

THE CONTACT LAYER

The Contact Layer: Vision-Based Tactile Perception Where Line-of-Sight Ends

A camera under three colored lights becomes a geometry-and-force sensor, and the fingertip complement to a whole-body event skin.

David Jean Charlot, PhD

Dean of Physical AI · The Charlot Lab, Institute for Physical AI @ BMI

Correspondence: contact@physicalai-bmi.org · physicalai-bmi.org

Interactive companion: physicalai-bmi.org/research/charlot-lab#topic-contact

ABSTRACT. Every remote sense goes to zero at the moment of contact. The last millimeter, the forces that hold a grasp, whether a surface is slipping, and the micron texture that distinguishes a bolt from a screw are not knowable from across a room. Touch is the perception layer that begins exactly where remote sensing ends. This report reviews the vision-based tactile fingertip, in which an elastomer gel deforms against an object and a camera under three colored lights records only color, while geometry and force are computed, not measured, on the device. Photometric stereo inverts the colored image into a field of surface normals; the normals integrate into a depth map by solving a Poisson equation; a grid of printed markers carried by the gel reads shear and the onset of slip against the friction cone. Section 2 states the review method. Sections 3 through 5 develop the principle and its two computations. Section 6 covers on-device recognition and a slip reflex. Section 7 positions the micron fingertip beside the whole-body event skin of the printed-body corpus as one energy-accounted touch system, and relates both to tactile foundation models. The report reports no new experimental measurements.

1. Introduction

A camera resolves a scene, a microphone resolves a sound field, a depth sensor resolves range, and each of these degrades to nothing as an object approaches and finally occludes itself against a gripper. At the instant of contact the informative signal is no longer light or sound arriving from a distance but the mechanical state of the interface itself: how deep the object has pressed into the finger, in what direction it is pushing, whether the point of contact is holding or sliding, and what fine geometry the surface presents at the scale of tens of microns.

None of this is recoverable by any remote sensor, because the quantities are defined by the contact and did not exist before it.

For Physical AI, meaning embodied systems that sense, model, and act within a body's power budget, this gap is where manipulation succeeds or fails. A grasp is a claim about forces, and a remote sensor cannot check the claim. Touch is the perception layer that begins exactly where remote sensing ends. This report reviews one concrete and now widely reproduced way to build that layer: the vision-based, or optical, tactile fingertip, in which a soft surface is watched from the inside by a small camera and the mechanics of contact are recovered by computation rather than by an array of discrete force transducers. The device is attractive precisely because it converts a mechanical measurement problem into an imaging problem, for which cheap, high-resolution, low-power hardware already exists.

2. Scope and method

This is a review and a research position. It surveys published work on vision-based tactile sensing and the classical computer-vision and grasping results the method rests on, drawing on peer-reviewed papers, preprints, and primary hardware disclosures. For each claim it favors a primary source. The report reports no original experiments; the interactive companion referenced below is a simulation of the sensing pipeline built for teaching, not a measurement of a physical device, and where the text describes device behavior it attributes it to a cited source. The report's principal limitation is that optical tactile sensing is an active area with many competing gel and illumination geometries, and quantitative comparisons across designs remain difficult; the survey therefore emphasizes the shared operating principle over any one implementation, and marks where a capability is a prototype or a design proposal rather than an established result.

3. The vision-based tactile principle

The core idea is retrographic: coat the far side of a transparent elastomer with an opaque, matte membrane, illuminate that membrane from within, and photograph it with a camera behind the gel. Because the membrane is opaque, the camera never sees the object; it sees only the membrane's own surface as the object presses the gel into a relief of the contact. The membrane's reflectance is engineered to be nearly Lambertian, so the image encodes surface orientation rather than the object's color or its own texture, which cancels the confounds that make ordinary shape-from-shading unreliable^[1]. The GelSight family of sensors established that this arrangement yields tactile images at a spatial resolution far finer than the human fingertip, sufficient to read surface texture, printed characters, and the thread of a screw^[2].

Illumination is the design's second idea. The membrane is lit from three or more directions, each with a distinct color, typically red, green, and blue from the sides. A single color channel then behaves like an image taken under a single directional light, so one exposure of the color camera captures what would otherwise require three separate monochrome exposures. This is what makes the geometry recovery a single-shot computation and keeps the sensor fast enough for closed-loop control. Compact, low-cost realizations such as DIGIT packaged this

principle into a fingertip suited to in-hand manipulation^[3], and later multimodal fingertips such as DIGIT 360 added many co-located channels, including a hemispherical field and additional contact modalities, on the same optical base^[4]. Figure 1 shows the full pipeline from the colored image to the quantities a controller consumes.

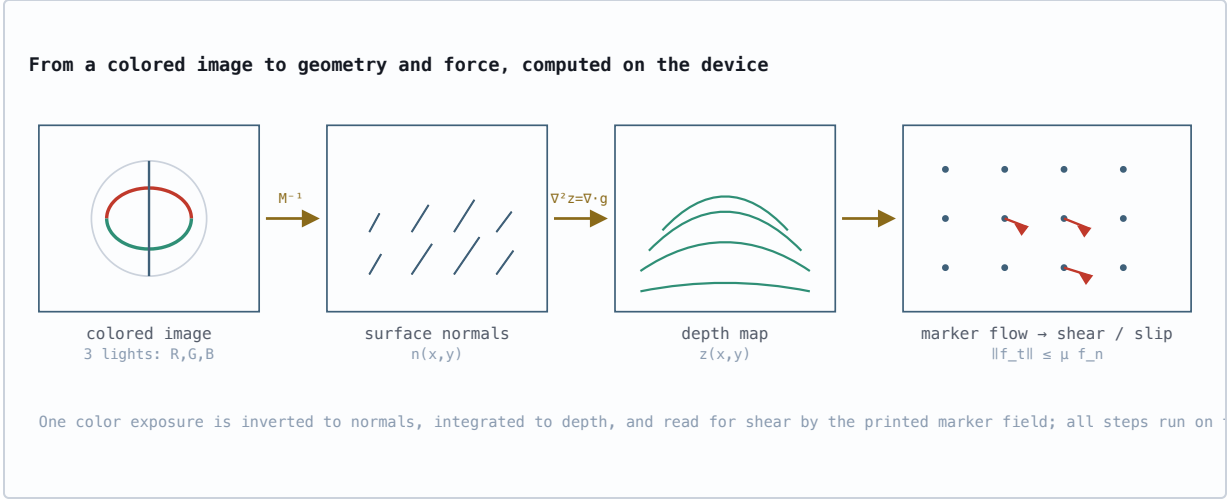


Figure 1. The vision-based tactile pipeline. A single colored image is inverted to a field of surface normals by photometric stereo, the normals are integrated to a depth map by solving a Poisson equation, and the displacement of a printed marker grid embedded in the gel reports shear and the onset of slip relative to the friction cone. An interactive version of the pipeline is at physicalai-bmi.org/research/charlot-lab.

4. Geometry computed, not measured: photometric stereo to normals to depth

The first computation recovers surface shape from the colored image. Under the Lambertian model, the intensity a pixel returns under a directional light of unit direction $\mathbf{l}$ is proportional to the cosine of the angle between the light and the local surface normal $\mathbf{n}$, that is $I = \rho (\mathbf{l} \cdot \mathbf{n})$, where ρ absorbs albedo and light intensity. Photometric stereo, introduced by Woodham in 1980, observes the same point under several known light directions and solves the resulting linear system for the normal^[5]. In the vision-based tactile sensor the three lights are the color channels, so the three observations arrive in one exposure. Collecting the red, green, and blue intensities of a pixel into a vector and the corresponding light-and-color mixing into a matrix M , the per-pixel recovery is a single inversion:

$$\mathbf{n} = \frac{M^{-1} [R, G, B]^T}{\|M^{-1} [R, G, B]^T\|}.$$

In practice M is not derived from first principles but calibrated once, by pressing a sphere of known radius into the gel and tabulating the mapping from color to normal, which folds the real reflectance, the color crosstalk, and the lighting nonuniformity into a lookup. The output is a dense field of unit normals $\mathbf{n}(x, y)$ over the sensor's image, at the camera's full resolution. This is the sense in which geometry is computed and not measured: no element of

the sensor reports a height, yet a height field is recoverable because orientation has been recorded everywhere.

The second computation turns orientation into shape. Writing the normal as $\mathbf{n} = (n_x, n_y, n_z)$, the surface gradient is $\mathbf{g} = (p, q)$ with $p = -n_x/n_z$ and $q = -n_y/n_z$. A depth map $z(x, y)$ whose gradient matches $\mathbf{g}$ satisfies, in the least-squares sense that best tolerates measurement noise and slight non-integrability, the Poisson equation

$$\nabla^2 z = \nabla \cdot \mathbf{g} = \frac{\partial p}{\partial x} + \frac{\partial q}{\partial y},$$

which is solved on the pixel grid under natural boundary conditions. Enforcing integrability this way is the standard route from a normal field to a surface and was placed on a firm footing by Frankot and Chellappa^[6]; on a regular grid the solve is a single sparse linear system or an FFT, cheap enough to run at camera frame rate. The result is a depth image of the contact accurate to the tens of microns, from which local curvature, edges, and fine texture follow directly.

5. Force and slip: the marker field and the friction cone

Geometry alone does not tell a controller whether a grasp will hold. For that the sensor must read force, and specifically the tangential force at the contact relative to what friction can supply. Vision-based sensors obtain this from a second structure printed into the gel: a regular grid of markers, visible to the same camera, that moves with the elastomer. Normal load presses the markers apart in a characteristic bloom; tangential load shears the grid; and the displacement field of the markers is, to first order, a readout of the surface traction distribution. Tracking the markers between frames is an optical-flow problem the sensor already has the hardware to solve.

The quantity that governs stability is the relation between the tangential and normal components of the contact force. Coulomb friction admits a contact without sliding only while the tangential load stays within the friction cone,

$$\|\mathbf{f}_t\| \leq \mu f_n,$$

where f_n is the normal force, $\mathbf{f}_t$ the tangential force, and μ the coefficient of friction. As a grasp is loaded, the marker field reports the ratio $\|\mathbf{f}_t\|/f_n$ climbing toward the cone's edge. Incipient slip appears before gross sliding as a partial slip: the periphery of the contact patch begins to move while the center still sticks, so the marker field shows a growing annulus of displacement around a stationary core. Detecting that signature is the earliest reliable warning that a grasp is about to fail, and it is exactly the regime a good controller wants to act in. The friction cone and the conditions for a stable, force-closed grasp are classical results in the grasping literature^[7]; the contribution of the optical fingertip is to make the cone observable in real time from an image.

6. On-device recognition and the auto-grip slip reflex

Because the sensor's native output is an image, the mature machinery of on-device vision applies unchanged. A small convolutional or transformer model can classify the contacted surface, estimate object pose from the depth patch, or regress the contact force, all from the tactile image and all within a fingertip's power budget. The most useful of these is a fast reflex rather than a full recognizer: a lightweight detector watching the marker field for the onset of partial slip, wired to increase grip force the moment the annulus of slip appears. This closes a loop entirely within the hand, at a latency set by the camera frame time rather than by any round trip to a central policy, and it is the tactile analogue of the stabilizing reflexes that let a biological hand hold a cup without crushing it. Neuromorphic and event-based implementations push this reflex toward microsecond latency and microwatt power by reporting only the pixels that change, and recent work demonstrates incipient-slip detection on such an event-driven substrate with a biomimetic fingertip morphology^[8].

A separate development concerns generalization across sensors. Every gel geometry, illumination, and marker layout produces a slightly different image, which historically meant that a model trained on one fingertip did not transfer to another. Tactile foundation models attack this directly: the Transferable Tactile Transformer learns a shared representation across diverse sensors and tasks, so that recognition and control skills acquired on one device carry to others^[9]. This matters for an institute that intends to teach the method rather than sell one sensor, because it decouples the perception skill from the particular fingertip a student happens to build.

7. Position: one touch system at two scales

The optical fingertip is not a whole-body solution, and it is not meant to be. Its virtues, micron resolution and dense force readout, are exactly what a fingertip needs and exactly what would be wasteful spread over a square meter of a robot's shell. The printed-body corpus of the lab pairs it with the opposite instrument: a whole-body magnetic event skin that reports coarse, sparse contact events over large area at very low power, saying that something touched here and roughly how hard, without resolving its texture. The two are complementary along the same axis that organizes the lab's view of computation, the axis of energy per decision. Micron detail is purchased with an imaging pipeline where dexterity needs it, at the fingers; an energy-accounted event skin covers everywhere else at the cost of a switch.

The proposal is a single touch system read at two resolutions: a computed, high-resolution optical fingertip where the hand meets the world, and a sparse, event-driven skin over the rest of the body, both priced in the same currency of energy per reported event. This report treats that pairing as a design position, not a demonstrated system.

Read this way, the contact layer is the perceptual complement to remote sensing rather than a competitor to it. A remote sensor guides the hand to the object; the fingertip takes over at the last millimeter, where the remote signal has gone to zero, and reports the geometry and forces

that only exist once contact is made. Tactile foundation models then let the skill learned on one fingertip transfer across a fleet, so the layer is a body-wide capability and not a per-device calibration. The claims here that are established are the sensing principle and its two computations, all of which rest on the cited primary work; the claims that are positions, and are marked as such, are the specific pairing of the optical fingertip with a whole-body event skin and the energy-accounting that would unify them.

8. Conclusion

Remote sensing ends at contact, and touch is the perception layer that begins there. The vision-based tactile fingertip makes the trade concrete: an image sensor watching a colored, marked gel is turned by computation into a geometry-and-force sensor, with photometric stereo recovering surface normals in one exposure, a Poisson solve integrating them to a depth map at micron scale, and a printed marker field reading shear and the onset of slip against the friction cone. None of these steps measures shape or force directly; all of them compute it from an image, on the device, fast enough to close a grip reflex within the hand. Placed beside a whole-body event skin and unified under a common energy budget, the fingertip is the high-resolution end of a single touch system, and the natural continuation of remote perception into the millimeter it cannot reach.

References

1. M. K. Johnson, E. H. Adelson. Retrographic sensing for the measurement of surface texture and shape. IEEE CVPR, 2009. doi:10.1109/CVPR.2009.5206534.
2. W. Yuan, S. Dong, E. H. Adelson. GelSight: high-resolution robot tactile sensors for estimating geometry and force. Sensors 17(12):2762, 2017. doi:10.3390/s17122762.
3. M. Lambeta, P-W. Chou, S. Tian, et al. DIGIT: a novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. arXiv:2005.14679, 2020.
4. M. Lambeta, T. Wu, A. Sengül, et al. Digitizing touch with an artificial multimodal fingertip (DIGIT 360). arXiv:2411.02479, 2024; code: github.com/facebookresearch/digit360.
5. R. J. Woodham. Photometric method for determining surface orientation from multiple images. Optical Engineering 19(1):139–144, 1980. doi:10.1117/12.7972479.
6. R. T. Frankot, R. Chellappa. A method for enforcing integrability in shape from shading algorithms. IEEE Trans. Pattern Analysis and Machine Intelligence 10(4):439–451, 1988. doi:10.1109/34.3909.
7. A. Bicchi, V. Kumar. Robotic grasping and contact: a review. IEEE Int. Conf. Robotics and Automation (ICRA), 2000. doi:10.1109/ROBOT.2000.844081.
8. A neuromorphic incipient slip detection system using papillae morphology. arXiv:2509.09546, 2025.
9. J. Zhao, Y. Ma, L. Wang, E. H. Adelson, et al. Transferable Tactile Transformers for representation learning across diverse sensors and tasks. arXiv:2406.13640, 2024.

AI-use disclosure. Preparation of this report used a large language model (Claude, Anthropic) for drafting and editing text, organizing the reviewed literature, and preparing the figures and the interactive companion. Cited references were checked to resolve to their sources. The author reviewed the content and is solely responsible for it. Consistent with ICMJE, COPE, and IEEE guidance, the model is a tool and is not credited as an author.

