured independently. However, due to the high motion sensitivity, with high probability, different independently moving objects are assigned to different bins.

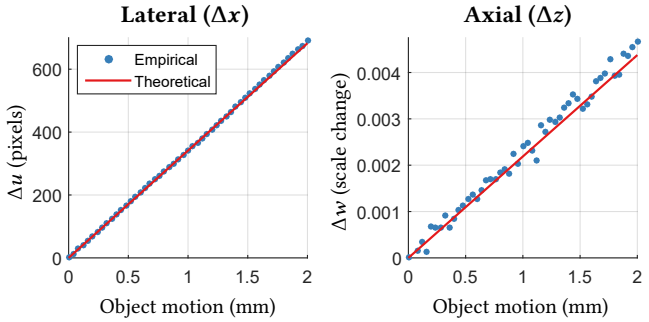


Fig. 8. Empirical observation vs. theoretical prediction: lateral and axial translations for a single object. Lateral motion measurement is more precise than axial motion measurement. In both cases, observed speckle movement agrees well with our model.

7 HARDWARE PROTOTYPE AND EXPERIMENTS

We have developed a hardware prototype for CoLux, and performed experiments in a controlled setting in order to validate the theoretical results in Sections 5 and 6.

7.1 Hardware Prototype

The left two images in Fig. 1 show our hardware prototype and experimental setup. We illuminate the scene with a 200 mW 532 nm DPSS laser from CivilLaser (0.1 nm linewidth, ~3 mm coherence length). The laser beam is defocused to illuminate the entire scene via two bi-concave lenses, a $f = -9$ mm (Thorlabs LD2568-C) and a $f = -15$ mm (Thorlabs LD2060-C) separated by 12 mm. A Point Grey Grasshopper 3 sensor (GS3-U3-23S6M-C) with a resolution of 1920×1200 captures the reflected light. The side of each pixel is $p = 5.86$ μm. The sensor is bare (lensless) except for a 532 ± 2 nm bandpass filter (Thorlabs FL532-10) used to block ambient light. We used general-purpose laboratory equipment for our experiments, which artificially inflates the cost of our prototype, but the hardware technology itself is typically inexpensive (e.g., used in optical laser mice). Ideally, the sensor must be co-located with the source. However, in our experiments, the scene depth is large (~50 cm) compared to the offset between source and sensor (4.5 cm), and the model in Section 4 closely approximates the observed behavior. In scenarios with close-range objects, a beam-splitter can be used for co-locating the sensor and the source.

7.2 Experiments

In the following controlled experiments, we used small 5 mm-diameter pieces of white chalk as the target objects. The surface of the chalk is microscopically rough, and has negligible sub-surface scattering. Each target was mounted on a linear stage (OpenBuilds C-Beam)

via a thin matte black rod. Unless note otherwise, the sensor-to-object distance was $d = 50$ cm. We used matte black velvet cloth as background so that most of the received light comes from the target object. However, this is not required: background speckle produces a cross-correlation peak at the origin, which can be easily removed.

7.2.1 Lateral Motion Measurement. The target object was moved along the x axis (y axis can be considered in a similar manner) in increments of 40 μm, and a speckle image was recorded after each increment. We measured the amount of speckle shift between pairs of frames using the technique described in Section 5.1. Over a 2 mm motion sequence, we measured the mean shift to be 13.98 pixels per 40 μm of lateral motion, or a slope of 0.348 pixels/μm. This agrees with the theoretical prediction (Eq. (8)) of 0.341 pixels/μm, as shown in Fig. 8 (a). We observe that the speckle motion model for lateral motion is quasi-invariant to scene geometry (e.g., depth, lateral offset) and object properties (e.g., size, shape, wide range of materials), as shown in the supplementary technical report.

To measure the lateral motion sensitivity ratio between a conventional camera and the CoLux sensor, we compared the first frame in the sequence and each subsequent frame: Frame 1 vs. Frame 2, Frame 1 vs. Frame 3, etc. We used a local correlation-based optical flow algorithm for the conventional camera to compute image motion. The top row of Fig. 9 compares the image motion for the conventional camera (left column), and the speckle motion (right column), for the same object x-translation. The measured lateral motion sensitivity ratio is approximately 60:1, consistent with the theoretical prediction in Eq. 10: $\mathcal{R}_x^{\text{sensitivity}} = 62.5$ for $d = 50$ cm and $f = 16$ mm.

7.2.2 Axial Motion Measurement. The target object was moved from $d = 50$ cm toward the sensor in increments of 40 μm. We measured the image scale factor χ between pairs of frames using the technique described in Section 5.2. For a 2 mm motion (Frame 1 vs. Frame 51), we measured the scale to be 4.66×10^{-3}, which agrees with the theoretical prediction (Eq. (8)) of 4.38×10^{-3}, as shown in Fig. 8 (b). The prediction here includes a small correction factor due to the 4.5 cm lateral offset between the coherent light source and the sensor. Strictly speaking, axial motion measurement is dependent on scene depth. However, if the axial motion is significantly smaller than d, then the depth can be considered approximately constant, making the axial motion estimates quasi-invariant to scene depths. Please see the supplementary report for additional details.

For measuring the sensitivity ratio $\mathcal{R}_z^{\text{sensitivity}}$ along the axial direction, we compared the first frame in the sequence and each subsequent frame: Frame 1 vs. Frame 2, Frame 1 vs. Frame 3, etc. We used a simple scale space algorithm for computing change in object size due to object motion as observed by the conventional camera. The bottom row of Fig. 9 compares the object motion required to create the same image scale change for the conventional camera (left column), and the CoLux sensor (right column). The units of image scale change are given in image size increments. While the theoretically predicted value for the axial motion sensitivity ratio is 1, our scale space speckle analysis technique (Section 5.2) achieves a significantly higher *measured* sensitivity ratio of approximately 15:1. This is because CoLux uses a bare sensor, and thus, can measure

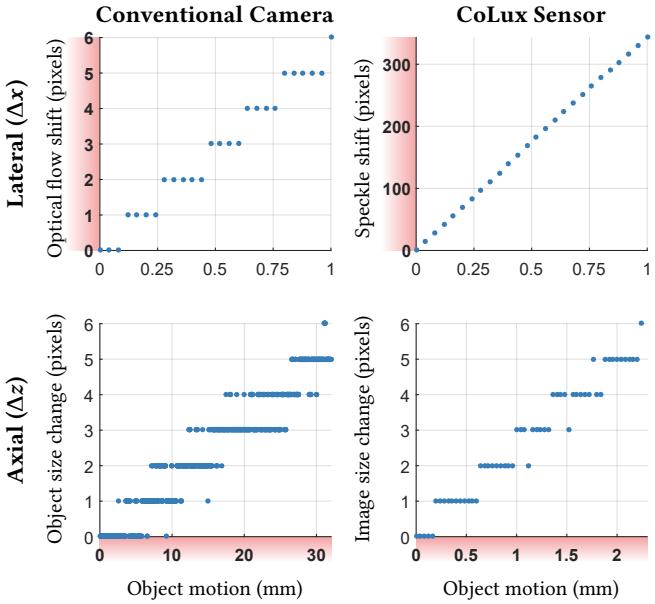
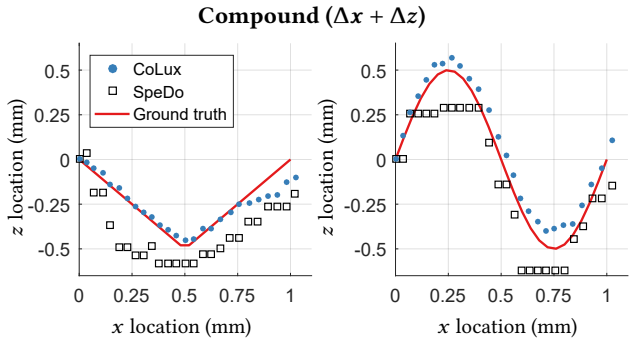


Fig. 9. **Top row:** Lateral (x axis) motion sensitivity. **Bottom row:** Axial (z axis) motion sensitivity toward the camera. **Left column:** Lateral and axial motions were computed by optical flow and change in object width, respectively, using images from a conventional cameral. **Right column:** Scene motion as measured by CoLux co-located with the conventional camera. Axial motion was measured in units of additive pixel size increments relative to the first frame in the sequence (e.g., 1200×1200 = 0, 1201×1201 = 1, etc.). CoLux measures object motion with an order of magnitude more sensitivity than a conventional camera. The lateral motion sensitivity ratio ~60:1 and the axial motion sensitivity ratio is ~15:1 in this case.

speckle scale change *over the entire image*, including the periphery (where the speckle motion is larger), and not just a small patch centered on the object (where the speckle motion is smaller), resulting in high motion sensitivity along all three axes for a wide range of real-world scenes and imaging scenarios.

7.2.3 Compound Motion Measurement. Fig. 10 shows tracking results for single-object trajectories that include a varity of compound (lateral+axial) motions. Tracking is a useful and challenging test case for compound motion estimation, but we emphasize that object tracking is not the focus of our work, as CoLux has no spatial specificity and cannot assign motions to multiple objects. The full trajectory can be estimated in this case because there is only one moving object. CoLux is able to measure compound motions with low mean absolute error: $\Delta x \leq 3.11$ μm (Δy estimation is analogous and has similar error) and $\Delta z \leq 18.36$ μm.

We compared CoLux's performance with SpeDo's [Jo et al. 2015] on the same input data. SpeDo employs a local optical flow algorithm to measure speckle motion. For fair comparison, we reduced the degrees of freedom of SpeDo from 6 to 3 (x, y, and z, no rotation). SpeDo requires two parameters: the size and number of patches used to compute local optical flow. As suggested in [Jo et al. 2015], we used 25 patches, each of size 41×41 pixels (empirically chosen to maximize performance). As seen in Fig. 10, CoLux and SpeDo produce similar



	Δx error	Δz error
CoLux	3.10 μm	17.25 μm
SpeDo	3.11 μm	49.24 μm

	Δx error	Δz error
CoLux	3.40 μm	18.36 μm
SpeDo	3.31 μm	74.90 μm

Fig. 10. Compound (lateral+axial) motion estimation. Starting at x=0, z=0 mm (camera at d=50 cm), a single object was moved in the x-z plane along the path shown in red. The trajectory was estimated by accumulating Δx and Δz computed between consecutive observations. The tables show mean absolute Δx and Δz errors. CoLux and SpeDo produce similar lateral motion estimates, but CoLux (global correlation analysis) significantly outperforms SpeDo (local optical flow) for small (e.g., 40 μm) axial motions.

lateral motion estimates, but CoLux (global correlation analysis) significantly outperforms SpeDo for small (e.g., 40 μm) axial motions. The speckle-based out-of-plane motion estimation method proposed by [Jakobsen et al. 2012] cannot handle compound motion, and they reported an average absolute axial error of 0.5 mm, which is an order of magnitude less accurate than CoLux, at much closer standoff distances (d < 4.9 cm).

7.2.4 Simultaneous Multiple Motion Measurement. In order to demonstrate the ability of CoLux to accurately measure motion of multiple simultaneously moving objects, we recorded three moving objects. Object 1 was moved from left to right along the x axis in increments of 160 μm per frame, Object 2 was moved downward along the y axis in increments of 160 μm per frame, and Object 3 was moved toward the sensor along the z axis in increments of 640 μm per frame. Using the technique described in Section 5, we measured the frame-to-frame motion of all three objects. Fig. 7 (b) shows a 3D motion histogram computed over a 100-frame sequence. The motion modes have been colored differently for illustration purposes (in reality, CoLux cannot assign motion modes to specific objects). Fig. 7 (c) shows the 1D motion histogram of each object separately along each axis. As expected, although the axial motion measurement estimates has more variance than lateral motion, CoLux can achieve high motion measurement accuracy for multiple objects, with an average error of < 20 μm. Note that previous speckle-based motion measurement techniques [Jakobsen et al. 2012; Jo et al. 2015] *can recover only a single rigid motion*, and are thus not applicable in such multi-object motion scenarios.

Practical limits on the number of objects: Although our model allows for an arbitrary number of moving objects in the scene, in practice the correlation peak intensities decrease with additional moving objects. In the extreme case, they can fall below the noise

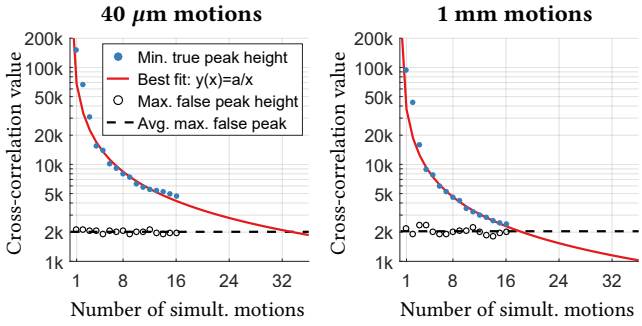


Fig. 11. Practical limits on the number of objects. The correlation peak intensities decrease with additional moving objects until the minimum true peak value falls below the maximum false (noise) peak value. We extrapolate the values of the minimum true peak value as a function of the number of motions by using a least square fit of the function $y(x) = \frac{a}{x}$, where x is the number of motions, $y(x)$ is peak value, and a is a scaling parameter. Comparing the two plots, we see that the number of motions that can be detected reliably decreases as the motion magnitude increases. Under controlled conditions (e.g., small chalk objects moving 40 μm, as shown on the left), CoLux can estimate up to ~32 unique motions.

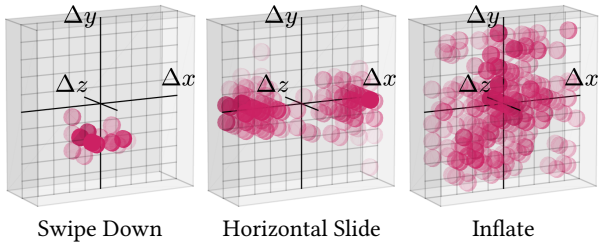


Fig. 12. 3D motion histograms for different gestures. Each cell represents 20 μm in each direction. CoLux correctly measures a single motion mode for the *Swipe Down* gesture (one finger), two motion modes for *Horizontal Slide* (two fingers), and multiple motion modes for *Inflate* (all fingers).

level, as shown in Fig. 11. For practical reasons (e.g., we have only three translation stages), we generated larger numbers of motions as follows. A single object was recorded multiple times, each time moving in a unique lateral direction (left, up, diagnoal, etc.). From Section 6, we know that the speckle pattern produced by multiple objects is the sum of the speckle patterns produced by each individual object. Therefore, to emulate a speckle sequence with M independent object motions, we averaged M individual speckle sequences. Under these controlled conditions, and assuming 40 μm motions, CoLux can accurately estimate up to ~32 unique simultaneous motions. The maximum number of allowable moving objects is strongly dependent on scene, illumination, and sensor characteristics. For example, Fig. 11 shows that the number of allowable motions decreases as the magnitude of the motions increases. We leave a full investigation as future work.

8 APPLICATION: GESTURE RECOGNITION

In order to demonstrate CoLux's motion measurement capabilities, we design and implement a prototype gesture recognition system

Fig. 13. Experimental gesture set. Arrows convey motion paths.

based on the proposed techniques. This application forms an interesting test case because hand gestures often involve subtle motion of multiple fingers. We consider hand *motion* gestures [Lien et al. 2016], as opposed to *pose* gestures, which involve recognizing the spatial configuration of the hand [Shotton et al. 2013]. Due to the subtlety of finger motion, conventional computer vision systems often fail to detect and recognize such gestures reliably. Indeed, CoLux offers complementary benefits: extreme motion sensitivity, which is good for subtle-motion gesture recognition, but no spatial specificity, whereas conventional cameras offer spatial specificity, which is good for pose recognition, but lower motion sensitivity. Fig. 12 shows the 3D motion histograms for three example gestures, involving one-, two-, and multiple-finger motions. The corresponding motion histograms clearly show one, two, and multiple motion modes, respectively. Note that the histogram bin-size is 20 μm, indicating the ability of CoLux to measure multi-object 3D micro-motions, thereby enabling fine-grained subtle gesture classification.

8.1 Gesture Set and Data Acquisition

Fig. 13 illustrates the 7 hand gestures used in our recognition system. **For a clearer demonstration, see the supplementary video.** These gestures were chosen due to their motion diversity: *Swipe Down* and *Swipe Up* feature small single-finger lateral motions, *Button Press* and *Button Release* feature small axial motions, *Horizontal Stretch* and *Vertical Stretch* feature simultaneous motions of two fingers, and *Inflate* features simultaneous motions of multiple fingers.

We asked 5 subjects to perform each gesture 5 times. Subjects were shown an example of each gesture, but were not trained otherwise. They performed each gesture ~0.5 meter from the sensor. Gestures were recorded by the CoLux sensor at 660 FPS and 256 × 256-pixel resolution. To construct our testing and training sets, we selected frames corresponding to each gesture instance and discarded others (e.g., during inter-gesture transitions). This selection was achieved by a conventional camera that concurrently recorded the scene at 60 FPS. The conventional camera was used only to manually segment gestures; for recognition we used only data from CoLux.

8.2 Feature Design and Extraction

For each gesture, we computed the 3D motion histogram between every pair of consecutive frames. In order to concisely represent the motion, we designed compact spatio-temporal features from the motion histograms. Each feature vector consists of the top M motion modes (M strongest cross-correlation peaks) from each motion histogram within an N-frame window. Each motion mode is represented as a 4-vector $(x, y, z, \text{intensity})$, consisting of the peak location and intensity (Section 5). The motion modes from a histogram are ordered by peak intensity and concatenated to form a vector of length $4M$. Each $4M$-length vector is then concatenated within the temporal window to form our final $4MN$-length feature vector, which is the input to the gesture classifier. Empirically, we found that $M = 10$ and $N = 200$ produced the best classification accuracy. This corresponds to $M = 10$ dominant independently moving objects, and a temporal duration of $N/660 \approx 0.3$ seconds.

	SD	SU	BP	BR	HS	VS	IF
SD	84%	2%	1%	0%	5%	4%	3%
SU	9%	82%	0%	1%	2%	4%	1%
BP	4%	1%	61%	10%	3%	7%	14%
BR	2%	5%	8%	52%	7%	11%	15%
HS	0%	0%	1%	5%	83%	4%	7%
VS	1%	1%	1%	2%	4%	89%	1%
IF	0%	0%	0%	0%	4%	7%	89%

(a) Confusion matrix

	SD	SU	BP	BR	HS	VS	IF
S1	100%	100%	100%	100%	90%	86%	99%
S2	94%	69%	53%	39%	69%	87%	100%
S3	91%	62%	40%	26%	79%	99%	69%
S4	39%	75%	3%	0%	86%	85%	100%
S5	90%	100%	92%	67%	86%	91%	88%

(b) Accuracy for five subjects (S1, S2, S3, S4, S5)

Table 1. Gesture recognition accuracy.

8.3 Random Forest Based Classification

CoLux is sufficiently general and can be used with a wide range of machine learning algorithms for classification, for example, support vector machines, hidden Markov models based on temporal time-series analysis and convolutional neural networks [Wang et al. 2016]. In our implementation, we used a random forest classifier (as implemented by the scikit-learn library: http://scikit-learn.org) because of its high computational efficiency, and low memory usage, with an eye on future implementation in low-power devices, such as cell-phones. Using a 2.60 GHz 32-core machine, test-time classification took ~0.27 seconds for ~9k feature vectors (samples). We defined a gesture instance as a set of 480 samples spanning ~0.7 seconds. To maximize the amount of training and testing data, we extracted multiple feature vectors from each gesture instance by shifting the N-frame sample window one frame at a time. This resulted in ~280 feature vectors (samples) per gesture instance. For training, samples from 4 out of the 5 trials of each subject gesture pair were used. Testing was performed using leave-one-trial-out cross validation (~40k training samples and ~9k testing samples).

Table 1 (a) summarizes the gesture recognition accuracy. The overall multi-class sample-level classification accuracy was 78%. The overall gesture-level classification accuracy (modal class label for each gesture trial) was 83%. Not surprisingly, gestures involving axial motion (*Button Press* and *Button Release*) are classified

with the lowest accuracy. Please see the supplementary video for a demonstration of our gesture recognition system.

Learning subject-specific gesture signatures: Table 1 (b) shows the subject-wise gesture classification accuracies. Subjects tended to perform gestures self-consistently, but in subtle different ways relative to one another. For example, Subject 4's performance of *Button Press* resulted in low recognition accuracy (3%) when included with other subjects. However, when the system was trained and tested on Subject 4 gestures only, the accuracy of *Button Press* increased to 62% (please see supplementary material for these results). This suggests that information from CoLux could be used as a signature for security applications, where both self-consistency and measuring subtle inter-subject variability can be highly informative features.

9 LIMITATIONS AND FUTURE WORK

There is little prior work on multi-object 3D micro-motion analysis using speckle imaging. This paper should be seen as a first step towards exploring this novel imaging modality, which opens many directions for further exploration.

Measuring rotation: So far, we have focused on measuring small object translation. In [Jo et al. 2015], a *single* rigid body rotation was measured by computing local patch-wise speckle motion and solving a linear system. Unfortunately, such a local technique is not applicable for multiple rotations. An interesting future direction is to extend the speckle motion model to measuring multiple rotations using techniques similar to those described in Sections 5 and 6.

Effect of material properties: If the scene consists of translucent or scattering objects, the homology conditions do not hold, i.e., speckle decorrelates and changes rapidly due to small object motion. This can make motion estimation challenging in the presence of highly translucent materials. Although skin is partially translucent, the speckle patterns remain correlated over large motions, thereby allowing accurate finger motion measurement. Given the interplay between material properties and motion, a promising avenue is to develop similar techniques to jointly recover motion and material.

Optimizing gesture recognition system: Our unoptimized subtle hand gesture recognition system is a proof-of-concept demonstration of the capabilities of CoLux, which leave several avenues for further optimizations. A possible next step is to incorporate conventional vision into the system, and explore more powerful machine learning techniques, such as deep neural networks, to potentially achieve significant performance improvement over our simple random forest classifier [Wang et al. 2016].

REFERENCES

E. Archbold and A. E. Ennos. 1972. Displacement measurement from double exposure laser photographs. *Optical Acta* (1972).

J. Bertolotti, E. G. van Putten, C. Blum, A. Lagendijk, W. L. Vos, and A. P. Mosk. 2012. Non-invasive imaging through opaque scattering layers. *Nature* 491 (2012).

O. Cossairt, N. Matsuda, and M. Gupta. 2014. Digital refocusing with incoherent holography. In *IEEE ICCP*.

J. C. Dainty. 1975. *Laser Speckle and Related Phenomena.* Springer.

A. Davis, M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand, and W. T. Freeman. 2014. The Visual Microphone: Passive Recovery of Sound from Video. *ACM Trans. Graph.* (2014).

L. Ek and N. E. Molin. 1971. Detection of the nodal lines and the amplitude of vibration by speckle interferometry. *Opt. Commun.* 2 (1971).

J. Engel, T. Schöps, and D. Cremers. 2014. LSD-SLAM: Large-Scale Direct Monocular SLAM. In *Proc. ECCV*.

M. Françon. 1979. *Laser Speckle and Applications in Optics.* Academic Press.

I. Freund. 1990. Looking Through Walls and Around Corners. *Physica A: Statistical Mechanics and its Applications* (1990).

J. García, Z. Zalevsky, P. García-Martínez, C. Ferreira, M. Teicher, and Y. Beiderman. 2008. Three-dimensional mapping and range measurement by means of projected speckle patterns. *Applied Aptics* (June 2008).

W J Godinez and K. Rohr. 2015. Tracking multiple particles in fluorescence time-lapse microscopy images via probabilistic data association. *IEEE Trans Med Imaging* (2015).

J. W. Goodman. 2000. *Statistical Optics.* Wiley-Interscience.

J. W. Goodman. 2007. *Speckle Phenomena in Optics: Theory and Applications.* Roberts and Company Publishers.

D. A. Gregory. 1976. Basic physical principles of defocused speckle photography: A tilt topology inspection technique. *Optics and Laser Technology* (1976).

D. A. Gregory. 1978. Topological Speckle and Surface Inspection. In *Speckle Metrology.* Chapter 8.

S. Gupta, D. Morris, S. Patel, and D. Tan. 2012. SoundWave: Using the Doppler Effect to Sense Gestures. In *Proc. ACM CHI.*

P. Jacquot and P. K. Rastogi. 1979. Speckle motions induced by rigid-body movements in freespace geometry: an explicit investigation and extension to new cases. *Applied Optics* (1979).

M. L. Jakobsen, H. T. Yura, and S. G. Hanson. 2012. Spatial filtering velocimetry of objective speckles for measuring out-of-plane motion. *Applied Optics* 51, 9 (2012).

K. Jo, M. Gupta, and S. Nayar. 2015. SpeDo: 6 DOF Ego-Motion Sensor Using Speckle Imaging. In *Proc. ICCV*.

B. Judkewitz, R. Horstmeyer, I. M. Vellekoop, I. N. Papadopoulos, and C. Yang. 2015. Translation correlations in anisotropically scattering media. *Nature Physics* (2015).

J. C. Silveira Jacques Junior, S. R. Musse, and C. R. Jung. 2010. Crowd Analysis Using Computer Vision Techniques. *IEEE Signal Processing Magazine* 27, 5 (2010).

K. Kapinchev, A. Bradu, F. Barnes, and A. Podoleanu. 2015. GPU Implementation of Cross-Correlation for Image Generation in Real Time. In *Proc. ICSPCS.*

O. Katz, P. Heidmann, M. Fink, and S. Gigan. 2014. Non-invasive single-shot imaging through scattering layers and around corners via speckle correlations. *Nature Photonics* 8 (2014).

O. Katz, E. Small, and Y. Silberberg. 2012. Looking around corners and through thin turbid layers in real time with scattered incoherent light. *Nature Photonics* 6 (2012).

J. Lien, N. Gillian, M. E. Karagozler, P. Amihood, C. Schwesig, E. Olson, H. Raja, and I. Poupyrev. 2016. Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar. *ACM Trans. Graph.* (2016).

S. Rollie and K. Sundmacher. 2010. Tracking the clustering dynamics in ternary particle mixtures by flow cytometry. *Powder Technology* 202 (2010).

J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. 2013. Real-time human pose recognition in parts from single depth images. *Commun. ACM* (2013).

S. K. Sinha. 1988. Improving the accuracy and resolution of particle image or laser speckle velocimetry. *Experiments in Fluids* (1988).

S. W. Smith. 2002. *Digital Signal Processing: A Practical Guide for Engineers and Scientists.* California Technical Publishing.

P. Synnergren. 1997. Measurement of three-dimensional displacement fields and shape using electronic speckle photography. *Optical Engineering* (1997).

H. J. Tiziani. 1972. A study of the use of laser speckle to measure small tilts. *Opt. Commun.* 5 (1972).

H. J. Tiziani. 1978. Vibration Analysis and Deformation Measurement. In *Speckle Metrology.* Chapter 5.

N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman. 2013. Phase-Based Video Motion Processing. *ACM Trans. Graph.* (2013).

S. Wang, J. Song, J. Lien, I. Poupyrev, and O. Hilliges. 2016. Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum. In *Proc. ACM UIST.*

F. Weichert, D. Bachmann, B. Rudak, and D. Fisseler. 2013. Analysis of the Accuracy and Robustness of the Leap Motion Controller. *Sensors* (2013).

H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman. 2012. Eulerian Video Magnification for Revealing Subtle Changes in the World. *ACM Trans. Graph.* (2012).

C. Xu, N. Ashwin, X. Zhang, and L. Cheng. 2015. Estimate Hand Poses Efficiently from Single Depth Images. *IJCV* (2015).

Z. Zalevsky, Y. Beiderman, I. Margalit, S. Gingold, M. Teicher, V. Mico, and J. Garcia. 2009. Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern. *Optics Express* (2009).

C. Zhao, K.-Y. Chen, M. T. I. Aumi, S. Patel, and M. S. Reynolds. 2014. SideSwipe: Detecting In-air Gestures Around Mobile Devices Using GSM Signal. In *Proc. ACM UIST.*

J. Zizka, A. Olwal, and R. Raskar. 2011. SpeckleSense: Fast, Precise, Low-cost and Compact Motion Sensing using Laser Speckle. In *Proc. ACM UIST.*