

Research

Abe Davis

My research connects computer graphics, computer vision, and human-computer interaction with domain knowledge in different fields to explore new applications of visual computing. Cameras have become our most ubiquitous source of data about the physical world, providing a tremendous opportunity to develop new computational solutions to real-world problems. However, most problems in the real world do not map well to visual inference alone. Some are difficult to relate to visual structure; for example how do we recover something like sound from visual data—or a simulation of physical behavior? Others simply lack sufficient constraints in their general form; creative tasks often fall into this category, as ideal constraints vary with the preferences of different users. My work addresses these and other challenges by leveraging domain knowledge to adapt how we capture, model, and interact with visual data to the needs of new applications.

Below I describe two categories of such problems from my past work. The first explores how careful analysis of nearly imperceptible vibrations in video can help us reason about the physical properties and behavior of visible objects. The second focuses on how insights from different communities across art and science can help us build more powerful tools for manipulating visual data. The last part of this statement discusses future directions and describes specific problems that would make suitable PhD topics.

Project Videos: Most of my work deals with video, leaving many results difficult to convey through text alone. I have included hyperlinks to various project [videos](#)  throughout this document, which can also be found at abedavis.com, and on my [faculty application page](#). These videos are the best way to get a quick sense of my work and I strongly encourage the reader to check them out.

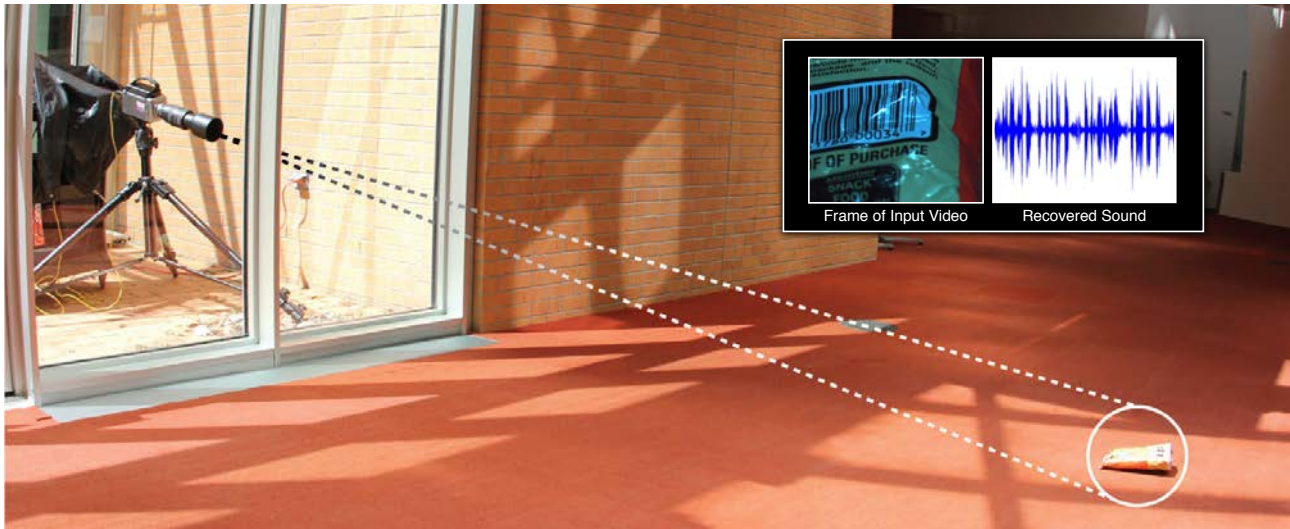
— Visual Vibration Analysis

Nothing is ever *completely* still. Real objects are always subject to some kind of force, and if one looks closely enough, *everything* is moving. Humans are amazingly adept at detecting some of this movement: our eyes pick up large motions, like the passing of a vehicle or the wave of a hand, while our ears alert us to the smaller, faster motion of sound. Our senses are limited though, and we are constantly surrounded by motion that eludes perception. Much of the movement we don't see is vibration—that is, motion that does not change the average shape or location of an object. Audible vibrations are an essential part of how we observe and communicate about the world, but most vibration is silent and what little we do hear is heavily integrated over our surroundings, limiting our ability to locate and reason about distinct sources of sound. The quiet, invisible vibrations of objects carry a tremendous amount of information, but most of that information is hidden from us. My dissertation [1]¹ showed that regular cameras can be used to capture and analyze such imperceptible vibrations visually, offering us a powerful new way to learn about visible objects and the forces that drive their motion.

— The Visual Microphone

When sound hits an object, it causes that object to vibrate, creating a subtle visual signal that is usually invisible to the naked eye. In my work [2] we showed how, using only silent video of an object and a suitable processing algorithm, we can extract these minute vibrations to partially recover the sounds that created them, allowing us to turn certain everyday objects—e.g., a glass of water, a potted plant, or a bag of chips—into visual microphones from a distance. This work was significant for several reasons: it demonstrated that high-speed video could reliably be used to recover audio; it mapped out how to predict

¹MIT Sprowls Award for best thesis in EECS | Runner-up, ACM SIGGRAPH Dissertation Award.

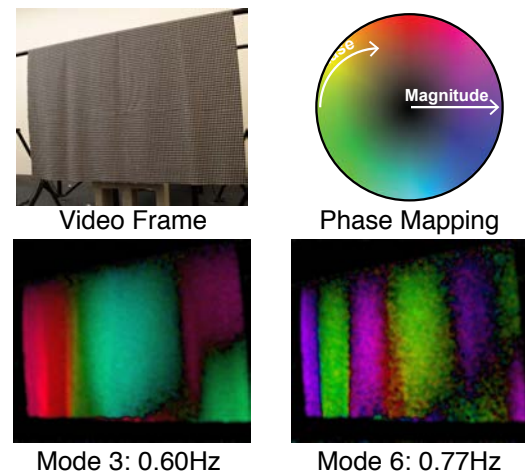


The Visual Microphone: This photo shows an experiment where a high-speed camera placed behind sound-proof glass (left) was used to recover human speech from silent video of a potato chip bag (circled, bottom right).

the SNR of sounds recovered in this way; and it introduced methods to recover frequency response curves and static vibration mode visualizations from video of an object, demonstrating new potential for computer vision in applications of vibration analysis.

— Visual Vibrometry [↗](#) [📺](#)

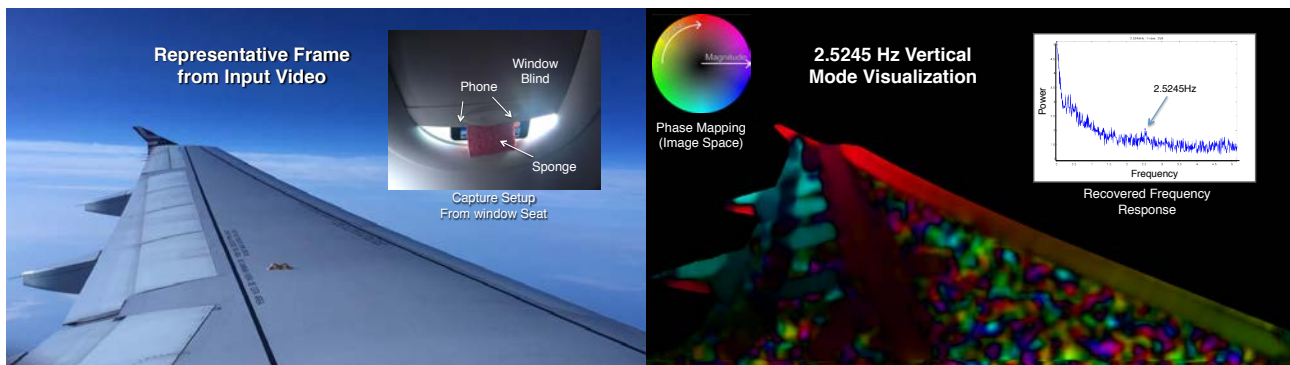
While my work on visual microphones used subtle visual signals to reason about forces that cause objects to vibrate (namely, sound), much of my subsequent work focused on using those same signals to learn about the objects themselves. Solid objects tend to vibrate in a set of preferred modes, the shapes and frequencies of which depend on structural and material properties. My work on visual vibrometry [3, 4] showed how information about an object's modes of vibration can be extracted from video and used to learn about various physical properties: e.g., density, Young's modulus, stiffness, and bulk damping. We also showed that the signals used to estimate these properties were highly invariant to changes in viewpoint, making them particularly robust cues for scene understanding. Finally, we showed that our approach could be used to detect physical changes in hidden or occluded parts of an object—information previously thought to be hidden from purely visual sensors.



Visual Vibrometry: This photo shows visualization for two modes of a hanging curtain captured with regular video. Using vibration spectra recovered from video, we were also able to infer bulk density and stiffness properties.

— Civil and Structural Engineering

Much of my work on visual vibrometry was motivated by a desire to connect computer vision with disciplines that use vibration analysis to analyze physical structures, especially civil and structural engineering. I have had the opportunity to explore several applications in these spaces through collaboration with researchers in Structural Health Monitoring (SHM) and Non-Destructive Testing (NDT)—areas that are rapidly becoming more important as we struggle to deal with aging infrastructure. SHM and NDT often rely on special-purpose contact or laser-based vibration sensors to detect damage in objects or structures. These sensors are expensive and difficult to deploy outside of laboratory settings, particularly when dense spatial measurements are needed. My work on visual vibration analysis shows that cameras can offer a low-cost, convenient way to trade temporal resolution for higher spatial sampling [5, 6, 7]. Most recently, I presented



In-Flight SHM with a Cell Phone: A flight from Boston to Palo Alto. During flight, a cell phone wedged under the window blind was used to record the left wing (setup shown left). Afterward, the video was analyzed to create a visualization of the wing's dominant flex mode (right), which was verified against published reference measurements for the same model aircraft [8]. This work, along with a similar example used to predict the response of a metal beam to localized forces, won the SHM in Action Award for Civil Engineering at the 2017 International Workshop on Structural Health Monitoring.

a paper [8] at IWSHM2017 (described in the figure above) on in-flight monitoring of aircraft wings using a mobile phone camera.² The ability to monitor large structures with inexpensive, ubiquitous hardware also opens up new opportunities for data-driven SHM.

— Manipulating Visual Data

As a representation, pixels tend to be far-removed from the objects and ideas they depict. This makes manipulating visual data a significant challenge for most humans. Several of my projects have looked at how to address this challenge in order to provide more powerful and intuitive tools for manipulating images and video.

— Dynamic Video

In the real world, we learn a lot about objects by through physical manipulation – by pushing, pulling, poking, prodding, and seeing what happens. Unfortunately, regular images and video don't allow for this type of interaction. My work on Interactive Dynamic Video (IDV) [9] extended visual vibration analysis to build physical models of how objects move in the image plane. Prior to this work, it was generally thought that geometry was needed to simulate objects in video. My insight was that the structural information needed for certain simulations could be derived directly from observed motion—specifically, by extracting projections of an object's vibration modes from its motion in video. In addition to allowing physically-plausible interaction with video this work can be used to



Interacting with Video: This photo shows a screenshot of a user interacting with an IDV of a jungle gym, which was extracted from less than a minute of regular video. The user can grab parts of the jungle gym and 'pull' them in different directions with their mouse. The physical response of the jungle gym is simulated in real time on a laptop.

create low-cost special effects for film; reason about how the interference of multiple modes affects physical stress in structures; and allow virtual characters to interact with real objects in in Augmented Reality (AR).

²Winner, 2017 SHM in Action Award for Civil Engineering.

— Computational Video Editing

In film, it is standard practice to record several versions, or takes, of a scripted scene. Each new take provides a unique camera framing or performance, which editors can cut together strategically to build a stronger narrative than any single recording could capture. But the space of possible ways to edit a scene grows quickly with each new take, and exploration of that space is slow and largely manual. Much of the difficulty comes from a mismatch of representation: editing decisions tend to be driven by high-level narrative goals, but must be executed through low-level manipulation of individual clips and frames. In [10] we built a system for editing video based on high-level objectives derived from cinematography literature. We first used vision and NLP to label and align input takes with a shared script. We then modeled high-level editing objectives described in cinematography literature as composable distributions over the labels of possible clip sequences. The result was a system that could automatically edit large amounts of video based on user-defined objectives. Compared to a professional editor our system reduced the amount of manual effort required to generate rough cuts of dialogue-driven scenes by orders of magnitude, letting editors spend much less time on tedium and more time focused on exploring the creative decisions that matter most.



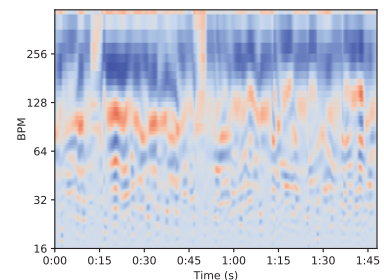
"Fluffles" : Input takes from one of our test scenes. Rows are different performances, and each column is a different framing of the scene. It took a professional editor three hours to manually produce a rough cut for this scene, while our system takes just three seconds to edit the scene once given a style objective. Further comparisons can be found on our [project website](#).

— Visual Rhythm and Beat

My work on Visual Rhythm and Beat [11] presented a visual analogue for musical rhythm derived from an analysis of motion in video, and showed that alignment of visual rhythm with its musical counterpart results in the appearance of dance. In other words, given a video with sufficient motion, we can manipulate that video to look like dance by time-warping its visual rhythm into alignment with a musical target. This work connects concepts in music information retrieval with computer vision and dance. We demonstrate a variety of applications, making video of everything from cats to politicians dance. You can find [code](#) and explore [hundreds of example results](#) linked from our [project website](#).



Source Video



Visual Tempogram

Accidental Dance: The visual tempogram on the right shows the rhythmic strength of President Obama's hand gestures at different tempos over the course of a short speech. In one experiment, we analyzed hours of footage like this from CSPAN to find moments where politicians were moving most dance-like ("accidental dance"). We then aligned these moments with a common song to create a dance music video.

Additional Past Work: Information about other projects is available on my [personal website](#). In particular, I also work on interactive computational camera systems, augmented reality, and image-based rendering [12, 13, 14, 15]. Of particular note is my work on [computational bounce flash](#), which Canon turned into the [Speedlite 470EX-AI AI Bounce](#) last year. I also developed [one of the first mobile AR apps](#) back in 2012, which let users capture, view, and exchange unstructured light fields.

— Future Work

Below I describe several possible future projects that build on my previous research experience. I am also excited to explore collaborations with other labs and departments, and eager to connect visual computing to more applications and domains.

Imaging Dynamics in the Field:

The practical impact of approaches like visual vibration analysis depends on how effectively they can be deployed in real-world settings. I want to design systems and protocols for using cameras and portable vibration devices to more quickly characterize the dynamics of different structures. A key challenge is how to ensure that all important deformation modes of a structure are recorded. Formulating this as a coverage problem, we should be able to guide a user's placement of vibration devices using augmented reality, similar to approaches I've used for light field capture [14] and laser-speckle imaging [12] in the past.

Imaging Non-linear Dynamics:

A linear model of dynamics is compact and simple to fit with data. However, it is only accurate for small deformations around an object's rest state (the domain of most vibration). This limits both what behavior we can simulate and what data we can fit. By extending IDV to use more complex, non-linear dynamics, it should be possible to recover much richer physical simulations from a greater variety of videos. There are two promising strategies for doing this. The first is to derive image space analogues for more complex non-linear models used in 3D simulations and fit the parameters of those models to motion in video. The second is to learn a more general model of motion directly from data using a deep network. The latter approach may be able to express more general types of motion, but would require large amounts of training data that could be hard to obtain. The simulation-driven approach also lets us choose physically meaningful parameters, making it easier to integrate with forward models. Both approaches are worth pursuing.

Sparsity-Driven Capture:

The high spatial resolution of cameras comes at the cost of significant temporal resolution when compared to point-based sensors. On some level this cost cannot be avoided; limited bandwidth necessitates some balance of spatial and temporal resolution, and the available light in a scene imposes such limits even with hypothetically perfect hardware. This makes imaging the higher-frequency vibration modes of an object very difficult. However, most objects have frequency responses that are either band-limited (as with high damping) or sparse (low damping). With control over high-speed strobed illumination and narrow frequency excitation forces, it should be possible design systems that leverage sparsity to capture an object's dynamics with regular-speed cameras. Such capture systems would be invaluable for non-destructive testing and could provide a controlled way to collect large amounts of training data for learning-based approaches.

Remixing Massive Amounts of Media:

A few of my recent projects have explored the benefits of using computational tools based on optimization to edit video [10, 11]. These tools balance automation with expressivity to let users work creatively with larger amounts of media—an exciting direction for research in creative domains. One open problem is how to extend this idea to use more content than any one person could possibly review (for example, all of YouTube). Facilitating creative exploration at this scale poses an interesting mix of systems, theory, and design challenges that I have begun to explore in my postdoc and hope to continue investigating as faculty.

User-Centered Design for Visual Computing:

A strong theme in my work has been connecting the progress of visual computing to other fields—from material science [3, 4] and civil engineering [8, 5, 7, 6], to dance and music theory [11], special effects [9], film editing [10], and several others. Part of my success in addressing this breadth of topics comes from incorporating user-centered design (need finding, interviews, prototyping, iterative design, etc.) into how I approach imaging problems. I would like to better formalize this connection through case studies working with labs in other research areas, and by creating a graduate level project course on "user-centered visual computing".

— References

- [1] Abe Davis. *Visual Vibration Analysis*. PhD thesis, Massachusetts Institute of Technology, Sep 2016. URL <http://abedavis.com/thesis.pdf>.
- [2] Abe Davis, Michael Rubinstein, Neal Wadhwa, Gautham J. Mysore, Frédo Durand, and William T. Freeman. The visual microphone: Passive recovery of sound from video. *ACM TOG. SIGGRAPH 2014*, July 2014. doi: 10.1145/2601097.2601119. URL <http://people.csail.mit.edu/mrub/VisualMic/>.
- [3] Abe Davis, Katherine L. Bouman, Justin G. Chen, Michael Rubinstein, Fredo Durand, and William T. Freeman. Visual vibrometry: Estimating material properties from small motion in video. June 2015. URL <http://www.visualvibrometry.com/>.
- [4] Abe Davis*, Katherine L. Bouman*, Justin G. Chen, Michael Rubinstein, Oral Buyukozturk, Fredo Durand, and William T. Freeman. Visual vibrometry: Estimating material properties from small motions in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, April 2017. doi: 10.1109/TPAMI.2016.2622271. URL <http://www.visualvibrometry.com/>.
- [5] Oral Buyukozturk, Justin G Chen, Neal Wadhwa, Abe Davis, Frédo Durand, and William T Freeman. Smaller than the eye can see: Vibration analysis with video cameras. *19th World Conference on Non-Destructive Testing 2016 (WCNDT)*, 2016. URL http://people.csail.mit.edu/ju21743/docs/buyukozturk_WCNDT_2016.pdf.
- [6] Justin G Chen, Neal Wadhwa, Abe Davis, Frédo Freeman Durand, T William, and Oral Buyukozturk. Long distance video camera measurements of structures. *10th International Workshop on Structural Health Monitoring (IWSHM 2015)*, 2015. URL http://people.csail.mit.edu/ju21743/docs/chen_IWSHM_2015.pdf.
- [7] Justin G Chen, Abe Davis, Neal Wadhwa, Frédo Durand, T Freeman, William, and Oral Buyukozturk. Video camera-based vibration measurement for condition assessment of civil infrastructure. *International Symposium Non-Destructive Testing in Civil Engineering (NDT-CE 2015)*, 2015. URL http://people.csail.mit.edu/ju21743/docs/chen_NDTCE_2015.pdf.
- [8] Abe Davis, Justin G. Chen, Oral Buyukozturk, Frédo Durand, and Doug L. James. Structural health monitoring from the window seat of a passenger airplane. *11th International Workshop on Structural Health Monitoring (IWSHM 2017)*, 2017. URL <http://abedavis.com/papers/AbeDavisIWSHM2017.pdf>.
- [9] Abe Davis, Justin G. Chen, and Frédo Durand. Image-space modal bases for plausible manipulation of objects in video. *ACM TOG. SIGGRAPH Asia 2015*, October 2015. doi: 10.1145/2816795.2818095. URL <http://www.interactivedynamicvideo.com/>.
- [10] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. Computational video editing for dialogue-driven scenes. *ACM TOG. SIGGRAPH 2017*, July 2017. doi: 10.1145/3072959.3073653. URL <https://graphics.stanford.edu/papers/roughcut/>.
- [11] Abe Davis and Maneesh Agrawala. Visual rhythm and beat. *ACM TOG. SIGGRAPH 2018*, July 2018. doi: 10.1145/3197517.3201371. URL <http://abedavis.com/visualbeat/>.
- [12] Y. C. Shih, A. Davis, S. W. Hasinoff, F. Durand, and W. T. Freeman. Laser speckle photography for surface tampering detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012. doi: 10.1109/CVPR.2012.6247655. URL <http://groups.csail.mit.edu/graphics/speckle/>.
- [13] Lukas Murmann, Abe Davis, Jan Kautz, and Frédo Durand. Computational bounce flash for indoor portraits. *ACM TOG. SIGGRAPH Asia 2016*, November 2016. doi: 10.1145/2980179.2980219. URL <http://lmurmann.net/bounceflash>.
- [14] Abe Davis, Marc Levoy, and Fredo Durand. Unstructured light fields. *Eurographics 2012, IEEE Computer Graphics Forum*, May 2012. doi: 10.1111/j.1467-8659.2012.03009.x. URL <http://people.csail.mit.edu/abedavis/ULF/>.
- [15] Lixin Shi, Haitham Hassanieh, Abe Davis, Dina Katabi, and Fredo Durand. Light field reconstruction using sparsity in the continuous fourier domain. *ACM TOG. SIGGRAPH 2015*, December 2014. doi: 10.1145/2682631. URL <http://groups.csail.mit.edu/netmit/LFSparseRecon/>.