Visual Vibration Analysis

Abe Davis

August 31, 2016

Visual Vibration Analysis

by

Abe Davis

Submitted to the Department of Electrical Engineering and Computer Science
on September, 2016, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science and Electrical Engineering

Abstract

This dissertation shows how regular cameras can be used to record and analyze the vibrations of visible objects. Through careful temporal analysis, we relate subtle changes in video to the vibrations of recorded surfaces, and use that information to reason about the physical properties of objects and the forces that drive their motion.

We explore several applications of our approach to extracting vibrations from video – using it to recover sound from distant surfaces, estimate the physical properties of visible objects, and even predict how objects will respond to new, previously unseen forces. Our work impacts a variety of fields, ranging from computer vision, to long-distance structural health monitoring and nondestructive testing, surveillance, and even visual effects for film.

By imaging the vibrations of objects, we offer cameras as low-cost vibration sensors with dramatically higher spatial resolution than the devices traditionally used in engineering. In doing so, we turn every camera into a powerful tool for vibration analysis, and provide an exciting new way to image the world.

Thesis Supervisor: Fredo Durand
Title: Professor

Acknowledgments

I would like to thank my advisor, Professor Frédo Durand for his support and patience over the past six years; Professor William T. Freeman for welcoming me into the research that he and his lab were doing on small motion in video, and for his advice and collaboration over the years; I'd like to give special thanks to Miki Rubinstein, who was a wonderful mentor to me when I began this work, and has continued to be a great collaborator and source of advice over the years, as well as Neal Wadhwa and Justin G. Chen, both for their collaboration on this work and on exploring how it could be applied outside of our research; I'd also like to thank Oral Buyukozturk, Gautham Mysore, and Katie Bouman for their collaboration on this work; I also had the privilege of working with great collaborators on projects that arent in this thesis during my time in grad school, who I also want to thank: Dina Katabi, Haitham Hassanieh, Lixin Shi, Yichang Shih, Samuel W. Hasinoff, Marc Levoy, Jan Kautz, and Lukas Murmann. Thanks again to Fredo, Bill, and Dina, as well as Wojciech Matusik for serving on my thesis committee. I have to give special thanks to Andrew Owens for being an awesome room mate for 6 years, and Bryt Bradley for being so much more than just our lab's admin. I was also fortunate to have great office mates while I was here: Jiawen Chen, Rob Wang, Michael Gharbi, Tiam Jaroensri, Adriana Schulz, Emily Whiting, Forrester Cole. I'd like to thank all of the other labmates and fellow students who made my time in grad school so much more enjoyable. Finally, I'd like to thank the rest of my family and friends who supported my bizarre obsession with research. There are too many of you to list here, but you are too wonderful not to mention. Thank you.

Contents

List of Figures

List of Tables

1

Introduction

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 our 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 individual objects carry a tremendous amount of information – but most of that information is hidden from us.

This dissertation shows how regular cameras can be used to capture and analyze such imperceptible vibrations *visually*. Our strategy is to relate subtle variations in video to the vibrations of recorded surfaces. By applying careful temporal analysis and established theories on vibration to signals we extract from video, we can infer a great deal about visible objects and the forces that drive their motion.

1.1 Overview

This dissertation is about the capture and analysis of vibrations using cameras. In it, we explore a surprising range of applications – from the recovery of sound, to physical property estimation, structural health monitoring, and even low-cost special effects for film. We originally considered these applications in separate publications, presented to different communities in computer vision, graphics, and civil engineering. My goal for this text is to distill the concepts that unite our work, and present each application as a different use case of common techniques and theory.

Vibration is a very fundamental topic, and cameras are among the most ubiquitous technologies of our time. I hope and suspect that many applications of this work are yet to be discovered. For this reason, I have tried to make this text accessible to a variety of potential readers. Figure 1-1 to the right shows a roadmap of how this dissertation is structured. Chapters 2 provides background and context for the rest of the dissertation. Chapter 3 presents the common theory and algorithms that underlie our work. Finally, Chapters 4, 5, and 6 present different applications, focusing on experimentation and analysis.

The rest of this chapter contains more detailed descriptions of each chapter, as well as links to videos, data, code, and the original publications this dissertation is based on.
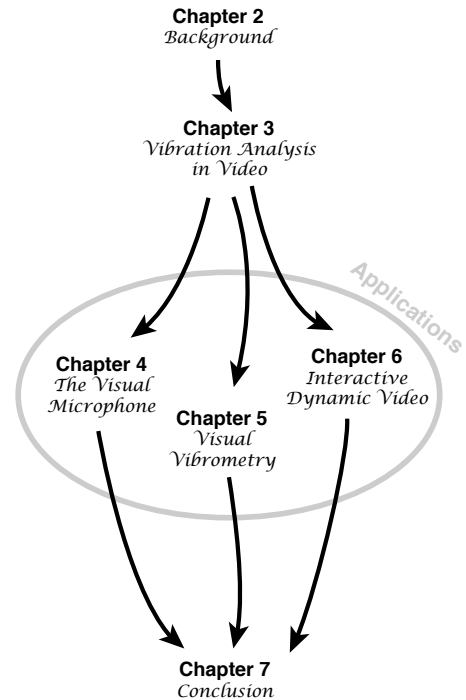
Figure 1-1: Dissertation Roadmap

Introduction (Ch 1) Chapter 1 includes an overview of the text, as as well as links to videos and data from our original publications. Readers are encouraged to watch the linked videos as an introduction to the work, and to see and hear results.

Background (Ch 2) Chapter 2 provides basic background on motion and vibration, how they are traditionally captured, and how they are used. We consider the strengths and weaknesses of different sensing devices, starting with human perception, then covering more traditional vibration sensors, and finally cameras. This chapter also serves to clarify our discussion of ideas that may be addressed differently in different related fields of study (e.g. computer vision and civil engineering), and review related work on small motion in video..

Vibration Analysis in Video (Ch 3) Chapter 3 outlines the theory behind our work, and derives our approach to vibration analysis in video. This chapter distills much of the common theory and algorithms that evolved over different publications in our original work. Much of our discussion here goes into greater depth and gives more examples than what appeared in our original publications, using consistent notation for derivations that we originally presented in different contexts. We also discus using modal imaging to analyze the motion of objects outside of laboratory settings – or, "modal imaging in the wild".

The Visual Microphone (Ch 4) When sound hits an object, it causes small vibrations of the object's surface. We show how, using only high-speed video of the object, we can extract those minute vibrations and partially recover the sound that produced them, allowing us to turn everyday objects — e.g. a glass of water, potted plant, box of tissues, bag of chips — into visual microphones. We recover sounds from high-speed footage of a variety of objects with different properties, and use both real and simulated data to examine some of the factors that affect our ability to visually recover sound. We evaluate the quality of recovered sounds using intelligibility and SNR metrics and provide input and recovered audio samples for direct comparison. We also explore how to leverage the rolling shutter in regular consumer cameras to recover audio from standard frame-rate videos.

Visual Vibrometry (Ch 5) Objects tend to vibrate in a set of preferred modes. The shapes and frequencies of these modes depend on the structure and material properties of an object. We show how information about an object's modes of vibration can be extracted from video and used to make inferences about that object's physical properties. We demonstrate our approach by using high-speed and regular framerate video to estimate physical properties for a variety of objects.

Interactive Dynamic Video (Ch 6) In the real world, we learn a lot about objects by interacting with them. Unfortunately, regular images and video don't allow for this kind of interaction. We show how, by recovering the shapes and frequencies of an object's vibration from video, we can build plausible image-space models of their dynamics. The result is an interactive video-based simulation of objects that can interactively predict how they will respond to new, unseen forces.

7

Conclusion (Ch 7) In Chapter 7 we review our contributions, and the applications that they impact. Our work bridges computer vision and a rich history of applications and theory dealing with vibration. By turning cameras into low-cost, high-spatial resolution vibration sensors, we offer a powerful and ubiquitous new tool for imaging the world.

1.2 Videos and Online Content

Chapters 4, 5, and 6 describe work that was originally published in the ACM Transactions of Graphics (SIGGRAPH and SIGGRAPH Asia), and the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Each chapter has an accompanying video, with an overview of the work and some results, as well as a project website.

The Visual Microphone [24]

Project Page:
people.csail.mit.edu/mrub/VisualMic
Video:
youtube.com/watch?v=FKXOucXB4a8
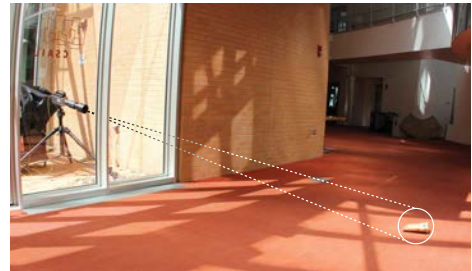
Visual Vibrometry [22]

Project Page:
VisualVibrometry.com
Video:
youtube.com/watch?v=5apFqYEx5ew
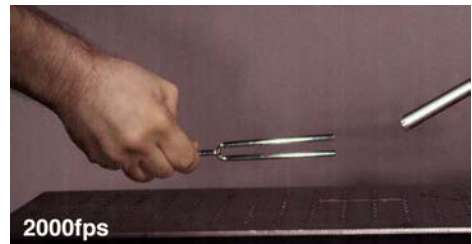
Interactive Dynamic Video [23]

Project Page:
InteractiveDynamicVideo.com
Video:
InteractiveDynamicVideo.com

The linked videos are meant to give a high-level introduction to the work, and are best viewed before reading each corresponding chapter.

Note: I also coauthored related papers [10, 13, 17] with researchers in civil engineering. Those papers are not discussed in the dissertation, but details and links to the paper can be found from on website: abedavis.com

2
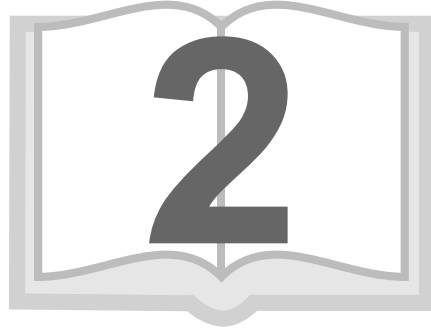
Background

We begin this chapter by building intuition and vocabulary for our discussion of motion. We then review some of the ways that vibration is captured, starting with human perception, then covering traditional vibration sensors and, finally, cameras.

2.1 Vibration as Motion

Vibrations are a type of motion – and yet, our intuition for vibration is very different from our intuition for other types of motion. Much of this difference has to do with the limits of how motion is *resolved* and *perceived* by different sensors.

$$f(t) = \vec{\mathbf{x}}$$

2.1.1 Resolving Motion

Motion is defined by a change in position over time. Our ability to capture motion is therefore limited by our ability to resolve positions in both space and time.

Spatial Resolution describes how well we can resolve shapes and locations in space. Discussions of spatial resolution are often complicated by the fact that a sensing device may sample at different resolutions in different dimensions of space. Our discussion of spatial resolution will mainly focus on the *number* of points being measured in space. This meaning is common in discussions of cameras, where the distance between points imaged by adjacent pixels varies depending on the magnification of a lens and the distance between camera and subject.

Temporal Resolution   describes our ability to resolve events in time, and is typically determined by a *sampling rate*, usually measured in Hertz (Hz). We will see throughout this dissertation that vibrations at different frequencies can carry different information. Sensors with higher sampling rates are able to capture higher frequencies of vibration, and therefore more information from the objects they record.

2.1.2   Perceiving Motion

The intuition that separates vibration from other types of motion is grounded in our own perception. Studies have shown that the lowest frequency humans can hear is about 20Hz, which sits at the upper limit of frequencies that we can see [21, 37].[1] This means that, even when we see and hear the same *object*, we rarely see and hear the same *motion*. The result is a surprising discontinuity in how we perceive the world: our intuition for low-frequency motion is highly spatial, while our intuition for high-frequency motion (which we associate with vibration) is largely built on temporal features, like tone and timbre. Both senses offer a useful perspective of motion, but each lacks the insight of the other. We don't hear shapes, or see tones – even though these concepts apply to all motion, regardless of how we perceive it.

At a high level, our work can be seen as applying the kind of *temporal* reasoning that most of us associate with sound to *visual* data. Chapters 4, 5, and 6 focus on demonstrating the value of such reasoning in the context of specific applications; but intuition is rooted in perception, and another useful way to look at our work is as a means of addressing certain limits of our own senses:

"Shape-Deafness"   Our ears favor temporal resolution over spatial resolution,[2] making it easy for us to hear high-pitched noises, but difficult to isolate, or locate distinct sources of sound. The vibrations we hear are averaged over many directions, leaving us deaf to their shapes. This is what makes it difficult to follow a specific conversations in a room full of talking people.

"Tone-Blindness"   As sensors, our eyes favor spatial resolution over temporal resolution. This is why we can see shapes in fine detail, while fast motions appear blurry. But, even at low speeds, we often struggle to recognize temporal frequencies in visible motion. This is why children use skipping songs when playing jump rope, or why CPR students are taught to perform chest compressions in time with the song "Stayin Alive"[3]. Our poor ability to recognize temporal frequencies in visible motion limits our ability to understand and predict the dynamic behavior of objects – which we will do computationally in Chapter 6.

---

[1]For reference, feature films are recorded at just 24Hz.

[2]The temporal resolution of human hearing is more complicated than a simple sampling rate, as different parts of the ear actually sense different frequencies of sound. This results in poor phase perception in certain contexts. Still the design of the human ear is one that strongly favors temporal resolution when compared to human eyes.

[3]The Bee Gees classic is currently recommended by cpr.heart.org. I first heard of this practice in 2009, during my own CPR certification.

2.2   Related Fields

Our work draws on and contributes to a variety of related fields. We discuss these related fields in more detail as we encounter relevant material in each chapter, but introduce a few key areas here.

Recording Audio:

As sound is such an important part of how we experience the world, the ability to record and reproduce sound as we hear it is very valuable. In Chapter 4 we show how to do this by analyzing the vibrations of visible objects in video. This is especially relevant to remote sound acquisition, which is an important part of surveillance.

Vibration Analysis in Engineering:

Several engineering disciplines rely on modal analysis of vibrations to learn about the physical properties of structures. Particularly relevant areas include structural health monitoring (SHM) and non-destructive testing (NDT).

Modal Analysis in Simulation:

Many techniques for physical simulation use modal analysis to define a modal basis that reduces the degrees of freedom in simulation, making computation more efficient. In Chapter 6 we use a similar concept to recover an image-space modal basis for simulating objects in video.

2.3   Traditional Vibration Sensors

Many devices have been used to measure and analyze vibrations. Here we review some of the most popular and effective devices.

Microphones

Traditional microphones are an example of *passive* sensors – meaning that they operate without projecting anything onto the objects they measure. They work by converting the motion of an internal diaphragm into an electrical signal. The diaphragm is designed to move readily with sound pressure, so that its motion can be recorded and interpreted as audio.

Traditional microphones often sample at very high rates, allowing them to recover frequencies of sound well above the limits of human hearing. However, microphones average aggressively over space, making it difficult to isolate and localize individual sources of sound.

Active Sensors

Active sensors take measurements by projecting something onto the object being measured, and observing a response or reflection. Laser vibrometers (also called laser microphones) are the most

common type of active sensor used for measuring vibration. Laser vibrometers measure the vibrations of an object by recording the reflection of a laser pointed at its surface. The most basic type records the phase of the reflected laser, which gives the objects distance modulo the lasers wavelength. A laser Doppler vibrometer (LDV) resolves the ambiguity of phase wrapping by measuring the Doppler shift of the reflected laser to determine the velocity of the reflecting surface [60]. Both types of laser-based sensors can recover high frequency vibrations from a great distance, but depend on precise positioning of a laser and receiver relative to a surface with appropriate reflectance.

Contact Sensors

Contact sensors work by placing a sensor directly on the object being measured. Accelerometers and piezoelectric pickups are examples of contact sensors that respond to acceleration. The main disadvantages of contact sensors come from the requirement that they be attached to the surface being measured. In many contexts, instrumenting an object with sensors is inconvenient or impossible. Furthermore, the weight of attached sensors may influence measurements by changing the way an object vibrates.

2.4 Cameras

Cameras are one of the most ubiquitous technologies of our time. Normally, we use them to capture the world as we see it, but cameras are not bound by the same limits as our vision. More an more, we use them to image phenomena outside the range of human perception: high-speed cameras capture video at frequencies too high for us to see, time-lapse photography reveals movement otherwise too slow to notice, and recent work in computer vision has shown that algorithms can be used magnify motion that is normally too small or subtle for the human eye to detect [79, 75, 77, 61]. Our work builds on this theme of using cameras to capture motion that would be otherwise invisible. However, where previous work has focused on visualizing such motion, we focus on quantitative analysis, bridging related works in computer vision with a rich history of vibration analysis in other fields.

2.5 Motion Magnification:

Our work builds on several recent works in vision and graphics that address small motions in video [79, 75, 77, 61]. As with many of these works, we use an Eulerian approach to motion estimation based on spatial phase variations of the complex steerable pyramid [67, 57]. These works also consider temporal properties of small motion, magnify certain frequencies to visualize phenomena that are otherwise too small to see. However, where these prior works focus on magnifying and visualizing small motion in video, we focus on analysis. In work that was partially concurrent with our own, Chen et al. [15, 16] used similar small motions to quantify the vibration modes of pipes and cantilever beams. By contrast, we focus on building general methods that can be applied to a greater variety of scenarios, often outside of controlled settings.

3

Vibrations in Video

This chapter describes the common theory and algorithms that underlie our work. Our goal is to take ideas that originally evolved over the course of several publications, and present them together to better show how they are connected.

Overview of Chapter 3:

Section 3.1: Local Motion Signals   We describe how to compute local motion signals and local contrast. These local signals are used throughout the rest of the dissertation.

Section 3.2: Global Motion Signals   We describe how to average local signals into global motion signals, which we use in Chapter 4 to recover sound from video.

Section 3.3: Theory of Vibration   We review established theory on modal analysis, which motivates our approach to computing global power spectra and modal images. This theory is also the basis for most of our analysis in Chapters 5 and 6.

Section 3.4: Global Power Spectra   We derive global power spectra as a way to represent the resonance of objects in video. These spectra are used in Chapter 5 to estimate physical properties of vibrating objects, and in Chapter 6 to find vibration modes of objects, which we use for simulation.

Section 3.5: Modal Images   We describe how to extract modal images, which visualize projections of vibration modes, from video. These images are used for visualization throughout the dissertation, and in Chapter 6 we use them to simulate the motion of objects.

3.1 Local Motion Signals

Here we describe how to compute local motion signals and local image contrast using the complex steerable pyramid (CSP) [67, 57, 31]. We use the CSP for convenience, and because previous work showed that it is effective when measuring small motion in video [75]. However, most of this dissertation does not depend on any specifics of the CSP – we assume only that motion and image contrast can be measured locally and at different orientations and scales. We therefore limit our discussion of the CSP to this section of the dissertation, referring more generally to local motion and contrast throughout the rest of the text.

> **Note About Phase:** This section (Section 3.1) describes how we measure motion using spatial phase variations in the CSP. In the past, the term 'phase' has been a common source of confusion in our work, as it may refer to spatial phases in the CSP, or temporal phases of motion over time. To avoid this confusion we will only use the term 'phase' to refer to spatial phases in this section and in Section 3.7 of the dissertation. For the rest of the text, all references to phase will refer to *temporal* phase information.

3.1.1 Measuring Local Motion with the CSP

We derive local motion signals from spatial phase variations in a CSP representation of the video V. The basis functions of the CSP are scaled and oriented Gabor-like wavelets with both cosine and sine phase components. Intuitively, these wavelets behave like a local Fourier transform: the texture around each point in an image is transformed into amplitudes and phases at different scales and orientations. Local amplitude roughly measures local contrast in an image, and translations of local texture result in shifts to local phase.

We compute a CSP for each input frame $V(x, y, t)$ of our video, giving us a complex image for each scale r and orientation θ, that can be expressed in terms of local amplitudes A and phases ϕ:

$$A(r, \theta, x, y, t)e^{i\phi(r,\theta,x,y,t)}. \tag{3.1}$$

Changes in local phase ϕ over time are linearly correlated with translations of local texture. To measure motion, we take the local phases ϕ in each frame and subtract the local phase of a reference frame t_0 (typically the first frame of the video), giving us the phase variations

$$u(r, \theta, x, y, t) = \phi(r, \theta, x, y, t) - \phi(r, \theta, x, y, t_0). \tag{3.2}$$

We refer to value $u(r, \theta, x, y, t)$ as a **local displacement**, and a signal $u_\ell(t)$ of local displacements over time (here ℓ indexes a particular location, orientation, and scale) as a **local motion signal**.

3.1.2 Noise and Local Contrast

In regions of high image contrast, the local motion signals given by Equation 3.2 are a good measure of motion in video. However, in regions of low image contrast, local phase information in the CSP is often dominated by noise. This quality is not specific to motion measured with the CSP – rather, it is a fundamental consequence of the classic aperture problem in computer vision. Fortunately, local amplitudes $A(r, \theta, x, y, t)$ in the CSP provide a convenient measure of image contrast that can be used as a confidence value to predict noise in motion signals. Points with low amplitude have high noise, and therefore low confidence – while points with high amplitude have low noise and therefore high confidence (as illustrated in Figure 3-1). We use these confidence values to weight and filter motion signals in different applications, which we

Figure 3-1: Gaussian noise around two points in the complex plane. Points with lower amplitude (P1) have higher variance in phase than points with higher amplitude (P2), which can be seen here in the distribution of angles to the origin covered by each gaussian.

describe later. To avoid confusing amplitudes in the CSP with other amplitudes in later chapters, we will hereafter refer to spatial amplitudes of the CSP as **local contrast**.

In most of our work, the local motion and contrast values that we have defined here (in terms of phases and amplitudes of the CSP) could likely be replaced with alternative measures of motion and contrast. Some investigation of advantages and disadvantages offered by alternative measures was made in [74], though the topic remains an ongoing area of research.

3.2 Global Motion Signals

Most traditional devices for measuring vibration record only a single point, and often in only a single dimension (e.g. laser vibrometers). One benefit of using video is that it gives us a way to measure motion across multiple points and dimensions in space. However, the resulting volume of data can be noisy, and unnecessarily cumbersome for many applications. The solution is to average local motions into a single global motion signal. Different strategies for averaging boost and attenuate different information, though – making the 'correct' strategy different for different applications.

The recovery of sound from video, which we describe in Chapter 4, is perhaps the simplest example of an application that calls for averaging, and it is our motivation for the global motion signals we describe in this section.

3.2.1 Sound and Coherent Motion

Intuitively, if our goal is to recover sound happening around a visible object, then our strategy for averaging local signals should cause motion correlated with that sound to add constructively without causing noise to add constructively.

Initially, we experimented with averaging strategies that attempted to force anticorrelated motion to add constructively. The logic was that resonant vibrations of an object (which we explore in subsequent sections) should provide useful information for recovering sound. However, we found that such motion tends to tell us more about objects and less about the forces – like sound – that drive their motion. Consider what happens when you tap on glassware with an eating utensil, as one might do when making a toast. The force driving vibration of the glassware is a short broad-spectrum impulse (the tap). The glass then resonates at a small number of frequencies to create a ringing sound which continues for some time after the glass has been struck. Our task of recovering sound from video is analogous to recovering the initial 'tap' from the ringing glass. The resonant vibrations that cause our glass to ring are large (and thus quite audible), but they are quite different from the force that initiated them.

Better visual microphones are objects that move *with* sound around them. And we found that, for such objects, motion *with* sound was typically coherent across the spatial dimensions of video.[1] Figure 3.2.1 shows one likely explanation for this. Consider the range of sound frequencies carried by a standard telephone (approx. 300Hz-3.4kHz). In room temperature air, the wavelengths of these frequencies range from about 10 centimeters to 1.13 meters. Most of this range is much larger than the portions of objects we recorded in our experiments, suggesting that motion of objects moving *with* sound should be relatively coherent across our image. Put another way, if we were able to see the sound itself, the motion of that sound would be coherent in our image, therefore motion that is closely correlated with that sound should be coherent as well.

Coherence Across Orientations: While the motion we are looking for is generally coherent across different *points* in an image, it may not be across different *orientations*. Consider a single point vibrating along the line defined by $y = -x$. If we simply average the motion of this point in x and y, we will always end up with a constant. For random choices of orientation, such a scenario may be rare, but it is a simple corner case to address. Our solution is to align motion signals that correspond to different orientations before averaging. This strategy leads to a surprisingly simple algorithm for computing global motion signals, which we describe in detail below.

3.2.2 Averaging Local Signals

We begin by calculating weighted local motion signals, where the square of local contrast is used as a measure of confidence:

$$\mathfrak{u}(r, \theta, x, y, t) = A(r, \theta, x, y, t)^2 u(r, \theta, x, y, t) \tag{3.3}$$

For each orientation θ and scale r, we then compute a sum of the weighted local motion signals to produce a single intermediate motion signal $a(r, \theta, t)$:

$$a(r, \theta, t) = \sum_{x,y} \mathfrak{u}(r, \theta, x, y, t). \tag{3.4}$$

---

[1] We observed this in our work described in Chapter 4, and believe it holds for most scenarios that are practical with today's cameras. However, our hypothesis for what causes this does not apply to significantly different scales or frequencies.
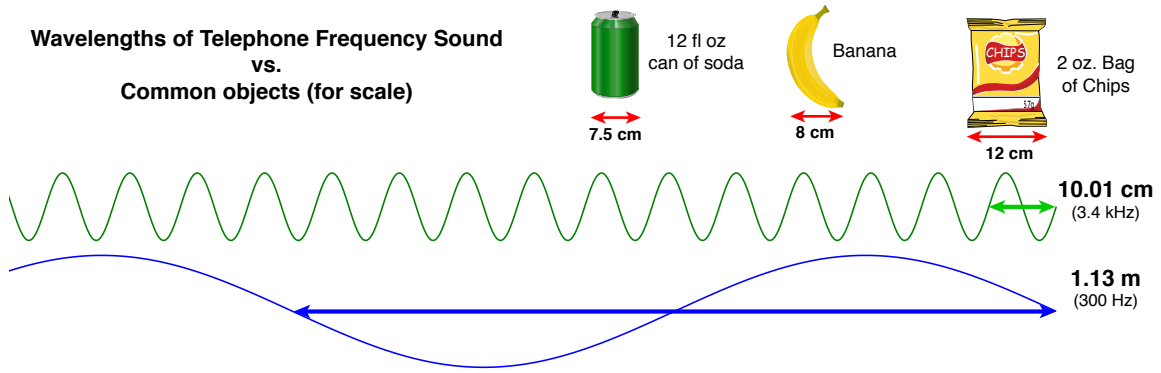
Figure 3-2: Here we see the wavelengths of sound at the high end of telephone frequencies (3.4kHz, 10cm) and the low end of telephone frequencies (300Hz 1.13m) next to several objects for scale. The three objects were taken from my afternoon snack, purchased at a cafe on MIT's campus, and may not be standard sizes.

Before averaging these $a(r, \theta, t)$ over different orientations and scales, we align them temporally to prevent destructive interference. The aligned signals are given by $a_i(t - t_\ell)$, such that

$$t_\ell = \operatorname*{argmax}_{t_\ell} a_0(t)^T a_\ell(t - t_\ell), \tag{3.5}$$

where ℓ in a_ℓ indexes all scale-orientation pairs (r, θ), and $a_0(t)$ is an arbitrary choice of reference scale and orientation. Our global motion signal is then:

$$\hat{s}(t) = \sum_\ell a_\ell(t - t_\ell), \tag{3.6}$$

Weights and Scale Ambiguity: Note that the global motion signals $\hat{s}(t)$ have ambiguous scale. If we normalize the intermediate signals a_ℓ by the sum of weights used in their calculation, then we can get values roughly proportional to pixel displacements at each corresponding scale. However, balancing normalization with a strategy that weighs against noise at different scales can be difficult, and the relationship between motion in pixels and metric motion is generally unknown. For simplicity, we chose to allow for scale ambiguity and weigh all signals according to local contrast, arriving at the weights A^2 empirically. We note, however, that other functions of local contrast seemed to perform almost as well, and the optimal weighting strategy for computing $\mathfrak{u}(r, \theta, x, y, t)$ remains an open problem.

> **Looking Ahead at Chapter 4:** In Chapter 4, we explore the use of global motion signals to recover sound from the vibrations of distant objects. Most of that chapter focuses on controlled experiments, and strategies for filtering global signals to recover audible sound.

3.3 Theory of Vibration

Our global motion signals $\hat{s}(t)$ are designed to target motion that is correlated with the sound around a visible object. In Chapters 5 and 6, we focus instead on using vibrations in video to learn

about objects themselves. For this we leverage established theory on vibration. This section reviews relevant theory on modal analysis, focusing on parts that relate directly to our work. For more detailed derivations we refer to the book [64].

3.3.1 Vibrations of Objects

Objects tend to vibrate in a set of preferred modes. Bells, for instance, vibrate at distinct audible frequencies when struck. We cannot usually see these vibrations because their amplitudes are too small and their frequencies are too high – but we hear them. Intuitively, we know that large bells tend to sound deeper than small ones, and that a bell made of wood will sound muted compared to one made of silver. This intuition is built on the close relationship between the physical properties of objects, and the way those objects vibrate.

While small motion like vibration is often difficult to see, it can be surprisingly simple to analyze. The general motion of an object may be governed by complex nonlinear relationships, but small deformations around a rest state (like vibration) are often well-approximated by linear systems. The theory of such linear systems is well established, and used in work spanning a variety of disciplines. This section reviews basic modal analysis, which is especially relevant to our work. In Sections 3.4 and 3.5 we use this theory to derive global power spectra and modal images, which we use to summarize the shapes and frequencies of vibration in video.

3.3.2 Modal Analysis

In modal analysis, a solid object with homogeneous material properties is modeled as a system of point masses connected by springs and dampers. Intuitively, rigid objects are approximated with stiff springs, highly damped objects approximated with strong dampers, and dense objects are approximated with heavy masses. Consider the mass matrix $\mathbf{M}$ of inertias between points, $\mathbf{C}$ of viscous damping values between points, and the matrix $\mathbf{K}$ of spring stiffnesses. The differential equation of motion for this system is given by:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{C}\dot{\mathbf{x}} + \mathbf{K}\mathbf{x} = 0, \tag{3.7}$$

where $\mathbf{x}$, $\dot{\mathbf{x}}$, and $\ddot{\mathbf{x}}$ are vectors describing the displacement, velocity, and acceleration of our points, respectively. Under the common assumption of Rayleigh damping, the matrix $\mathbf{C}$ is a linear combination of $\mathbf{M}$ and $\mathbf{K}$ given by $\mathbf{C} = \alpha\mathbf{M} + \beta\mathbf{K}$. In this case, the eigenmodes of the system are the solutions to the generalized eigenvalue problem given by $\mathbf{K}\phi_i = \omega_i^2\mathbf{M}\phi_i$. The eigenmodes $\phi_1...\phi_N$ define a modal matrix $\mathbf{\Phi}$ that diagonalizes the mass and stiffness matrices into modal masses $\mathbf{m}_i$, and stiffnesses $\mathbf{k}_i$:

$$\mathbf{\Phi} = \begin{bmatrix} \phi_1 & \phi_2 & \dots & \phi_N \end{bmatrix} \tag{3.8}$$

$$\mathbf{\Phi}^T\mathbf{M}\mathbf{\Phi} = diag(\mathbf{m}_i) \tag{3.9}$$

$$\mathbf{\Phi}^T\mathbf{K}\mathbf{\Phi} = diag(\mathbf{k}_i) \tag{3.10}$$

The matrix $\mathbf{\Phi}$ defines modal coordinates $\mathbf{q}(t)$, where $\mathbf{x}(t) = \mathbf{\Phi}\mathbf{q}(t)$, which decouple the system into single degree of freedom systems defined by modal masses $\mathbf{m}_i$, stiffnesses $\mathbf{k}_i$, and dampings

$\mathbf{c}_i = \alpha\mathbf{m}_i + \beta\mathbf{k}_i$. Defining the undamped natural frequency of a mode as $\omega_i = \sqrt{\frac{\mathbf{k}_i}{\mathbf{m}_i}}$, we get the decoupled equation of motion for each mode in terms of its corresponding modal coordinate, $q_i(t)$:

$$\ddot{q}_i(t) + 2\xi_i\omega_i\dot{q}_i(t) + \omega_i^2 q_i(t) = 0 \tag{3.11}$$

where ξ_i is a modal damping ratio, defined as:

$$\xi_i = \frac{\mathbf{c}_i}{2\mathbf{m}_i\omega_i} = \frac{1}{2}\left(\frac{\alpha}{\omega_i} + \beta\omega_i\right). \tag{3.12}$$

Important Takeaways from Equations 3.7-3.12:

Mode Shapes $\phi_1...\phi_N$: Each mode shape ϕ_i represents a different way the object can vibrate, and the set of mode shapes $\phi_1...\phi_N$ form an orthogonal basis where object motion can be decoupled into independent 1DOF systems.

Mode Frequencies ω_i: Each mode is associated with a particular frequency ω_i, and these frequencies are global properties of the object (meaning all motion associated with a particular mode happens at the same frequency).

Geometry: Both mode shapes and frequencies depend on an object's geometry. For example, if a piece of an object is removed, the sparsity of $\mathbf{M}$ and $\mathbf{K}$ changes, potentially changing both the eigenvectors and eigenvalues for our system.

Material Properties: If geometry is held constant and only material properties are changed (say by making the object uniformly denser or stiffer), this simply scales the eigenvalues of our system, leaving eigenvectors unchanged. This implies two things: 1) different objects with the same geometry have the same set of mode shapes, and 2) resonant frequencies scale in proportion to material properties, leaving the ratios of mode frequencies unchanged.

3.4 Global Power Spectra

Most of our strategy for learning about objects in video will focus on using the motion signals $u_\ell(t)$ to reason about the vibration modes of visible objects. This becomes much easier if we consider motion in the frequency domain. In this section we derive the transfer functions of modal systems, and relate those transfer functions to the spectra of our motion signals $u_\ell(t)$.

3.4.1 Transfer Functions of Modal Systems

Just as we did with motion in Section 3.3.2, we can use modal coordinates to decouple the impulse response of our system into a superposition of simpler impulse responses for individual modes. We obtain the unit impulse response for the ith mode by solving Equation 3.11

$$h_i(t) = \left(\frac{e^{-\xi_i \omega_i t}}{\mathbf{m}_i \omega_{di}} \right) \sin(\omega_{di} t) \tag{3.13}$$

where the damped natural frequency is $\omega_{di} = \omega_i \sqrt{1 - \xi_i^2}$. The Fourier transform of the unit impulse response $h_i(t)$ in Equation 3.13 is then the convolution

$$H_i(\omega) = \left(\frac{1}{\mathbf{m}_i \omega_{di}} \frac{\xi_i \omega_i}{\xi_i^2 \omega_i^2 + \omega^2} \right) * \left(\frac{\delta(\omega - \omega_{di}) - \delta(\omega + \omega_{di})}{i} \right). \tag{3.14}$$

Examining Equation 3.14, we see that the transfer function of a single mode is the convolution of a spike at its resonant frequency and a Lorentzian distribution (the Fourier transform of the decaying exponential) with a width that depends on modal frequency and damping. The impulse response and transfer function for our system are then simple sums of their modal components:

$$h(t) = \sum_i h_i(t) \tag{3.15}$$

$$H(\omega) = \sum_i H_i(\omega) \tag{3.16}$$

which tells us that the transfer function of the system is a collection of spikes at resonant frequencies, convolved with Lorentzians that depend on modal frequencies and damping.

3.4.2   Modal Frequencies and Motion Spectra

Equations 3.14 and 3.16 give us a convenient way to identify modal frequencies ω_i of an object by observing how the object responds to a broad-spectrum force. For an impulse force, the motion of the object is simply its impulse response $h(t)$, and the Fourier transform of that motion is the transfer function $H(\omega)$. In this case, finding modal frequencies ω_i amounts to finding peaks in the power spectrum of observed motion.

In the more general case, this peak estimation is analogous to the color constancy problem in computer vision[2]: we observe the product of a forcing spectrum (analogous to illumination) and the transfer function of our object (analogous to reflectance), and our task is to estimate the transfer function alone. In Chapter 5 we will solve this ambiguity by exciting objects with a broad-spectrum force that we control. In Chapter 6 we will sometimes instead assume that an unknown force is approximately broad-spectrum, and examine the potential consequences of that assumption.

3.4.3   Averaging Local Spectra

Recall that the frequencies ω_i are global properties of an object – they do not vary across the object's surface. This means that the power spectra of all local motion signals $u_\ell(t)$ that measure the same object should have spikes at the same resonant frequencies. This holds even if two points on the object move with opposite phase or in different directions at a given frequency. This observation

---

[2]Color constancy refers to the problem of estimating the reflectance of objects when the color we see is actually a product of reflectance and illumination. This problem is under-constrained, but can often be addressed by assuming that illumination is approximately broad-spectrum (i.e. white or gray).

allows us to average across the spectra of all local motion signals and find mode frequencies ω_i by looking for peaks in a single global spectrum.

As we did when computing $\hat{s}(t)$ in Section 3.2, we want to weigh our local motion signals according to their confidence values. To do this we use the same weighted motion signals $\mathfrak{u}_\ell(t)$ in our calculations. However, as was not the case with $\hat{s}(t)$, here we want motion with opposite phase to add constructively. To accomplish this, we simply add the power spectra of local motion signals. The resulting global power spectrum $P(\omega)$ is computed as:

$$P(\omega) = \sum_\ell \mid \mathcal{F}(\mathfrak{u}_\ell(t)) \mid^2 \qquad (3.17)$$

where $\mathcal{F}$ denotes the Fourier transform.

3.4.4   Viewpoint Invariance

An advantage of using the temporal spectra $P(\omega)$ in computer vision applications is that they offer invariance to changes in scale and viewpoint. This invariance comes from the fact that resonant frequencies are global properties of an object, meaning that we see the same frequencies from different perspectives. In Chapter 5 we use this to learn the material properties of objects in a scenario where training and testing data are captured from different cameras and perspectives.

3.4.5   Damping

Under a broad-spectrum excitation force, the recovered spectra $P(\omega)$ should take the shape of an object's transfer function. In Section 3.4.1 we showed that damping determines the width of resonant spikes in this transfer function. Therefore, by observing the width of resonant spikes in recovered motion spectra we can reason about damping in an object. The relationship between motion power spectra and damping can be learned through observation, or estimated explicitly by fitting Lorentzian distributions to spikes in $P(\omega)$.

> **Looking Ahead at Chapters 5 and 6:**   In Chapter 5 we use the motion power spectra $P(\omega)$ as a feature to estimate physical properties of objects. We explore learning and explicit measurement-based approaches, experimenting on different classes of objects and materials. In Chapter 6 we use $P(\omega)$ to help find vibration modes, which we then use to simulate objects in image-space.

3.5   Modal Images

While the modal frequencies ω_i are global properties of an object, local vibrations across the object are scaled according to the mode shapes ϕ_i. In video, this means that the contribution of each vibration mode to the local spectra $U_\ell(\omega)$ should be scaled by a projection of the corresponding shape ϕ_i. We can estimate this projected shape by computing what we call a *modal image*, which is an image $\mathcal{U}_{\omega_i}(x, y)$ of temporal Fourier coefficients corresponding to motion at a common frequency,
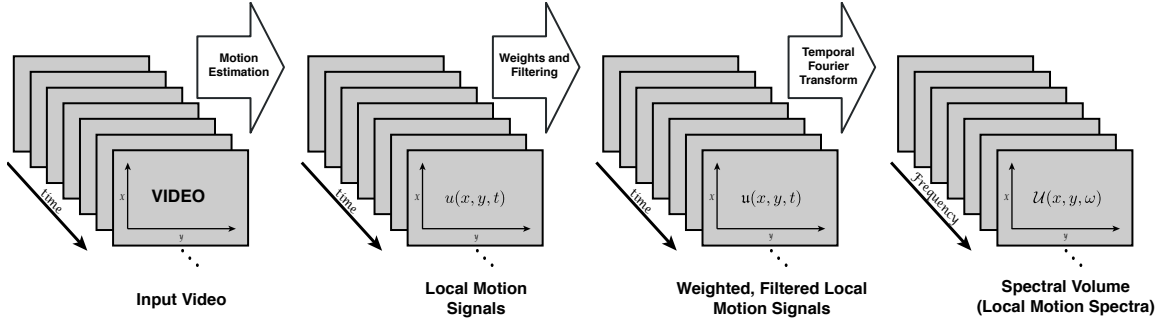
Figure 3-3: Steps for computing a spectral volume.

or set of frequencies associated with a common mode. In Chapter 6 we show that modal images can be used as a modal basis for motion in image space. Here we discuss how these images are computed and visualized.

3.5.1 Spectral Volumes

Up until now we have largely avoided any dependence on spatially-varying properties of our data, allowing us to simplify our derivations by operating on local motion signals $u_\ell(t)$ without regard for their position in an image. Mode shapes are inherently spatially-varying though. To address this we take a slightly different view of our data, illustrated in Figure 3-3.

In the simplest case, our modal images are frequency slices of a spectral volume $U(x, y, \omega) = \mathcal{F}_t(u(x, y, t))$, where $\mathcal{F}_t$ denotes a Fourier transform along the time dimension only. However, in practice we weigh and filter our local motion signals before taking their Fourier transform, using weights and filters that vary slightly from application to application. To simplify our discussion here, we denote the weighted and filtered motion signals $\mathfrak{u}_\ell(t)$ as before and discuss different ways to calculate them later, giving us weighted local spectra:

$$\mathcal{U}_\ell(\omega) = \mathcal{F}(\mathfrak{u}_\ell(t)) \tag{3.18}$$

and spectral volume:

$$\mathcal{U}(x, y, \omega) = \mathcal{F}_t(\mathfrak{u}(x, y, t)) \tag{3.19}$$

where $\mathcal{F}_t$ denotes a Fourier transform along the time dimension only. Our modal images are then frequency-slices of the filtered spectral volume $\mathcal{U}(x, y, \omega)$.

3.5.2 Weighing and Filtering

As before, we use local image contrast to weigh our signals. However, now we do additional filtering of these signals, and normalize the result differently depending on our application.

Filtering: Filtering is done in image space on a slice $u_t(x, y)$ of local motion signals at time t. We first weigh this image of local displacements $u_t(x, y)$ according to the squares of local contrast. We then apply a Gaussian blur with standard deviation σ_b to the resulting weighted image.
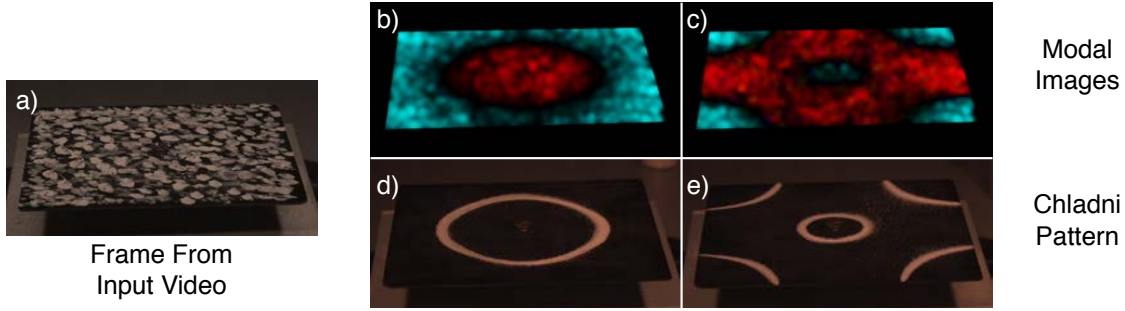
Figure 3-5: Modal Images: (left) A single frame of input video; (right top)Recovered modal images at ω_i = (b)180Hz and (c) 525Hz for motion in the y dimension. Each modal image is shown above a corresponding image obtained using the Chladni method ((d) for 180Hz and (e) for 525Hz). The Chladni method works by shaking sand away from vibrating parts of the plate, causing it to gather at nodal lines. We see that the nodal lines predicted by the Chladni method are recovered in our modal images.

Normalization: While scale ambiguity is not much of a problem in global signals, it can become a problem for modal images in some applications. This is because different parts of a modal image can scale differently depending on local weights, effectively warping the projected mode shape by a function of local image contrast. This is not a problem in applications where modal images are used only to visualize vibration modes (it can even help by acting as a mask for noise). However, when modal images are used to synthesize motion as we do in Chapter 6, additional normalization is necessary. For this we use an approach similar to [75] to normalizing each $\mathfrak{u}_t(x, y)$ by the sum of weights that contributed to it. We first filter an image $A_t^2(x, y)$ of local weights, separately from our image of local displacements. We then divide the filtered image of weighted local displacements by the filtered image of weights.

3.5.3   Visualization

We visualize modal images by mapping the amplitude and phase of each temporal Fourier coefficient in $\mathcal{U}_{\omega_i}(x, y)$ to intensity and hue, respectively. While phase maps naturally to hue, different maps for intensity can highlight different information about a mode. We typically map the minimum and maximum amplitudes to zero and one, applying an exponent within this range to adjust contrast as desired.
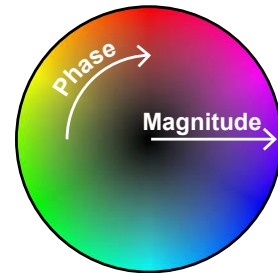
Figure 3-4: Modal Image Visualization Key

3.5.4   Visual Interpretation

The easiest way to interpret our visualization of modal images is to look for patterns in phase and nodal lines. Regions of the image with opposite phase have opposite colors on the color wheel, and nodal lines are dark lines, usually at the boundary between regions with opposite phase. In Figure 3-5 we compare the nodal lines in our visualization to lines found using the Chladni sand plate method. Both image the mode shapes of a rectangular metal plate that is shaken by a vibration generator in controlled settings. We compare visualizations at two modal frequencies, and see that
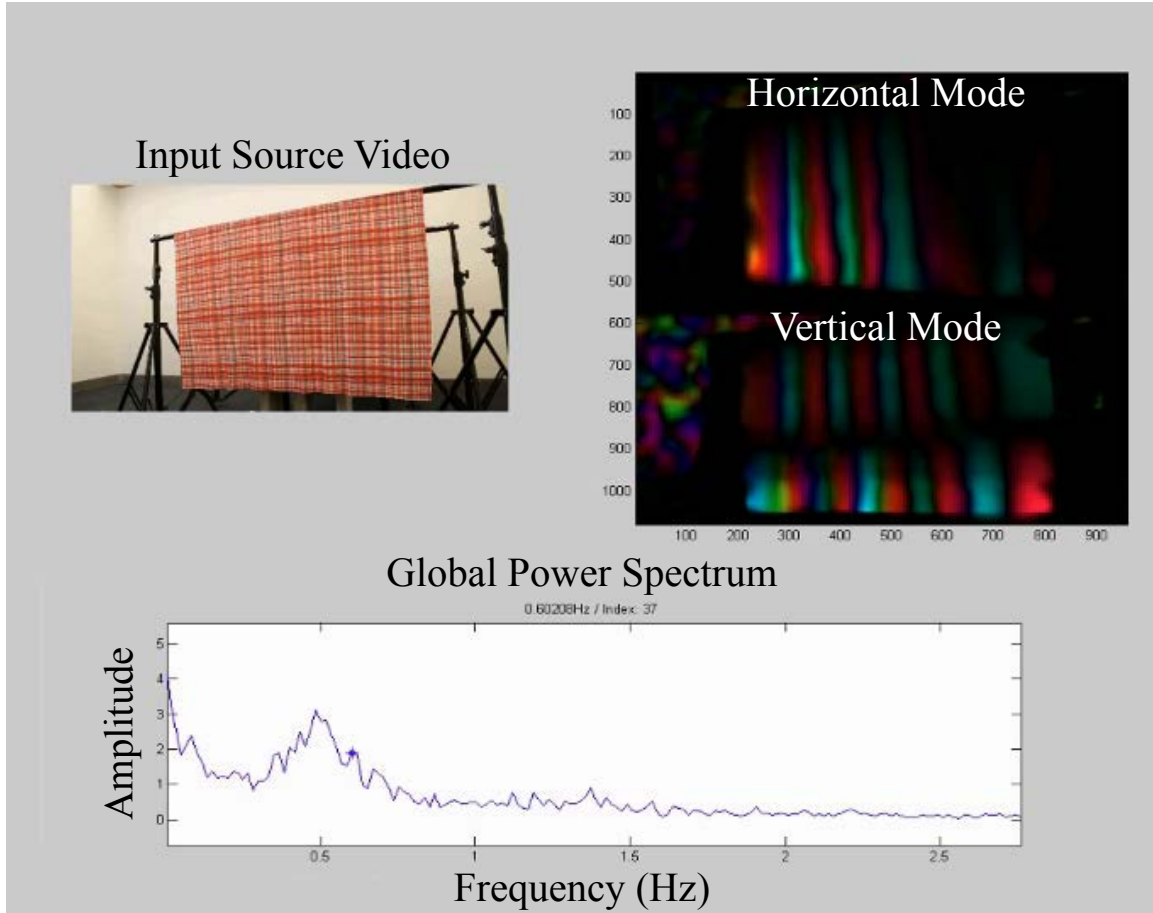
Figure 3-6: To use our mode selection interface, users click on a frequency in the video's global power spectrum (bottom) and are shown a visualization of the corresponding candidate modal image (right).

the nodal lines predicted by our visualization agree with the Chladni method. Unlike the Chladni method, our approach also visualizes the relative phase of different points on the object.

3.5.5 Mode Selection

It is often the case in practice that modal frequencies ω_i are unknown. This can make searching for the real vibration modes in a spectral volume a difficult task. When we don't have a prior on modal frequencies for an object, we address this by providing the user with a manual mode selection interface, as shown in Figure 3-6

Our interface displays a representative image from the input video, the global power spectrum recovered from the video, and a visualization of the current selected candidate mode, chosen by the user. When a user clicks on the spectrum in our interface, we find the frequency with maximum energy in a small window around the user's mouse, and display the corresponding candidate mode in our shape window. This selection process is similar to peak-picking methods that have been used for modal identification of structures in engineering [26].

3.5.6 Modal Imaging in the Wild

> Note: Much of this subsection is anecdotal, as it presents one experiment done in a very uncontrolled setting (from the window seat of a commercial aircraft). The point here is to show that our work is well suited for such settings. I have provided as much information as possible, but have intentionally omitted a few details, like the specific airline and aircraft, to avoid upsetting the FAA. Please contact me directly for more details.

Modal imaging can be a powerful tool for analyzing vibrations in real world setting. Cameras offer a number of significant advantages over other more traditional vibration sensors. Because cameras are passive and offer significant spatial resolution, it is often possible to simply point a camera at an object and analyze its vibrations with little setup or advanced planning. Figure 3-7 shows an anecdotal but compelling example.
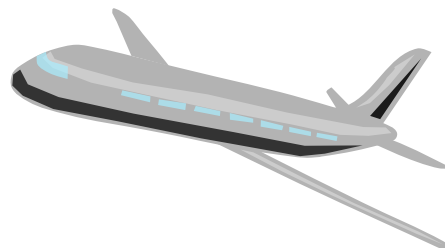
In the summer of 2015 I took a flight from Boston to San Francisco to present our paper "Long Distance Video Camera Measurements Of Structures" [17] at the International Workshop on Structural Health Monitoring (IWSHM 2015) at Stanford. Noting that several other papers at the conference focused on structural health monitoring of aircraft using modal analysis, I decided to run an experiment from the window seat of my flight to the conference. Using a sponge, I propped my cell phone up against the window (Figure 3-7 top left), pointed at the airplane's wing, and recorded about two minutes of 30fps video (Figure 3-7 bottom left).

Upon arriving at my hotel for the workshop, I selected 102 seconds of clean video (cropping portions where I had to adjust the camera at the beginning and end) and ran projection-only video stabilization on the remaining footage using Adobe After Effects CS6. I then uploaded the stabilized video and ran our modal imaging code. The right half of Figure 3-7 shows our mode selection interface, and a selected mode at 2.5245Hz. I predicted that this mode was the dominant flex mode of the airplane wing, and reported the result to Justin G. Chen, my frequent coauthor from the Civil Engineering Department at MIT.

Justin was able to find the model aircraft that I had taken by looking up the flight number. He was then able to find a reference for the dominant flex mode of that aircraft, which was reported as 2.584 Hz [9]. This is less than 3% off from the frequency we predicted from video. Considering the effects of temperature and fuel in the wing, and that the video was taken on the fly with no special equipment (other than a common dishwashing sponge) this is an extremely impressive result. I presented the experiment in my talk at IWSHM to a very positive response.

> **Looking Ahead at Chapter 6**   In Chapter 6 we examine modal images more closely, and show that they can be used as an image-space basis for the motion of objects in video.
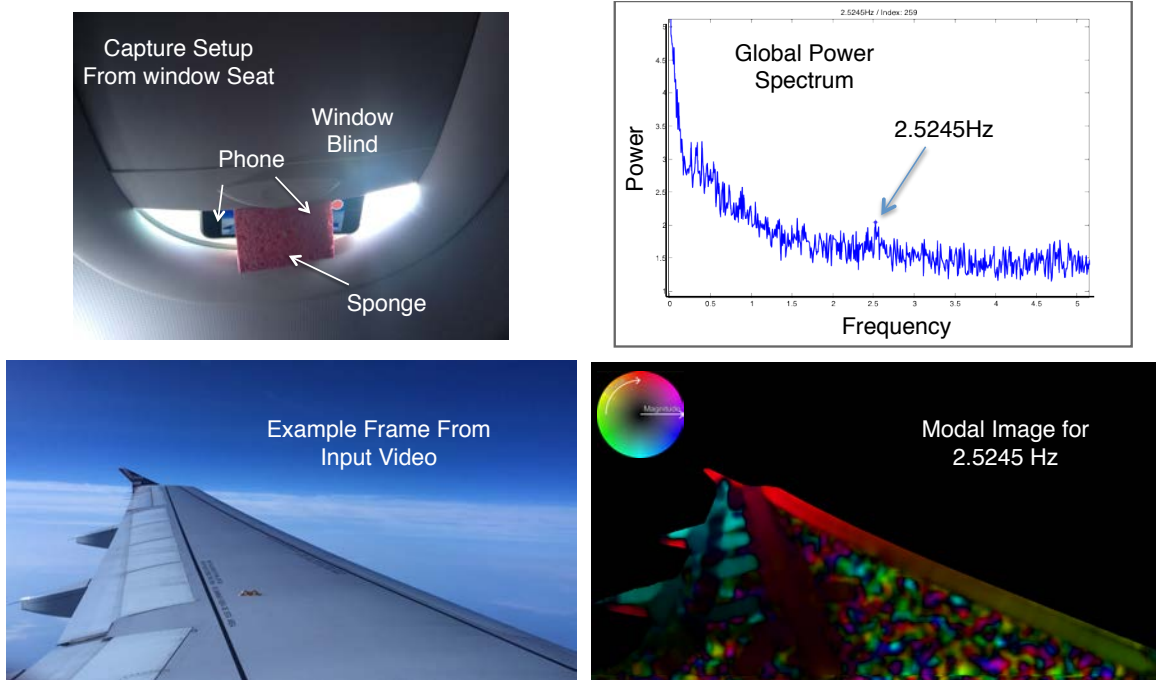
Figure 3-7: On a flight from Boston to San Francisco to present our paper [17] at the International Workshop on Structural Health Monitoring (IWSHM), I filmed the airplane wing from a window seat in the cabin using my cell phone. To record the video, I wedged my phone under the window blind with a common household sponge (top left). On the bottom left we see a frame from the captured video. On the top right we see the global power spectrum recovered from the video. On the bottom left we see the modal image for a predicted flex mode at 2.5245 Hz, which is less than 3% off from the reported value of 2.584 Hz [9]. The modal image is for the x dimension of the video. The opposite blue/red phase relationship on parts of the wing are likely the result of rotation relative to the optical axis of the camera.

3.6 Glossary of Variables

3.6.1 Object Variables

$x(t)$: Object motion

ω_i: The vibration mode frequency for the ith vibration mode.

ϕ_i: The vibration mode shape for the ith vibration mode.

Φ: The transformation from object degrees of freedom to modal coordinates given by a matrix with the mode shapes $\phi_1...\phi_N$ as columns.

3.6.2 Algorithm Variables

r, θ, x, y, t: The dimensions of our motion signals – scale, orientation, x, y, and time, respectively.

$A(r, \theta, x, y, t)$: Local Image Contrast

$u(r, \theta, x, y, t)$: Local displacement or deformation

$u_\ell(t)$: Local motion signal, where ℓ indexes a particular location, orientation, and scale.

$\mathfrak{u}_\ell(t)$: Weighted local motion signals (typically weighted by the square of local contrast).

σ_b: Standard deviation of image space blur for filtering local displacement images.

$U_\ell(\omega)$: An unweighted local motion spectrum.

$\mathcal{U}_\ell(\omega)$: A weighted local motion spectrum.

$P(\omega)$: The global power spectrum of a video.

$U(x, y, \omega)$: The unweighted spectral volume of motion.

$\mathcal{U}(x, y, \omega)$: The weighted and filtered spectral volume of motion.

$\mathcal{U}_\omega(x, y)$: The modal image corresponding to frequency ω

3.7 Implementation Details

The algorithms in this chapter originally developed over the course of several projects and publications, three of which are described in Chapters 4, 5, and 6. As such, some details may be slightly different in different chapters. Here we clarify some of the differences, and provide some additional details. In addition to this section, I have left some redundancy in subsequent chapters for further clarification.

3.7.1 Handling Orientation and Scales

The numbers of orientations and scales are additional parameters of our motion estimation (and more generally of the CSP). These parameters can be important in practice, but in ways that depend closely on the other aspects of the experimental setup. Here we discuss how orientations and scales were chosen.

Orientations   Two orientations is generally enough for any of the applications that we considered, as motion in the image plane cannot be orthogonal to both the x and y dimensions of a video. For the applications in Chapters 4 and (to a slightly lesser extend) 5, a single well-chosen orientation is often sufficient if it aligns with the dominant movement of an object. However, we use multiple orientations (generally 2) to ensure that we don't 'miss' any motion. If included, extra orientations beyond 2 generally increase processing time, but have little effect on results.

Scales   Each scale of the CSP represents a different spatial frequency. In theory, the motion at different scales could be different, but this tends not to be the case. Strong edges in an image have broad-spectrum frequency content, and when those edges move rigidly it results in similar phase shifts across all scales. As a result, motions at different scales tend to look very similar. However, it is possible for motion to be 'too big' or 'too small' for certain scales. To understand this, consider the size of a single filter in the CSP filter bank. If the motion in a video is bigger than this filter, it will cause phase-wrapping at the corresponding scale.[3]   On the other hand, if the motion is much smaller than the filter, then it spans a smaller range of phases, lowering precision and raising the noise floor. Our precise strategy for selecting and combining scales varied from application to application: in Chapter 4 we used the strategy described in 3.2, and in Chapters 5 and 6 we used a single scale. In Chapter 5 we simply chose the finest scale, but in Chapter 6 we sometimes chose other scales depending on our mode selection interface. A very simple strategy that worked on all of the applications we considered was to compute several scales at once and simply choose the best result after the fact. As we do this, we keep the standard deviation σ_b of our Gaussian filter for modal images constant in pixels, effectively letting us test out a few different standard deviations at once.

3.7.2 Modal Image Details

We first introduced a visualization of modal images in our work described in Chapter 4. However, our approach was a bit different at first. In Chapters 4 and 5 modal images are computed on the sums of motion in x and y, and the visualizations are not normalized. In 6 we visualize x and y separately, normalize the results, and frequently use a mask to cover noisy regions of the image. All of our modal imaging code and our selection interface are written in MATLAB.

---

[3]Actually, the effect is generally worse than simple phase wrapping, as it depends on adjacent image texture.

3.8   Conclusion

In this chapter we have presented many of the common threads that unite the following chapters. The work described here originally evolves over several projects and publications. It is my hope that presenting these ideas together will provide a perspective that better inspires future work.

4

The Visual Microphone

When sound hits an object, it causes small vibrations of the object's surface. In this chapter we show how these vibrations can be extracted from high-speed video and used to recover the sounds that produced then — letting us passively turn everyday objects into visual microphones from a distance.

Our approach is simple, but effective. To recover sound from an object, we film the object using a high-speed video camera and extract local motion signals from the recorded video. We then align and average these local signals into a single, 1D signal that captures global movement of the object over time. This global signal is then filtered and denoised to produce a recovered sound.

Most of this chapter focuses on experimentation and analysis, through which we validate our approach to extracting vibrations from video. We recover sounds from high-speed footage of a variety of objects with different properties, and use both real and simulated data to examine factors that affect the accuracy of what we recover. We evaluate the quality of recovered sounds using intelligibility and SNR metrics, and provide input and recovered audio samples for direct comparison. Finally, in Section 4.5 we explore how to leverage the rolling shutter in regular consumer cameras to recover audio from standard frame-rate videos.

4.1   Related Work

Traditional microphones work by converting the motion of an internal diaphragm into an electrical signal. The diaphragm is designed to move readily with sound pressure so that its motion can be recorded and interpreted as audio. Laser microphones work on a similar principle, but instead

---

Most of this chapter was originally published in our paper [24] in collaboration with Michael Rubinstein, Neal Wadhwa, Gautham Mysore, Frédo Durand, and William T. Freeman. (URL)
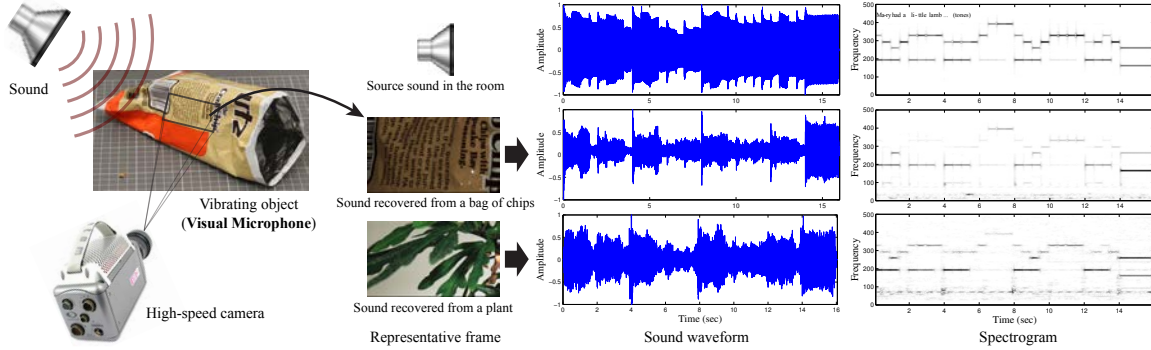
Figure 4-1: Recovering sound from video. Left: when sound hits an object (in this case, an empty bag of chips) it causes extremely small surface vibrations in that object. We are able to extract these small vibrations from high speed video and reconstruct the sound that produced them - using the object as a visual microphone from a distance. Right: an instrumental recording of "Mary Had a Little Lamb" (top row) is played through a loudspeaker, then recovered from video of different objects: a bag of chips (middle row), and the leaves of a potted plant (bottom row). For the source and each recovered sound we show the waveform and spectrogram (the magnitude of the signal across different frequencies over time, shown in linear scale with darker colors representing higher energy). The input and recovered sounds for all of the experiments in the chapter can be found on the project web page.

measure the motion of a distant object, essentially using the object as an external diaphragm. Laser microphone can recover high quality audio from great distances, but require precise positioning of a laser and receiver, and require that surfaces be at least partly retro-reflective.

Zalevsky et al. [80] address some of these limitations by using an out-of-focus high-speed camera to record changes in the speckle pattern of reflected laser light. Their work allows for greater flexibility in the positioning of a receiver, but still depends on recording reflected laser light. In contrast, our technique does not depend on active illumination.

4.2   Recovering Sound from Video

Figure 4-3 gives a high-level overview of how the visual microphone works. An input sound (the signal we want to recover) consists of fluctuations in air pressure at the surface of some object. These fluctuations cause the object to move, resulting in a pattern of displacement over time that we film with a camera. We then process the recorded video with our algorithm to recover an output sound.

The input to our method is a video, $V(x, y, t)$, of an object. In this section we consider high speed video (1kHz-20kHz). Lower frame rates are discussed in Section 4.5. We assume that the relative motion of our object and camera is dominated by vibrations due to a sound signal, $s(t)$. Our goal is to recover $s(t)$ from V.

Our method is to first compute the global motion signals $\hat{s}(t)$ we discussed in Section 3.2 of Chapter 3, and then apply audio denoising and filtering techniques to obtain our recovered sound.
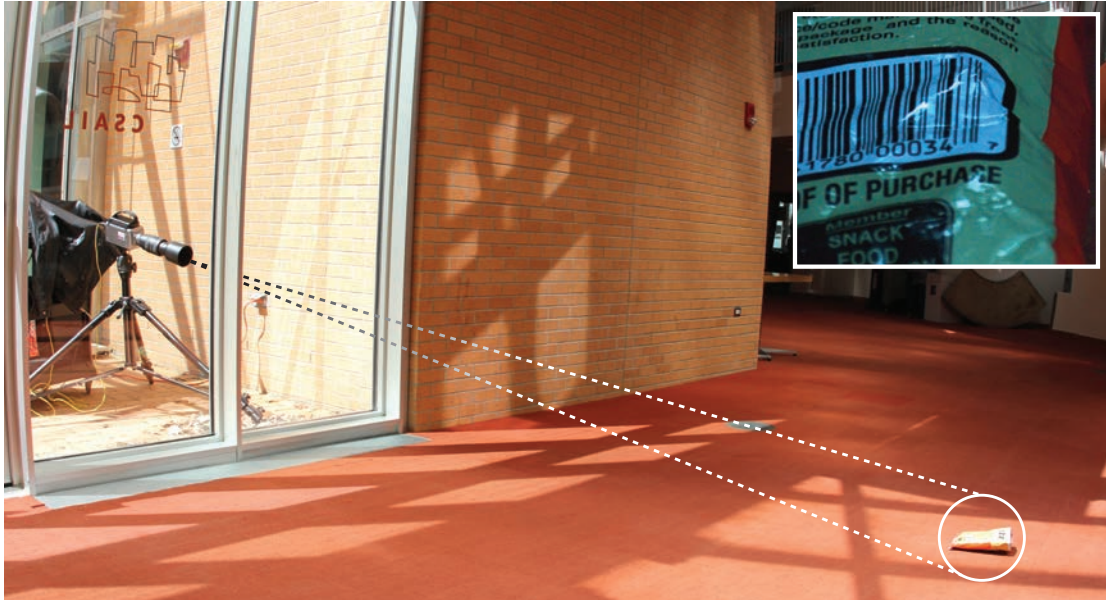
(a) Setup and representative frame
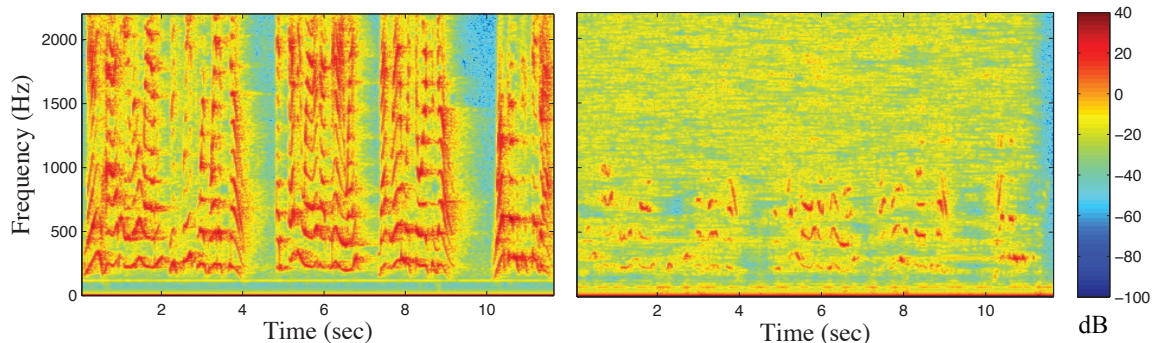
(b) Input sound   (c) Recovered sound

Figure 4-2: Speech recovered from a 4 kHz video of a bag of chips filmed through soundproof glass. The chip bag (on the floor on the bottom right in (a)) is lit by natural sunlight only. The camera (on the left in (a)) is positioned outside the room behind thick soundproof glass. A single frame from the recorded video (400×480 pixels) is shown in the inset. The speech "Mary had a little lamb ... Welcome to SIGGRAPH!" was spoken by a person near the bag of chips. (b) and (c) show the spectrogram of the source sound recorded by a standard microphone next to the chip bag, and the spectrogram of our recovered sound, respectively. The recovered sound is noisy but comprehensible (the audio clips are available on the project web page).
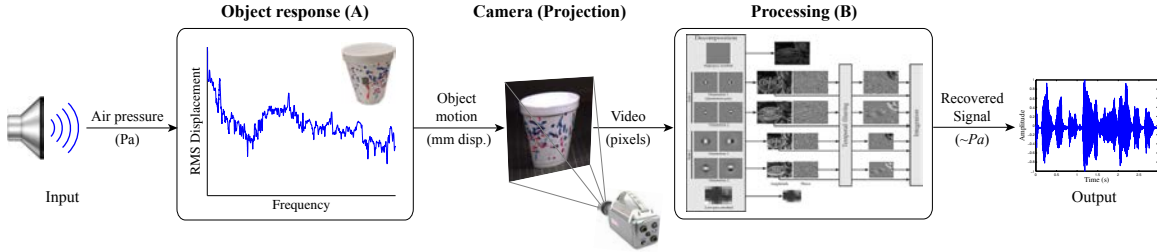
Figure 4-3: We model the visual microphone as a system that operates on sound. Component **A** (Section 4.4.1) models an object's response to sound, and is purely physical—taking as input changes in air pressure, measured in Pascals, and producing physical displacement of the object over time, measured in millimeters. The response of the object to the sound depends on various factors such as the sound level at the object, and the object's material and shape. A camera then records the object, transforming the physical displacements into pixel motions in a video. Component **B** (Section 5.3, Section 4.4.2) is our spatiotemporal processing pipeline, which transforms the motions in the video back into sound. The resulting 1D signal is unit-less, but is correlated with the input Pascals and can therefore be played and analyzed as sound.

4.2.1   Denoising

We further process the recovered global motion signal to improve its SNR. In many videos, we noticed high energy noise in the lower frequencies that typically did not correspond to audio. We address this by applying a high pass Butterworth filter with a cutoff of 20-100Hz (for most examples, 1/20 of the Nyquist frequency)[1].

Our choice of algorithm for additional denoising depends on our target application – specifically, whether we are concerned with accuracy or intelligibility. For applications targeting accuracy we use our own implementation of a technique known as spectral subtraction [6]. For intelligibility we use a perceptually motivated speech enhancement algorithm [50] that works by computing a Bayesian optimal estimate of the denoised signal with a cost function that takes into account human perception of speech. All of the results we present in this chapter were denoised automatically with one of these two algorithms. Our results may be further improved by using more sophisticated audio denoising algorithms available in professional audio processing software (some of which require manual interaction).

Different frequencies of our recovered signal might be modulated differently by the recorded object. In section 4.3.3, we show how to use a known test signal to characterize how an object attenuates different frequencies, then use this information to equalize unknown signals recovered from the same object (or a similar one) in new videos.

---

[1]For very noisy cases we instead apply this highpass filter to the intermediate signals $a(r, \theta, t)$ before alignment to prevent the noise from affecting the alignment.

4.3 Experiments

We performed a variety of experiments to test our technique. All the videos in this section were recorded indoors with a Phantom V10 high speed camera. The setup for these experiments consisted of an object, a loudspeaker, and the camera, arranged as shown in Figure 4-4. The loudspeaker was always placed on its own stand separate from the surface holding the object in order to avoid contact vibrations. The objects were lit with photography lamps and filmed at distances ranging from 0.5 meter to 2 meters. In other experiments we recover sound from greater distances without the aid of photography lamps (e.g. Figure 4-2). Video frame rates are in the range of 2kHz-20kHz, with resolutions ranging from 192x192 pixels to 700x700 pixels. Sounds were played at loud volumes ranging from 80 dB (an actor's stage voice) to 110 dB (comparable to a jet engine at 100 meter). Lower volumes are explored in Section 4.4, Figure 4-2, and additional experiments on our web page. Videos were processed using complex steerable pyramids with 4 scales and 2 orientations, which we computed using the publicly available code of Portilla and Simoncelli [57]. Processing each video typically took 2 to 3 hours using MATLAB on a machine with two 3.46GHz processors and 32GB of RAM.

Our first set of experiments tested the range of frequencies that could be recovered from different objects. We did this by playing a linear ramp of frequencies through the loudspeaker, then seeing which frequencies could be recovered by our technique. The second set of experiments focused on recovering human speech from video. For these experiments we used several standard speech examples from the TIMIT dataset [30] played through a loudspeaker, as well as live speech from a human subject (here the loudspeaker in Figure 4-4 was replaced with a talking human). Audio for these experiments and others can be found on the project website. Our results are best experienced by listening to the accompanying audio files through headphones.

4.3.1 Sound Recovery from Different Objects/Materials

In this first set of experiments we play a ramp signal, consisting of a sine wave that increases linearly in frequency over time, at a variety of objects. Figure 4-5(a) shows the spectrogram of our input sound, which increases from 100Hz to 1000Hz over 5 seconds. Figure 4-5(b) shows the spectrograms of signals recovered from 2.2kHz videos of a variety of objects with different material properties. The brick at the top of Figure 4-5(b) is used as a control experiment where we expect to recover little signal because the object is rigid and heavy. The low-frequency signal recovered from the brick (see the spectrogram visualized for *Brick* in Figure 4-5(b)) may come from motion of the brick or the camera, but the fact that this signal is very weak suggests that camera motion and other unintended factors in the experimental setup have at most a minor impact on our results. In particular, while almost no signal is recovered from the brick, much better signal is recovered from the other objects shown.

In almost all of our results the recovered signal is weaker in higher frequencies. This is expected, as higher frequencies produce smaller displacements and are attenuated more heavily by most materials. We show this more explicitly with data from a laser Doppler vibrometer in Section 4.4. However, the decrease in power with higher frequencies is not monotonic, possibly due to the excitement of vibration modes. Not surprisingly, lighter objects that are easier to move tend to support the recovery of higher frequencies better than more inert objects.
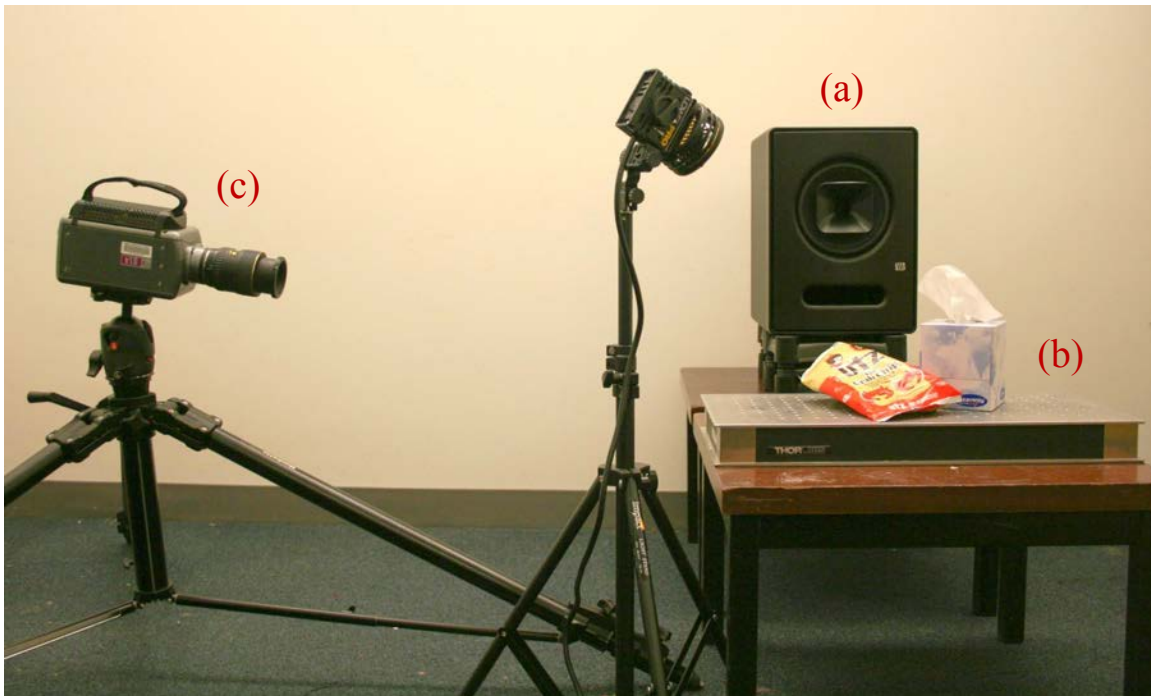
Figure 4-4: An example of our controlled experimental setup. Sound from an audio source, such as a loudspeaker (a) excites an ordinary object (b). A high-speed camera (c) records the object. We then recover sound from the recorded video. In order to minimize undesired vibrations, the objects were placed on a heavy optical plate, and for experiments involving a loudspeaker we placed the loudspeaker on a separate surface from the one containing the objects, on top of an acoustic isolator.

(a) Input sound (played in the room)

(b) Reconstructed sound
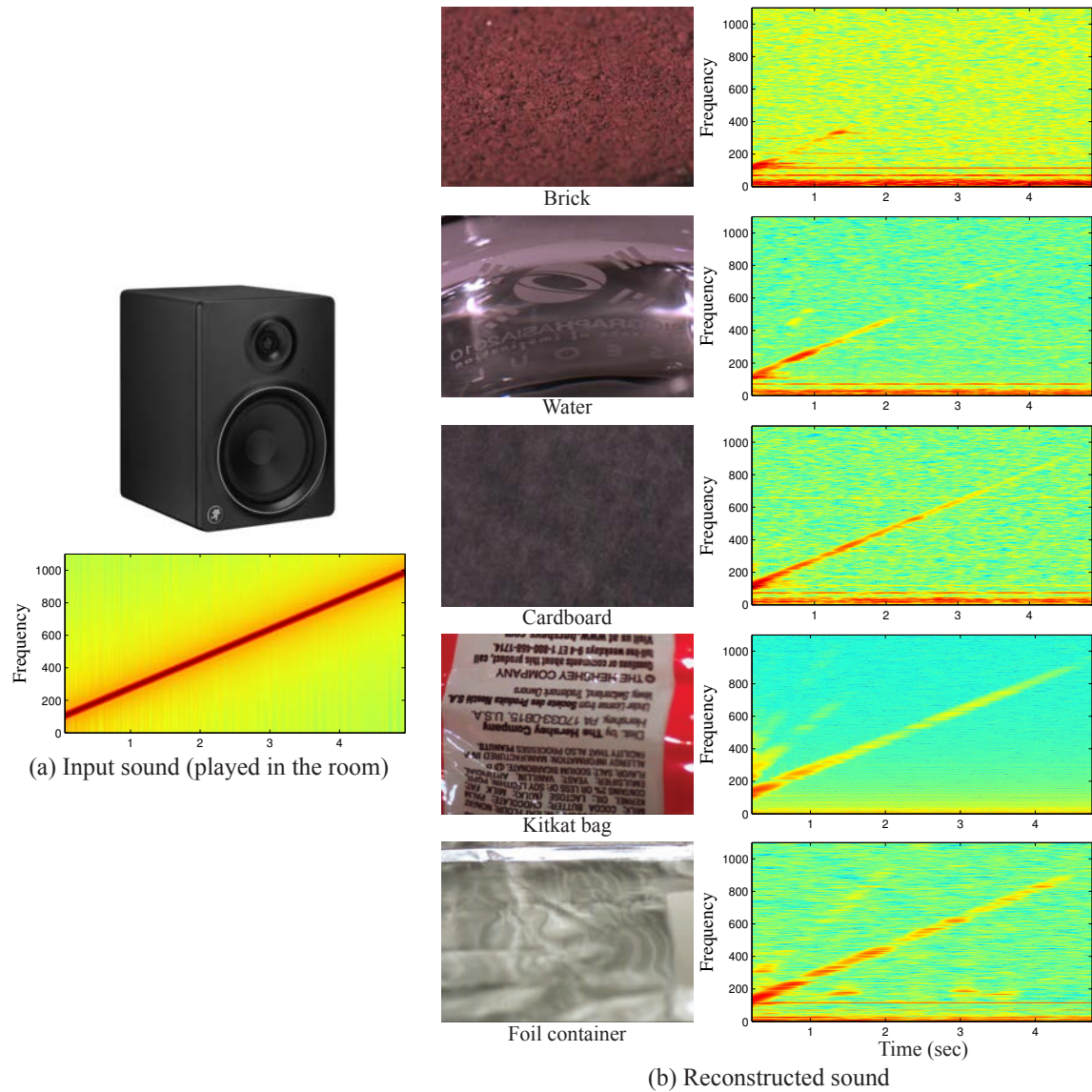
Figure 4-5: Sound reconstructed from different objects and materials. A linear ramp ranging from $100 - 1000$Hz was played through a loudspeaker (a), and reconstructed from different objects and materials (b). In *Water*, the camera was pointed at one side of a clear mug containing water, where the water surface was just above a logo printed on the side of the mug. Motion of the water's surface resulted in changing refraction and moving specular reflections. More details can be found on our project web page.

Sequence	Method	SSNR	LLR Mean	Intelligibility
Female speaker - fadg0, sa1	VM	24.5	1.47	0.72
	LDV	**28.5**	**1.81**	**0.74**
Female speaker - fadg0, sa2	VM	**28.7**	1.37	0.65
	LDV	26.5	**1.82**	**0.70**
Male speaker - mccs0, sa1	VM	20.4	1.31	0.59
	LDV	**26.1**	**1.83**	**0.73**
Male speaker - mccs0, sa2	VM	23.2	1.55	0.67
	LDV	**25.8**	**1.96**	**0.68**
Male speaker - mabw0, sa1	VM	23.3	1.68	**0.77**
	LDV	**28.2**	**1.74**	0.76
Male Speaker - mabw0, sa2	VM	25.5	1.81	0.72
	LDV	**26.0**	**1.88**	**0.74**

Table 4.1: A comparison of our method (VM) with a laser Doppler vibrometer (LDV). Speech from the TIMIT dataset is recovered from a bag of chips by both methods simultaneously. Both recovered signals are denoised using [50]. The recovered signals are evaluated using Segmental SNR (SSNR, in dB) [35], Log Likelihood Ratio mean (LLR) [59] and the intelligibility metric described in [72] (given in the range 0-1). For each comparison, the better score is shown in bold.

4.3.2 Speech Recovery

Speech recovery is an exciting application of the visual microphone. To test our ability to recover speech we use standard speech examples from the TIMIT dataset [30], as well as live speech from a human speaker reciting the poem "Mary had a little lamb," in reference to the first words spoken by Thomas A. Edison into the Phonograph in 1877. Additional speech experiments can be found on the project website.

In most of our speech recovery experiments, we filmed a bag of chips at 2200 FPS with a spatial resolution of 700×700 pixels. Recovered signals were denoised with a perceptually motivated speech enhancement algorithm [50], described in section 4.2.1.

The best way to evaluate our reconstructed speech is to listen to the accompanying audio files, available on our project website. In addition to providing these audio files, we also evaluate our results using quantitative metrics from the audio processing community. To measure accuracy we use Segmental Signal-to-Noise Ratio (SSNR) [35], which averages local SNR over time. To measure intelligibility we use the perceptually-based metric of Taal et al. [72]. For our results in Table 4.1 we also include Log Likelihood Ratio (LLR) [59], which is a metric that captures how closely the spectral shape of a recovered signal matches that of the original clean signal. Finally, our results can be evaluated visually by looking at the spectrograms of our input speech and recovered signals, shown in Figure 4-6.

Up to the Nyquist frequency of our videos, the recovered signals closely match the input for both pre-recorded and live speech. In one experiment, we captured a bag of chips at 20,000 FPS and were able to recover some of the higher frequencies of the speech (Figure 4-6, bottom right). The higher frame rate resulted in reduced exposure time and therefore more image noise, which is why the resulting figure is noisier than the results at 2200Hz. However, even with this added noise, we were able to qualitatively understand the speech in the reconstructed audio.

We also compare our results to audio recovered by a laser Doppler vibrometer (Table 4.1). Our
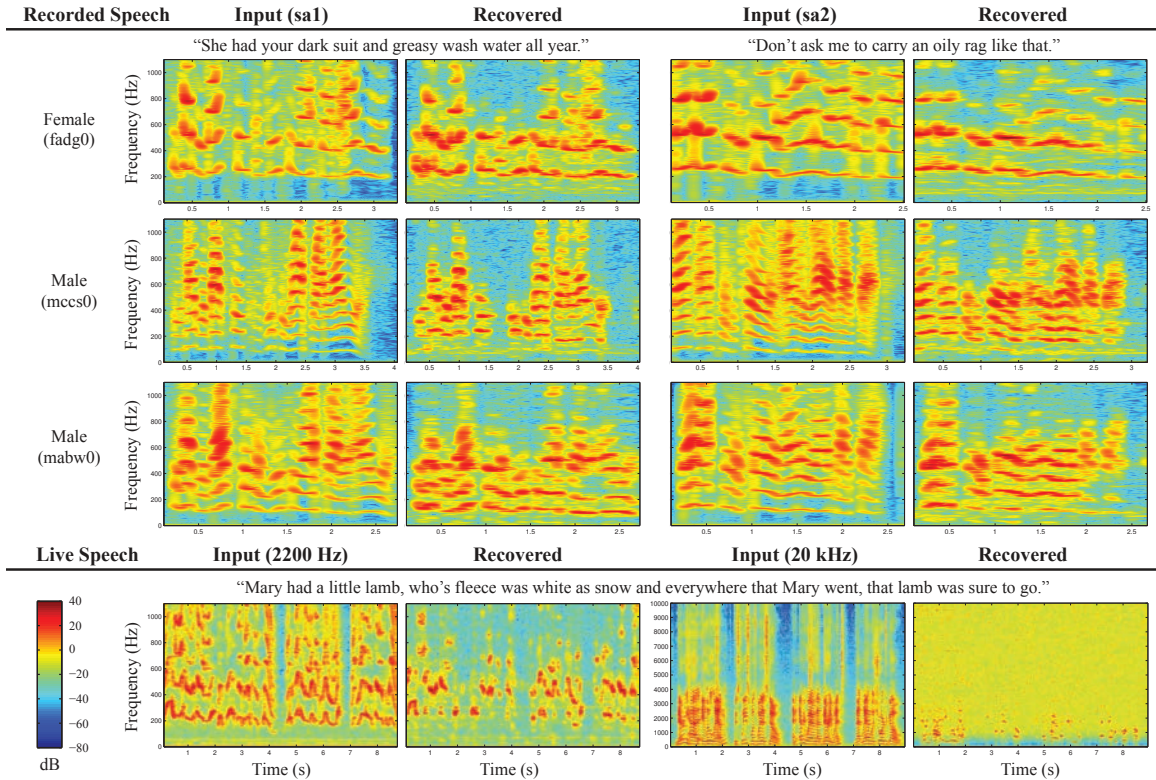
Figure 4-6: Speech recovered from a bag of chips. **Recorded Speech (top three rows):** We play recordings of three speakers saying two different sentences from the TIMIT dataset [30] through a loudspeaker near a bag of chips. We then recover audio from a 2,200Hz, 700×700 video of the bag of chips (see table 4.2(a)) for a representative frame) and display the spectrograms of both the input audio and the recovered signal. **Live Speech (bottom row):** In a separate experiment, a male speaker recites the nursery rhyme "Mary had a little lamb...", near the same bag of chips. We display the spectrograms of audio recorded by a conventional microphone next to the spectrograms of the audio recovered from video of the bag of chips using our technique. Results were recovered from videos taken at 2,200Hz, 700× 700 pixels (bottom left), and 20 kHz, 192× 192 pixels (bottom right). Input and recovered audio clips can be found on the project web page.

(a) Input   (b) Spectrogram   (c) Log Transfer coefficients
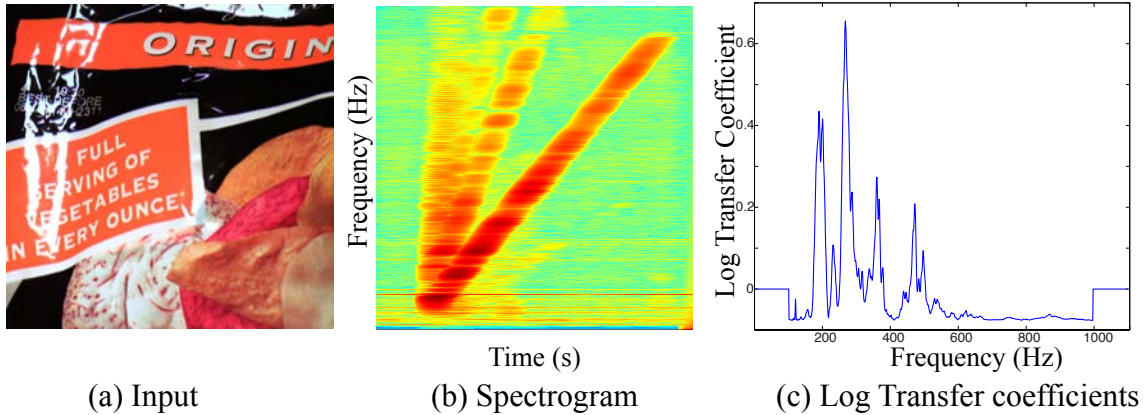
Speaker	fadg0		mccs0		mabw0	
Clip	sa1	sa2	sa1	sa2	sa1	sa2
SSNR w/o Eq.	33.2	29.7	29.8	30.4	19.6	**30.7**
SSNR with Eq.	**35.9**	**33.2**	**30.1**	**31.8**	**20.9**	27.8

Table 4.2: We use a known ramp signal to estimate the transfer coefficients for a bag of chips. We then use these transfer coefficients to equalize new unknown signals recovered from the same bag. a) One frame from a video of the bag of chips. b) The recovered ramp signal we use to compute transfer coefficients. c) The log transfer coefficients (set to 1 outside the range of frequencies in our ramp). The table shows SSNR for six speech examples with and without the equalization. Spectral subtraction is applied again after equalization, as boosting attenuated frequencies tends to boost noise in those frequencies as well. Note that the denoising method SSNR values reported here are different from Table 4.1, as our equalization focuses on accuracy over intelligibility (see text for details).

method recovered audio that was comparable to the laser vibrometer when sampled at the same rate as the video, as measured by the intelligibility metric. However, the LDV required active lighting, and we had to affix a piece of retro-reflective tape on the object for the laser to bounce off the object and go back to the vibrometer. Without the retro-reflective tape, the quality of the vibrometer signal was significantly worse.

4.3.3 Transfer Functions and Equalization

We can use the ramp signal from Section 4.3.1 to characterize the (visual) frequency response of an object in order to improve the quality of signals recovered from new observations of that object. In theory, if we think of the object as a linear system, Wiener deconvolution can be used to estimate the complex-valued transfer function associated with that system, and that transfer function could then be used to deconvolve new observed signals in an optimal way (in the mean squared error sense). In practice however, this approach can be highly susceptible to noise and nonlinear artifacts. Instead, we describe a simpler method that first uses the short time Fourier transform of a training example (the linear ramp) to calculate frequency transfer coefficients at a coarse scale, then equalizes new observed signals using these transfer coefficients.

Our transfer coefficients are derived from the short time power spectra of an input/output pair of signals (like the ones shown in Figure 4-5). Each coefficient corresponds to a frequency in the short time power spectra of the observed training signal, and is computed as a weighted average of that frequency's magnitude over time. The weight at every time is given by the short time power spectrum of the aligned input training signal. Given that our input signal contains only one frequency at a time, this weighting scheme ignores nonlinear artifacts such the frequency doubling seen in Figure 4.2(b).

Once we have our transfer coefficients we can use them to equalize new signals. There are many possible ways to do this. We apply gains to frequencies in the short time power spectra of the new signal, then resynthesize the signal in the time domain. The gain we apply to each frequency is proportional to the inverse of its corresponding transfer coefficient raised to some exponent k.

Figure 4.2 shows the results of applying an equalizer derived from a chip bag to speech sequences recovered from the same object. In the absence of noise, k would be set to 1, but broad spectrum noise compresses the range of the estimated transfer coefficients. Using a larger k can compensate for this. We manually tuned k on one of the female speech examples, then applied the resulting equalizer to all six speech examples. Since this equalization is designed to improve the faithfulness of a recovered signal rather than the intelligibility of speech, we use spectral subtraction for denoising and SSNR to evaluate our results.

Note that calibration and equalization are optional. In particular, all of the results in this chapter outside of Table 4.2 assume no prior knowledge of the recorded object's frequency response.

4.4   Analysis

In this section, we provide an analysis that helps predict when and how well our technique works, and estimate the scale of motions that we are able to recover. At a high level, our method tries to infer some input sound $s(t)$ by observing the motion it causes in a nearby object. Figure 4-3 outlines a series of transformations describing this process. A sound, $s(t)$, defined by fluctuations in air pressure over time, acts on the surface of an object. The object then moves in response to this sound, transforming air pressure into surface displacement. We call this transformation the object response, **A**. The resulting pattern of surface displacement is then recorded with a camera, and our algorithm, **B**, transforms the recorded video into a recovered sound. Intuitively, our ability to recover $s(t)$ will depend on the transformations **A** and **B**. In this section we characterize these transformations to help predict how well the visual microphone will work in new situations.

4.4.1   Object Response (A)

For each object we recorded motion in response to two signals in a calibrated lab setting. The first was a 300Hz pure tone that increased linearly in volume from [0.1-1] Pascals (RMS) (~57 to 95 decibels). This signal was used to characterize the relationship between volume and object motion. To get an accurate measure of volume we calibrated our experimental setup (the loudspeaker, room, and position of the object being tested) using a decibel meter. Figure 4-7 (b) shows the RMS motion of different objects as a function of RMS air pressure in Pascals (at 300Hz). From this graph we see that for most of the objects we tested, the motion appears to be approximately linear in sound pressure. For each object we tested one or more additional frequencies and saw that this

	Measured displacement increase rate (μm/Pa)
crabchips	62.4
greenteabox	41.1
tissue	32.9
foiltogo	20.6
kitkat	21.6
afoil	24.6
rose	10.6
foamcup	8.6
chobani	1.2
teapot	1.1

(a) Displacement coefficients at 300Hz   (b) Motion vs. sound volume   (c) Frequency responses
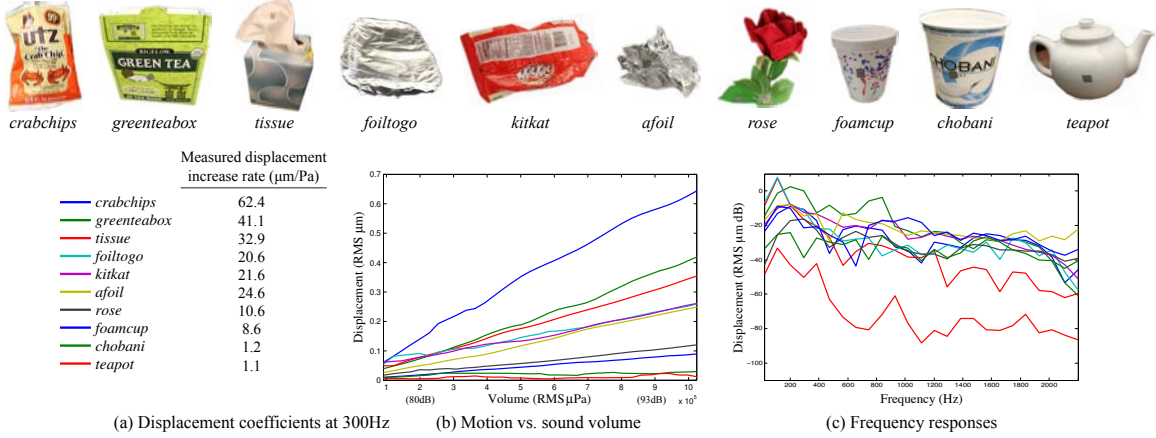
Figure 4-7: Object motion as function of sound volume and frequency, as measured with a laser Doppler vibrometer. Top: the objects we measured, ordered according to their peak displacement at 95 dB, from left (larger motion) to right (smaller motion). (b) The RMS displacement (micrometers) vs RMS sound pressure (Pascals) for the objects being hit by a calibrated 300Hz sine wave linearly increasing in volume from 57 decibels to 95 decibels. Displacements are approximately linear in Pascals, and are all in the order of a micrometer (one thousandths of a millimeter). (c) The frequency responses of these objects (Power dB vs frequency), based on their response to a ramp of frequencies ranging from 20Hz to 2200Hz. Higher frequencies tend to have weaker responses than lower frequencies. Frequency responses are plotted on a dB scale, so the relative attenuation of higher frequencies is quite significant.

relationship remained linear, suggesting that we may model the object response $\mathbf{A}$ as a linear time invariant (LTI) system.

Our second test signal was a ramp signal similar to the one used in Section 4.3.1, with frequencies in the range of 20Hz to 2200Hz. Modeling $\mathbf{A}$ as an LTI system, we used this ramp signal to recover the impulse response of that system. This was done by deconvolving our observed ramp signal (this time recorded by a LDV) by our known input using Wiener deconvolution. Figure 4-7 (c) shows frequency responses derived from our recovered impulse responses[2]. From this graph we see that most objects have a stronger response at lower frequencies than higher frequencies (as expected), but that this trend is not monotonic. This agrees with what we observed in Section 4.3.1.

We can now express the transformation $\mathbf{A}$ in the frequency domain as multiplication of our sound spectrum, $S(\omega)$, by the transfer function $\mathbf{A}(\omega)$, giving us the spectrum of our motion, $D_{mm}(\omega)$:

$$D_{mm}(\omega) \approx \mathbf{A}(\omega)S(\omega) \tag{4.1}$$

The magnitude of the coefficient $\mathbf{A}(\omega)$ for an object corresponds to the slope of its respective volume vs. displacement curve (like the ones shown in Figure 4-7(b)) at frequency ω.

4.4.2   Processing (B)

The relationship between object motion D_{mm} and pixel displacement, D_p, is a straightforward one given by the projection and sampling of a camera. Camera parameters like distance, zoom, viewing

---

[2]The frequency responses shown here have been smoothed to remove noise and intelligibly display all ten on one graph. Responses may also be affected by the responses of the room and speaker.

angle, etc., affect our algorithm's input (the video) by changing the number of pixels that see an object, n_p, the magnification of pixel motion (in $\text{mm}/pixel$), m, and the noise of captured images, σ_N. The relationship between object motion and pixel motion can be expressed as:

$$D_p(\omega) = D_{mm}(\omega) \times m \times \cos(\theta) \tag{4.2}$$

where θ is the viewing angle of our camera relative to the object's surface motion and m is the magnification of our surface in $\frac{\text{mm}}{pixel}$.

Through simulations we also studied the effect of the number of pixels imaging an object (n_p), the amplitude (in pixels) of motion ($D_p(\omega)$), and image noise (given by standard deviation σ_n), on the SNR of our recovered sounds. The results of these simulations (available on our webpage) confirmed the following relationship:

$$\frac{\sigma_S(\omega)}{\sigma_N(\omega)} \propto |D_p(\omega)| \frac{\sqrt{n_p}}{\sigma_n}, \tag{4.3}$$

which shows how the signal to noise ratio increases with motion amplitude and the number of pixels, and decreases with image noise.

To confirm this relationship between SNR and motion amplitude with real data and to test the limits of our technique on different objects, we conducted another calibrated experiment like the one discussed in Section 4.4.1, this time using the visual microphone instead of a laser vibrometer. In this experiment, the camera was placed about 2 meters away from the object being recorded and objects were imaged at 400×480 pixels with a magnification of 17.8 pixels per millimeter. With this setup, we evaluated SNR (dB) as a function of volume (standard decibels). For sufficiently large amplitudes of pixel displacement, our recovered signal becomes approximately linear in volume (Fig. 4-8(a)), confirming the relationship given in Equation 4.3.

To give a sense of the size of motions in our videos, we also estimated the motion, in pixels, for each of the of the corresponding videos using phase-based optical flow [33]. We found these motions to be on the order of one hundredth to one thousandth of a pixel (Fig. 4-8(b)).

4.5   Recovering Sound with Normal Video Cameras using Rolling Shutter

One limitation of the technique presented so far is the need for high speed video. We explore the possibility of recovering audio from video filmed at regular frame rates by taking advantage of the *rolling shutter* common in the CMOS sensors of most cell phones and DSLR cameras [52]. With rolling shutter, sensor pixels are exposed and read out row-by-row sequentially at different times from top to bottom. Compared to uniform global shutters, this design is cheaper to implement and has lower power consumption, but often produces undesirable skewing artifacts in recorded images, especially for photographs of moving objects. Previously, researchers have tried to mitigate the effect of rolling shutter on computer vision problems such as structure-from-motion [51] and video stabilization [34]. Ait-Aider et al. [1] used rolling shutter to estimate the pose and velocity of rigid objects from a single image. We take advantage of rolling shutter to effectively increase the sampling rate of a camera and recover sound frequencies above the camera's frame rate.

crabchips   foamcup   foiltogo

Object	Freq. (Hz)	RMS (px)
crabchips	100	0.029
foiltogo	100	0.010
foamcup	100	0.007
crabchips	300	0.006
foiltogo	300	0.012
foamcup	300	0.005
crabchips	500	0.005
foiltogo	500	0.007
foamcup	500	0.002

(a) SNR vs. Volume   (b) Motion in Pixels
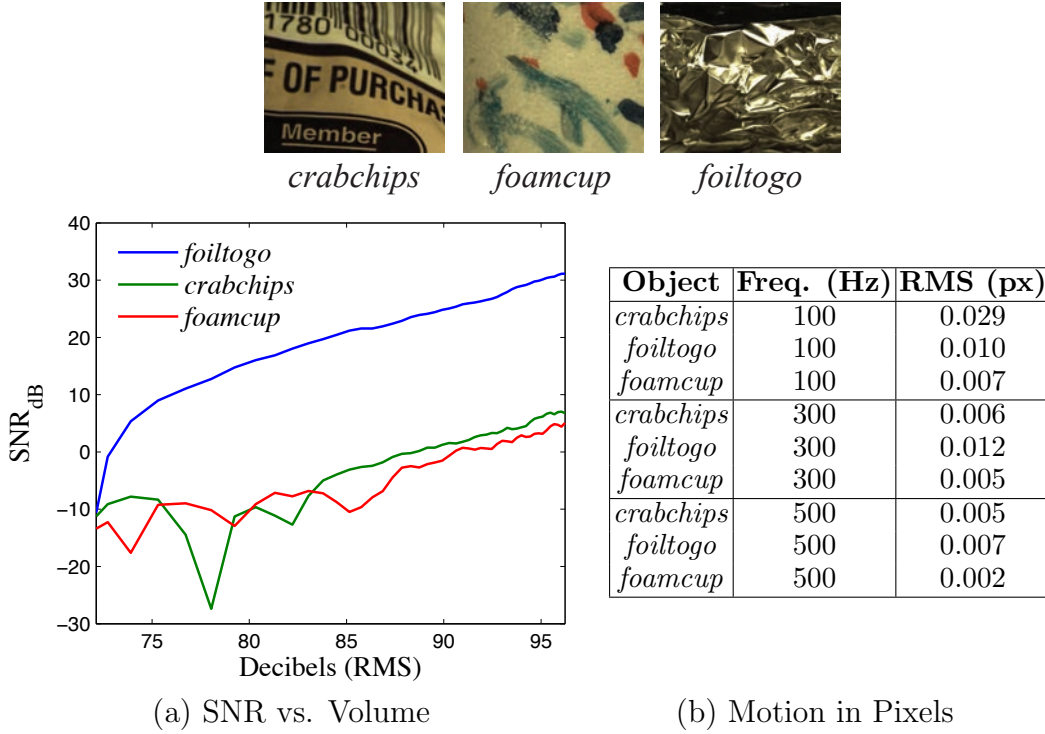
Figure 4-8: The signal-to-noise ratio of sound recovered from video as a function of volume (a), and the absolute motion in pixels (b), for several objects when a sine wave of varying frequency and volume is played at them.

Because each row in a sensor with rolling sensor is captured at different times, we can recover an audio signal for each row, rather than each frame, increasing the sampling rate from the frame rate of the camera to the rate at which rows are recorded (Fig. 4-9). We can fully determine the mapping of the sensor rows to the audio signal by knowing the exposure time of the camera, E, the line delay, d, which is the time between row captures, the frame period T, the time between frame captures, and the frame delay, D (Fig. 4-9). The rolling shutter parameters can be taken from the camera and sensor specs, or computed (for any camera) through a simple calibration process [51], which we also describe on our project web page. We further assume a forward model in which an object, whose image is given by $B(x, y)$, moves with coherent fronto-parallel horizontal motion described by $s(t)$, and that the motion reflects the audio we want to recover, as before. If we assume that the exposure time $E \approx 0$, then the nth frame I_n taken by the camera can be characterized by the equation

$$I_n(x, y) = B(x - \alpha s(nT + yd), y). \tag{4.4}$$

We use this equation to produce a simulation of rolling shutter.

If we assume that the yth row of B has sufficient horizontal texture, we can recover $s(nT + yd)$ using 1D eulerian motion analysis. If the frame delay, the time between the capture of the last row of one frame and the first row of the next frame, is not zero, then there are be times when the camera is not recording anything. This results in missing samples or "gaps" in the audio signal. In Fig. 4-9(b), we show how a triangular wave is recovered from a rolling shutter camera. Each frame contributes eleven samples, one for each row. There are five missing samples, denoted in light gray, between each frame corresponding to the nonnegligible frame delay. To deal with the missing

(a) Rolling shutter in a video

(b) Converted to audio signal
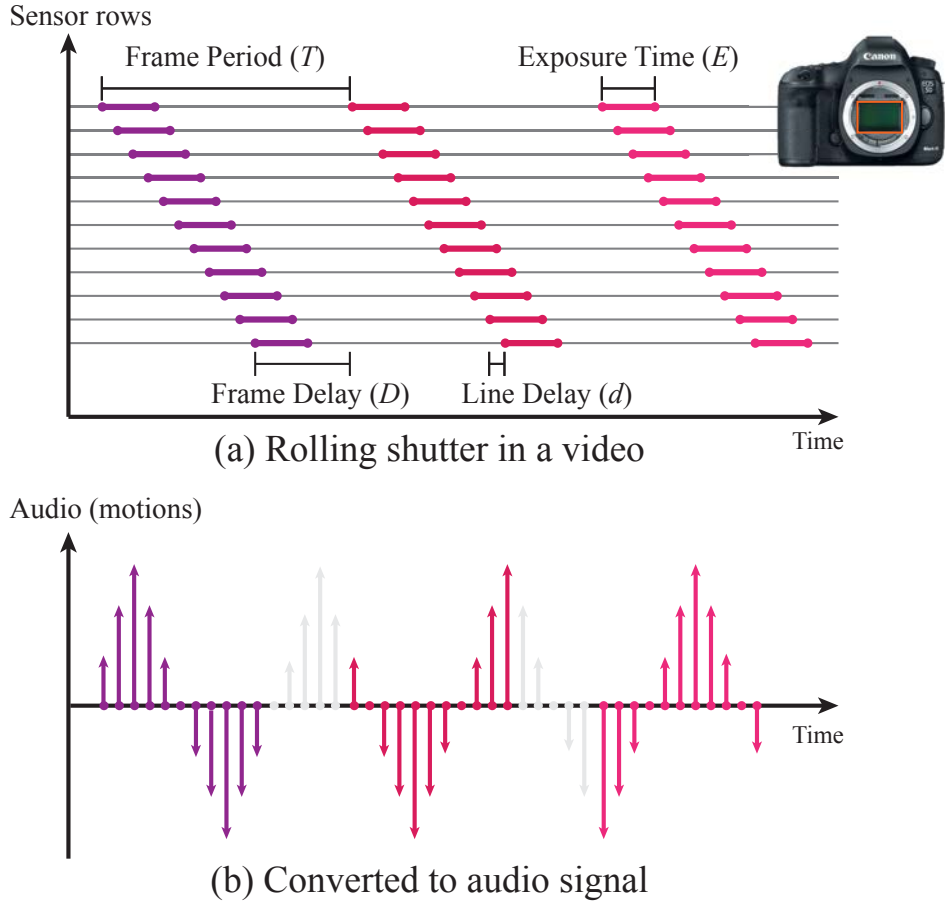
Figure 4-9: Motions from a rolling shutter camera are converted to an audio signal. Each row of the video is captured at a different time. The line delay d is the time between the capture of consecutive rows. The exposure time E is the amount of time the shutter is open for each row, the frame period is the time between the start of each frame's capture and the frame delay is the time between when the last row of a frame and the first row of the next frame are captured. The motion of each row corresponds to a sample in the recovered audio signal (b). Samples that occur during the frame delay period are missing and are denoted in light gray.

samples in our audio signal, we use an audio interpolation technique by Janssen et al. [43].

In practice, the exposure time is not zero and each row is the time average of its position during the exposure. For sinusoidal audio signals of frequency $\omega > \frac{1}{E}$, the recorded row will approximately be to the left of its rest position for half of the exposure and to the right for the other half. Therefore, it will not be well-characterized by a single translation, suggesting that E is a limit on the maximum frequency we can hope to capture with a rolling shutter. Most cameras have minimum exposure times on the order of 0.1 milliseconds (10 kHz).

We show an example result of sound recovered using a normal frame-rate DSLR video in Figure 4-10. We took a video of a bag of candy (Fig. 4-10(a)) near a loudspeaker playing speech, and took a video from a viewpoint orthogonal to the loudspeaker-object axis, so that the motions of the bag due to the loudspeaker would be horizontal and fronto-parallel in the camera's image plane. We used a Pentax K-01 with a 31mm lens. The camera recorded at 60 FPS at a resolution of 1280×720 with an exposure time of $\frac{1}{2000}$ seconds. By measuring the slope of a line, we determined it to have a line delay of 16 μs and a frame delay of 5 milliseconds, so that the effective sampling rate
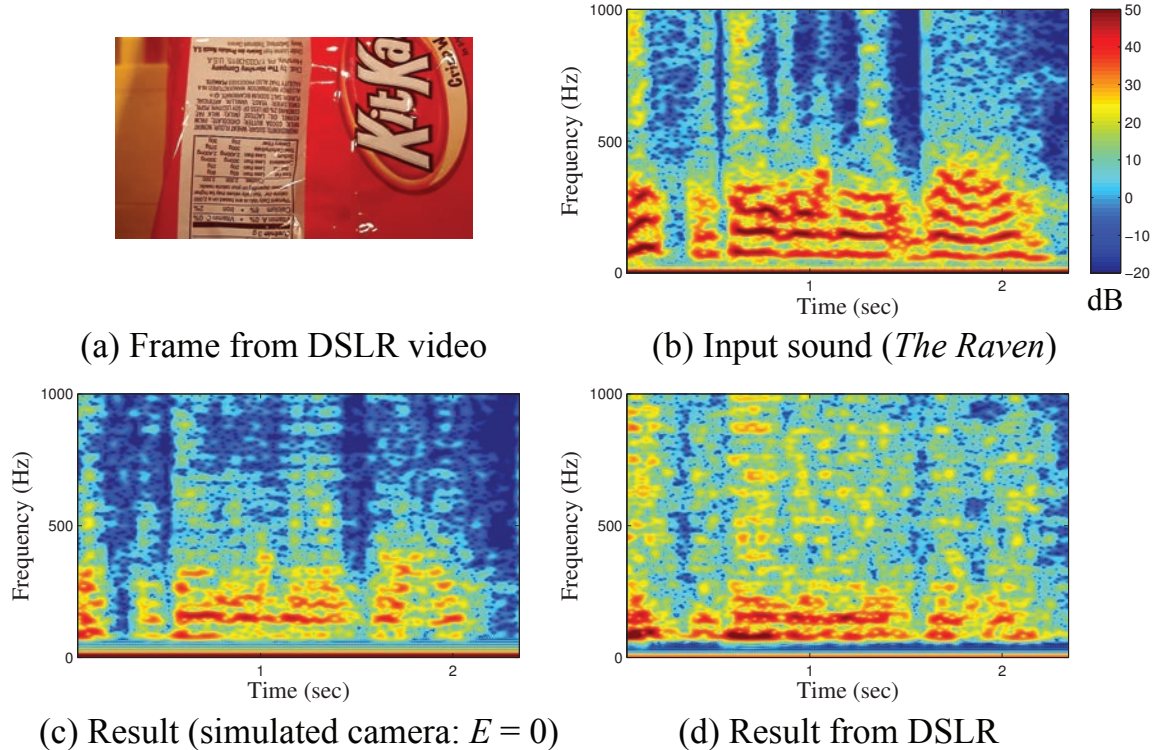
(a) Frame from DSLR video

(b) Input sound (*The Raven*)

(c) Result (simulated camera: $E = 0$)

(d) Result from DSLR

Figure 4-10: Sound recovered from a normal frame-rate video, shot with a standard DSLR camera with rolling shutter. A frame from the DSLR video is shown in (a). James Earl Jones's recitation of "The Raven" by Edgar Allan Poe [56] (spectrogram shown in (b)) is played through a loudspeaker, while an ordinary DSLR camera films a nearby Kit Kat bag. The spectrogram of the signal we manage to recover from the DSLR is shown in (d). In (c) we show the result from our rolling shutter simulation that used parameters similar to the DSLR, except for exposure time (E) that was set to zero.

is $61920Hz$ with 30% of the samples missing. The exposure time caps the maximum recoverable frequency at around 2000Hz. In addition to audio interpolation to recover missing samples, we also denoise the signal with a speech enhancement algorithm and a lowpass filter to remove out-of-range frequencies we cannot recover due to the exposure time. We also performed a simulated experiment with identical camera parameters, except for an instant (zero) exposure time. The recovered audio clips are available online.

4.6 Discussion and Limitations

Information from Unintelligible Sound Many of our examples focus on the intelligibility of recovered sounds. However, there are situations where unintelligible sound can still be informative. For instance, identifying the number and gender of speakers in a room can be useful in some surveillance scenarios even if intelligible speech cannot be recovered. Figure 4-11 shows the results of an experiment where we were able to detect the gender of speakers from unintelligible speech using a standard pitch estimator [25]. On our project web page we show another example where we

(a) Pitch trajectory   (b) Estimated pitch
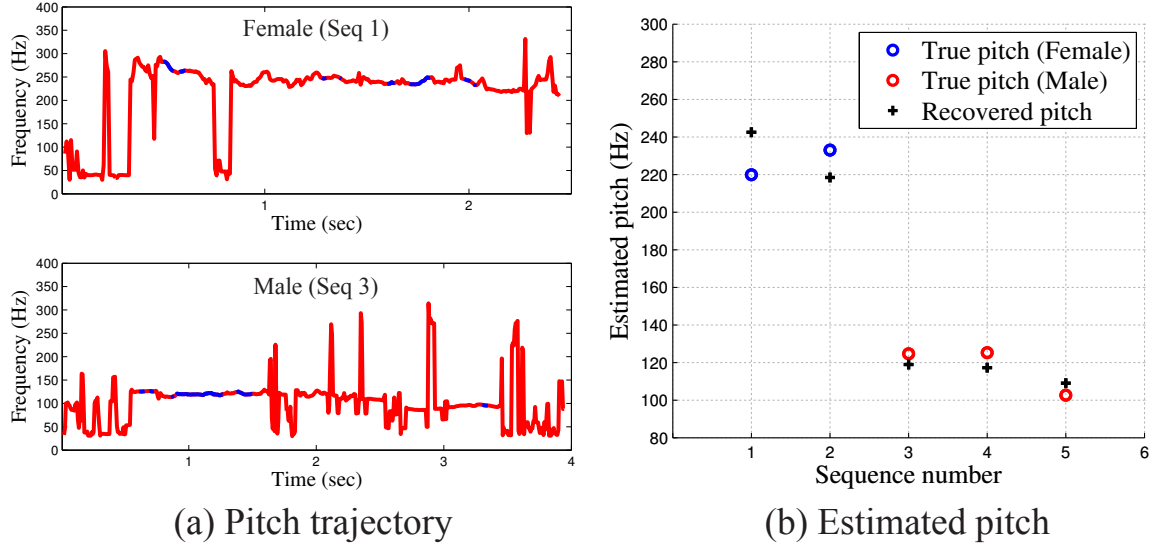
Figure 4-11: Our method can be useful even when recovered speech is unintelligible. In this example, we used five TIMIT speech samples, recovered from a tissue box and a foil container. The recovered speech is difficult to understand, but using a standard pitch estimator [25] we are able to recover the pitch of the speaker's voice (b). In (a) we show the estimated pitch trajectory for two recovered speech samples (female above, male below). Blue segments indicate high confidence in the estimation (see [25] for details).

recover music well enough for some listeners to recognize the song, though the lyrics themselves are unintelligible in the recovered sound.

Visualizing Vibration Modes Because we are recovering sound from a video, we get a spatial measurement of the audio signal at many points on the filmed object rather than a single point like a laser microphone. We can use this spatial measurement to recover the vibration modes of an object. This can be a powerful tool for structural analysis, where general deformations of an object are often expressed as superpositions of the object's vibration modes. As with sound recovery from surface vibrations, most existing techniques for recovering mode shapes are active. Stanbridge and Ewins [69], for instance, scan a laser vibrometer in a raster pattern across a surface. Alternatively, holographic interferometry works by first recording a hologram of an object at rest, then projecting this hologram back onto the object so that surface deformations result in predictable interference patterns [58, 44]. Like us, Chen et al. [15] propose recovering mode shapes from a high-speed video, but they only look at the specific case of a beam vibrating in response to being struck by a hammer.

Vibration modes are characterized by motion where all parts of an object vibrate with the same temporal frequency, the modal frequency, with a fixed phase relation between different parts of the object. We can find the modal frequencies by looking for peaks in the spectra of our local motion signals. At one of these peaks, we will have a Fourier coefficient for every spatial location in the image. These Fourier coefficients give the vibration mode shape with amplitude corresponding to the amount of motion and phase corresponding to fixed phase relation between points. In Figure 4-12, we map amplitude to intensity and phase to hue for two vibration modes of a drum head. These recovered vibration modes (Fig. 4-12(b)) closely correspond to the theoretically-derived modal shapes (Fig. 4-12(c)).

Limitations   Other than sampling rate, our technique is mostly limited by the magnification of the lens. The SNR of audio recovered by our technique is proportional to the motion amplitude in pixels and the number of pixels that cover the object (Eq. 4.3), both of which increase as the magnification increases and decrease with object distance. As a result, to recover intelligible sound from far away objects, we may need a powerful zoom lens. The experiment in Figure 4-2 used a 400mm lens to recover sound from a distance of 3-4 meters. Recovery from much larger distances may require expensive optics with large focal lengths.

4.7   Conclusion

We have shown that the vibrations of many everyday objects in response to sound can be extracted from high speed videos and used to recover audio, turning those objects into "visual microphones". We integrate local, minute motion signals across the surface of an object to compute a single motion signal that captures vibrations of the object in response to sound over time. We then denoise this motion signal using speech enhancement and other techniques to produce a recovered audio signal. Through our experiments, we found that light and rigid objects make especially good visual microphones. We believe that using video cameras to recover and analyze sound-related vibrations in different objects will open up interesting new research and applications. Our videos, results and supplementary material are available on the project web page: `http://people.csail.mit.edu/mrub/VisualMic/`.

(a) Example frame from input
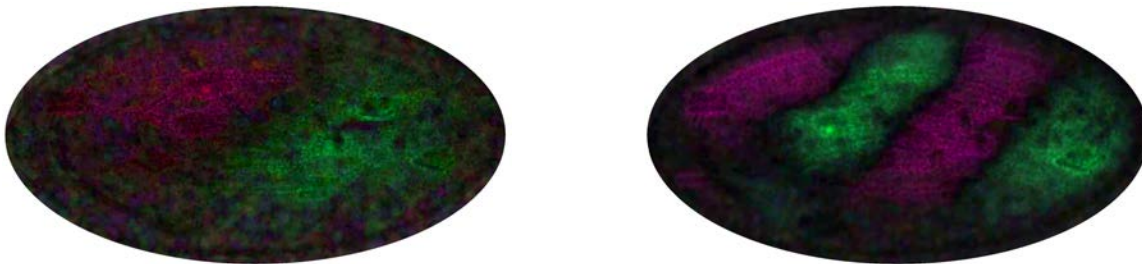
(b) Our recovered mode shapes
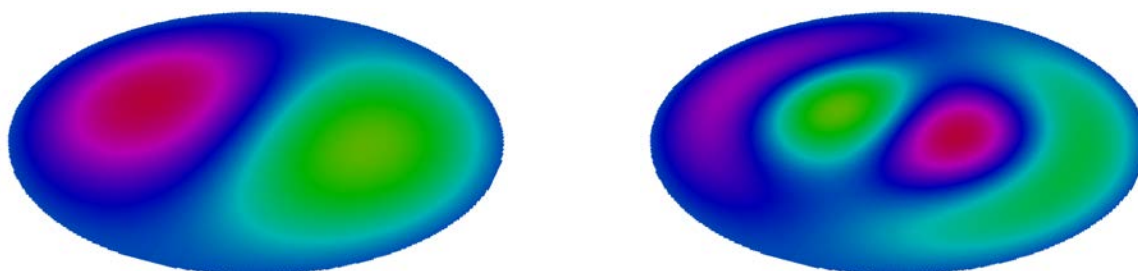
(c) Theoretically-derived mode shapes

Figure 4-12: Recovered mode shapes (b) from a video of a circular latex membrane excited by a chirp playing from a nearby audio source (a). Our recovered mode shapes (b) are similar to the theoretically-derived mode shapes (c). For the modes shown in (b), the phase of surface motion across the membrane is mapped to hue, while the amplitude of vibrations across the surface is mapped to saturation and brightness.

5

Visual Vibrometry

The estimation of material properties is important for scene understanding, with many applications in vision, robotics, and structural engineering. This chapter connects fundamentals of vibration mechanics with computer vision techniques in order to infer material properties from small, often imperceptible motion in video. Objects tend to vibrate in a set of preferred modes. The frequencies of these modes depend on the structure and material properties of an object. We show that by extracting these frequencies from video of a vibrating object, we can often make inferences about that object's material properties. We demonstrate our approach by estimating material properties for a variety of objects by observing their motion in high-speed and regular framerate video.

5.1   Introduction

Understanding a scene involves more than just recognizing object categories or 3D shape. Material properties like density, stiffness, and damping can play an important role in applications that involve assessing or interacting with the world. In the field of non-destructive testing (NDT), these properties are often recovered by analyzing the vibrations of an object. Typically, these vibrations are measured with contact sensors or expensive laser vibrometers, which limit sampling to only a small number of discrete points on an object's surface. We propose an alternative approach to vibration analysis that instead uses cameras to measure vibrations and make inferences about the object's underlying physical properties.

Objects tend to vibrate in a set of preferred modes. These vibrations occur in most materials, but often happen at scales and frequencies outside the range of human visual perception. Bells, for
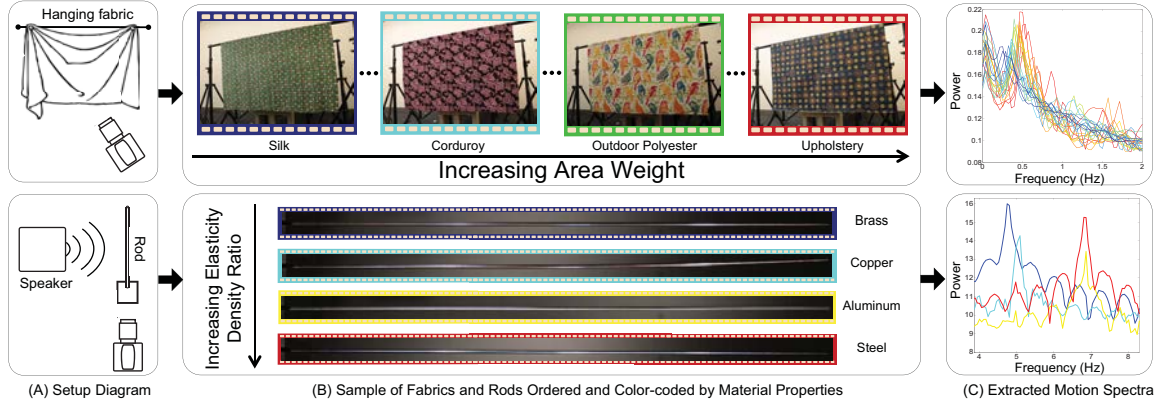
Figure 5-1: We present a method for estimating material properties of an object by examining small motions in video. (A) We record video of different fabrics and clamped rods exposed to small forces such as sound or natural air currents in a room. (B) We show fabrics (top) color-coded and ordered by area weight, and rods (bottom) similarly ordered by their ratio of elastic modulus to density. (C) Local motion signals are extracted from captured videos and used to compute a temporal power spectrum for each object. These motion spectra contain information that is predictive of each object's material properties. For instance, observe the trends in the spectra for fabrics and rods as they increase in area weight and elasticity/density, resp (blue to red). By examining these spectra, we can make inferences about the material properties of objects.

instance, vibrate at distinct audible frequencies when struck. We cannot usually see these vibrations because their amplitudes are too small and their frequencies are too high - but we hear them. Intuitively we know that large bells tend to sound deeper than small ones, or that a bell made of wood will sound muted compared to one made of silver. This is because an object's modes of vibration are closely and predictably related to its material properties. We show how this connection can be used to learn about the material properties of an object by analyzing its vibrations in video.

In this chapter we first review established theory on modal vibrations, and connect this theory to features that can be extracted from video. Our features provide an ambiguous combination of structural and material information that can be used directly to make relative measurements, or in combination with additional information to make absolute measurements. We present three experiements showing how these features can be used to estimate structural or material properties given some prior information about an object. The first experiment, using a set of clamped rods, is designed to resemble typical engineering applications, and shows how our features can be used to resolve material properties in situations where geometry can be precisely measured. Our second experiment, using a set of hanging fabrics, explores the idea of *learning* the relationship between our features and material properties when objects naturally occur with similar geometry – demonstrating the potential for successful data-driven approaches to material estimation. Finally, our third experiment, using a set of wine glasses, is a simple demonstration of how our technique can be used to estimate relative properties even without a prior on geometry by comparing the resonance of objects within a group, or of a single object over time.

5.2 Related Work

This chapter connects related works in computer vision, graphics, and civil engineering through common theory and uses these connections to extend existing methods.

5.2.1 Traditional Vibration Analysis

Vibration analysis is an established tool used in a variety of engineering disciplines. Especially related to this chapter is work in the field of NDT, where techniques based on ultrasound are common. However, these techniques often require direct contact with the object being measured [66]. Non-contact vibration measurement is usually accomplished with a laser Doppler vibrometer, which computes the velocity of a surface by measuring the Doppler shift of a reflected laser beam [28]. Laser vibrometers have been used to non-destructively examine valuable paintings [12, 19], detect land mines [36, 2], test fruit [62], find defects in composite materials [11, 14, 29], and even test vibration modes of small structures [69]. However, laser vibrometers are active in nature and generally only measure the vibration of a single surface point. While scanning or multi-beam laser vibrometers exist [69, 2], they are still active and can be prohibitively expensive - costing several times more than even the most expensive high-speed camera used in this work.

5.2.2 Material Property Estimation from Video

Previous work in computer vision has focused on estimating material properties from static images [65, 49, 39, 32]. In contrast, our goal is to use video in order to estimate material properties that characterize the motion of an object.

A number of works in vision and graphics have been used to estimate properties of fabric, which we also do in this chapter. Early approaches worked by fitting the parameters of cloth-specific models to video and depth information [5, 45]. Bouman et al. [7] adopted a learning approach that allowed them to estimate material properties from a video of fabric moving under wind forces. As with our experiments in Section 5.5, they estimate material properties directly from video statistics using a regression strategy. Their work found the local autocorrelation of optical flow to be especially predictive of a fabric's area weight and stiffness, suggesting a possible connection between material properties and the spectrum of an object's motion in video. Our work uses established vibration theory to explain this connection and improve on the features used in their chapter.

5.3 Method

Our task is to estimate the material properties of objects using the motion spectra described in Section 3.4 of Chapter 3. Our method has three components that vary depending on the object being observed.

5.3.1 Excitation

An object must move in order for us to observe its vibration modes. Some very deformable objects, like hanging fabric, may move enough with natural air currents for no additional forces to be necessary. For more rigid objects, like wine glasses or metal rods, we use sound to induce motion. The excitation should be strong enough to create a recoverable motion signal, and should contain energy at each of the objects resonant frequencies. Sound has been used for this purpose previously in NDT [14, 19, 11, 29, 36].

5.3.2 Video Capture

To estimate an object's resonant frequencies we need to record at a high enough framerate to place these frequencies under the Nyquist limit. We should also ensure that videos capture enough periods at each mode frequency to sufficiently localize corresponding spikes in the Fourier domain. For objects with high resonant frequencies this can be accomplished with short clips of high speed video. Objects with low resonant frequencies (like hanging fabric) can be captured with longer, lower-framerate video.

5.3.3 Inference

The motion spectrum of an object provides us with an ambiguous combination of structural and material information. In some cases, this combination is directly useful (e.g. tuning an instrument or identifying a source of unwanted noise). In others, it provides constraints from which we can infer more specific properties. This inference depends on the type of information available and the properties being inferred. We explore three different strategies in this chapter, each with different strengths and weaknesses. The first strategy is to use measured or known geometry to directly estimate material properties. This strategy can be very precise, but requires additional measurement (usually through some means other than video). The second strategy alleviates the need for careful measurement by learning the relationship between recovered motion spectra and material properties from training data. This approach is convenient, but depends on the availability and accuracy of a learned prior. Finally, the third strategy is to sidestep the need for any prior on geometry by simply comparing spectra to detect changes over time or variations within a group objects. This strategy is simple, and a promising approach for applications in structural health monitoring, where any significant change in resonance may indicate a problem, and reference spectra are often available.

5.4 Estimating Properties of Materials with Known Geometry: Rods

In our first set of experiments we estimate the material properties or geometry of various rods by extracting their resonant frequencies from video. The simple geometry of a clamped rod makes it easy to solve for vibration modes analytically as a function of length, diameter, density, and an elastic modulus. While length, diameter, and density can all be measured with a simple ruler and scale, the elastic modulus is usually measured with a tensile test, which requires expensive equipment

and usually damages the object being tested. In these experiments we first show how this elastic modulus can instead be measured with a speaker and high-speed camera. Just as our recovered spectra can be used to resolve unknown material properties (i.e. elastisity) given known geometry, we also show that they can be used to resolve unknown geometry given known material properties. This second case could be used to resolve an ambiguity of scale when a filmed object is made of a known material.

Setup

We filmed rods made from four different metals - steel, aluminum, copper, and brass. Rods were clamped to a block of concrete next to a loudspeaker (see Figure 5-2), and each rod was tested twice: once clamped to a length of 15 inches and once clamped to a length of 22 inches. In Section 5.4.3 we compare material properties derived from our observations to estimates provided by the manufacturer. Recovered frequencies and mode shapes for all of the rods, as well as birch and fiberglass rods with unreported material properties, can be found in the provided supplemental material.

Image of setup
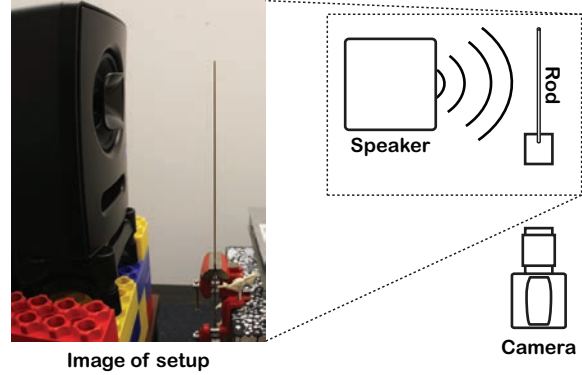
Figure 5-2: Rods were clamped to a concrete block next to a loudspeaker (shown left) and filmed with a high-speed camera. By analyzing small motions in the recorded video, we are able to find resonant frequencies of the rods and use them to estimate material properties.

Excitation

The excitation signal should be broad spectrum to ensure that multiple rod modes are activated. In [15, 16] this is accomplished by striking the beam with a hammer. To avoid damage to the rod, we instead use sound - specifically, a linear ramp of frequencies from 15 Hz to 2250 Hz played through the loudspeaker at each rod. We found that modes at frequencies below 15 Hz were still activated by this signal, possibly due to the presence of some signal components below 15 Hz and the relatively high sensitivity of lower modes.

Video Capture

Rods were filmed with a Phantom high-speed camera. Given the lengths and thicknesses of our rods, a conservative estimate of material properties put the fourth mode of each rod well below 1250 Hz. We filmed at 2500 fps to ensure a sampling rate high enough to recover this mode for each rod.

5.4.1   Finding Resonant Frequencies

The vibrations of clamped rods are well studied [64]. A rod's fundamental frequency ω_1 (corresponding to its first mode) is related to material properties by the equation:

$$\omega_1 = 0.1399 \frac{d}{L^2} \sqrt{\frac{E}{\rho}} \tag{5.1}$$
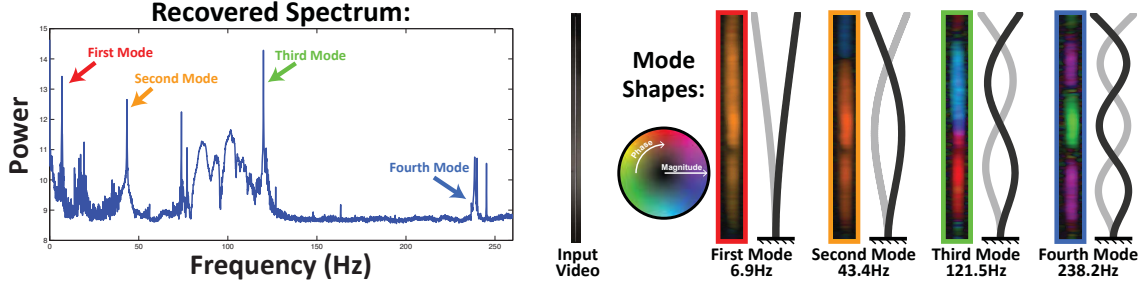
Figure 5-3: Finding vibration modes of a clamped brass rod: (Left) We recover a motion spectrum from 2.5 kHz video of a 22 inch clamped aluminum rod. Resonant frequencies are labeled. To distinguish resonant frequencies from other spikes in the spectrum, we look for energy at frequencies with ratios derived from the known geometry of the rod. (Middle) A sample frame from the 80×2016 pixel input video. (Right) Visualizations of the first four recovered mode shapes are shown next to the corresponding shapes predicted by theory.

where d is the diameter of the rod, L is its length, ρ is its density and E is its Young's modulus (measuring elasticity). Given the length and width of a rod, the task of estimating $\sqrt{\frac{E}{\rho}}$ can then be reduced to finding its fundamental frequency. Under ideal conditions this would amount to finding the largest spike in the rod's motion spectrum. However, real spectra tend to also contain spikes at non-modal frequencies (see Figure 5-3). To distinguish these from the rod's resonant frequencies we recall from Chapter 3 that changes in material properties only scale the modal frequencies - leaving their ratios constant. In clamped rods, ratios for the first four resonant frequencies can be found analytically[1], and are given by:

$$\omega_i = \eta_i \omega_1,$$
$$\eta_1 = 1, \quad \eta_2 = 6.27, \quad \eta_3 = 17.55, \quad \eta_4 = 34.39 \tag{5.2}$$

where again ω_i is the resonant frequency for the ith mode. To distinguish modal frequencies from other spikes, we look for energy in the recovered spectra that occurs in the ratios given by Equation 5.2. We assume that the probability of a rod mode at a given frequency is proportional to the power at that frequency. Given the recovered spectrum S, we then have:

$$P\left(\omega = \omega_1 \mid S\right) \propto \prod_{i=1}^{4} S(\omega \eta_i). \tag{5.3}$$

Using Equation 5.3, we can find the most likely fundamental frequency using a simple voting scheme. In practice, since we operate in the discrete Fourier domain, we achieve higher precision at the fundamental by using the relations of Equation 5.2 to vote for the fourth resonant frequency.

5.4.2 Estimating Damping

As discussed in Chapter 3, the damping of a mode appears in an object's transfer function as convolution with a Lorentzian distribution that depends on the damping ratio ξ. To find ξ, we fit a Lorentzian distribution around the modes identified by out voting scheme. Automatically fitting

---

[1]By solving the continuous analog to Equation 3.7 [64]
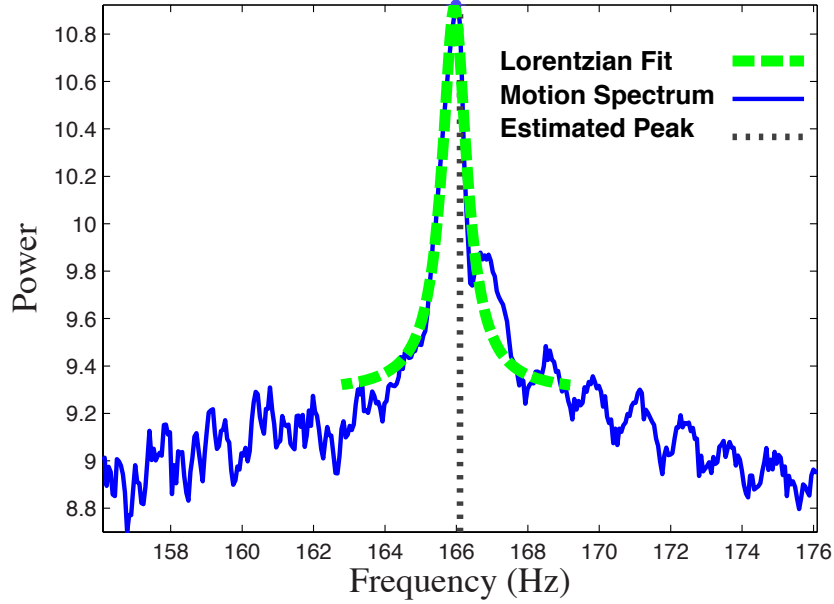
Figure 5-4: Our damping selection interface, inspired by the standard proceedure defined in [3], presents users with a view of the recovered motion spectra around a predicted rod frequency and asks them to click and drag over the spike region. A Lorentzian is fit to the selected region and presented for the user to evaluate.

these distributions using a fixed range of frequencies around identified modes produces poor fits, as different damping values affect different ranges of frequencies. We address this using a manual selection strategy, inspired by the procedures set by the ASTM for measuring the material damping or loss factor in materials [3]. Our selection interface is similar to the one used for mode selection in [23], where users are presented with the motion spectrum of a video and asked to click on peaks. However, our selection process uses the frequencies predicted with our voting scheme as an initial estimate, zooming in on each predicted frequency one at a time. Users are then asked to select the range of frequencies between the resonant peak and noise floor using their mouse. A Lorentzian is immediately fit to the selected region using non-linear least squares, and presented for the user to evaluate (Figure 5-4). If the fit looks good, the user proceeds to the next mode. If the fit does not look good, they press a button to indicate that the damping on the corresponding mode cannot be accurately estimated, a result often caused by mode masking. One of the parameters of the Lorentzian distribution is the full width at half maximum $\Delta\omega$, which can be used to calculate the modal damping ratio as $\xi_i = \frac{\Delta\omega}{2\omega_{di}}$.

5.4.3 Results

Young's Modulus

Under fixed but unknown geometry, the recovered fundamental frequencies provide a value proportional to $\sqrt{E/\rho}$. From this we can use Equation 5.1 with lengths and densities measured by a scale and measuring tape to compute the modulus of each rod. Figure 5-5a shows a plot of Young's
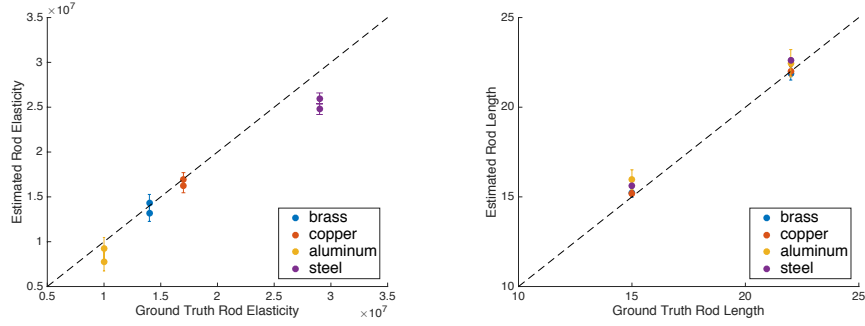
(a) Elastisity Estimation (force per squared inch)

(b) Length Estimation (inches)

Figure 5-5: Estimating the elastic modulus and length of clamped rods: (a) Young's moduli (force per squared inch) reported by the manufacturer plotted against values estimated using our technique. Estimated values are close to those reported by the manufacturer, with the largest discrepancies happening in 15 inch rods made of aluminum and steel. (b) The length (inches) of each rod measured to the base of the clamp plotted against values estimated using our technique.

moduli (in force per squared inch) reported by the manufacturer against the values estimated using our technique. Percent errors are given in Table 5.1.

Length

By massaging Equation 5.1, we see that the length of a rod can be estimated as a function of the fundamental frequency, rod diameter, elastisity, and density:

$$L = \sqrt{0.1399 \frac{d\eta_i}{\omega_i} \sqrt{\frac{E}{\rho}}} \tag{5.4}$$

which we can use to estimate length given our observed resonant frequencies and the Young's modulus reported by the manufacturer. Figure 5-5b shows a plot of the measured length (in inches) of each rod veruses the value estimated in this manor. Percent errors are given in Table 5.2.

Error

Error bars in Figure 5-5 are calculated for each Young's modulus and length estimate by propagating error bounds for each measured variable. Error propagation was done assuming independent variables [47]. Given a function $F(a, b, c, ...)$, the equation for the error σ_F depending on the errors $\sigma_a, \sigma_b, \sigma_c...$ is given as:

$$\sigma_F = \sqrt{\left(\frac{\partial F}{\partial a}\right)^2 \sigma_a^2 + \left(\frac{\partial F}{\partial b}\right)^2 \sigma_b^2 + \left(\frac{\partial F}{\partial c}\right)^2 \sigma_c^2 + ...} \tag{5.5}$$

Young's modulus estimates were calculated by propagating error from length, diameter, and density. Length estimates were calculated by propagating error from only diameter and density. The calcu-

lated errors for length estimation are smaller than expected due to the lack of reported tolerances on Young's modulus values. Refer to the Appendix for further information on error approximation.

Mode Shapes

For each rod, we can further verify recovered modes by visualizing the recovered shapes corresponding to estimated resonant frequencies (see Figure 5-3). Mode shapes are sometimes masked by vibrations from other parts of the experimental setup - for instance, vibrations of the camera or the frequency of lights powered by AC current. However, it is unlikely that a majority of resonant frequencies will be masked in any single single rod. In practice we see the predicted shapes of multiple modes in the data recovered for each rod. All 48 mode shapes recovered in our experiments can be found in the provided supplemental material.

Damping

Material damping properties are not as well characterized as other mechanical properties, such as Young's modulus for stiffness. This is, in part, because it is very difficult to control for external sources of damping. Additionally, damping can vary across the different modes of a given system. As a result, manufacturers do not typically report damping ratios. However, some general trends are accepted for different materials. For example, metals tend to have very low material damping compared rubber. In addition to our metal rods, for which the manufacturer reported Young's moduli, we also obtained a rod made of wood (birch). While material property values for wood are highly variable (likely the reason no Young's modulus was provided), wood is generally accepted to have higher damping than most metals, and quantitative studies of different vibrating systems (e.g. [20]) have supported this claim. Figure 5-6 shows our damping estimates of different rod modes as a function of frequency (damping was evaluated at each unmasked rod mode). As expected, we see that the wooden rod has the highest damping ratio at every mode.

Discussion

Our estimated moduli are close to, but consistently under, the reported values (Figure 5-5a and Table 5.1). One possible explanation for this is an incorrect estimate of where the clamp grabbed each rod in our setup. Similarly, Figure 5-5b and Table 5.2 show that our length estimates are close to, and correlated with, but consistantly longer than our measured values - which could be explained by the same source of measurement error.

Our damping results show that our wooden rod has consistantly higher damping than the metal rods, which is expected given their material differences. However, the relative damping ratios of our metal rods are less consistent across different modes. These results suggest that we are able to distinguish between materials with significantly different levels of damping (such as metal and wood), though additional experiments would be needed to better understand how well we distinguish damping between more similar materials (e.g. among the different metals).

Our Young's modulus and length results suggest both a strength and weakness of an approach that pairs recovered motion spectra with careful measurement for inference - high precision that is very sensitive to accurate modeling of the structure being tested. Our next experiments address

% Error	Brass	Copper	Aluminum	Steel
22 inches	2.13	-0.40	-7.82	-10.40
15 inches	-5.98	-4.69	-22.13	-14.53

Table 5.1: Percent errors in estimating the Young's modulus (force per squared inch) for each rod.

% Error	Brass	Copper	Aluminum	Steel
22 inches	-0.52	0.10	2.06	2.78
15 inches	1.55	1.21	6.45	4.00

Table 5.2: Percent errors in estimating the length (inches) for each rod.

this issue by instead attempting to learn the relationship between material properties and resonant frequencies.

5.5 Learning Properties of Materials with Unknown Geometry: Fabrics

The inference described in Section 5.4.1 relies on knowing the ratios between resonant frequencies, η_i. These ratios are simple to derive in clamped rods, but can be prohibitively difficult to compute in more general structures. As a result, many applications of vibrometry are limited to simple geometries that can be precisely measured (as is the case with rods) or man-made structures (airplanes, buildings, cars, etc) with resonant frequencies that can be derived from detailed CAD models through FEM analysis. The ubiquity and passive nature of video offers the potential to address this limitation by providing sufficient data to learn relationships between motion spectra and the material properties of objects. In this section, we explore that potential by using a learning approach to estimate the material properties of hanging fabrics from video. We show that our technique outperforms a previous video-based fabric property estimation method, even when trained using data captured from different viewpoints or using different excitation forces.

A number of metrics exist to describe the material properties of fabrics. These properties can be measured using setups such as the Kawabata system [46, 78]. In the work of Bouman, et al. [7], a dataset of 30 fabrics along with ground truth measurements of stiffness and area weight were collected. We extend this dataset to predict the material properties from videos exhibiting small motions that are often invisible to the naked eye, in contrast to [7] that relied on much larger motions produced by fans.

Setup

Each fabric specimen from [7] (width 43.5 to 44.5 inches across) was loosely draped over a bar and hung a length of 29.25 to 32.25 inches from the top of the bar. Notice that although the geometry was kept relatively constant, these measurements vary a great deal compared to those used in Section 5.4.
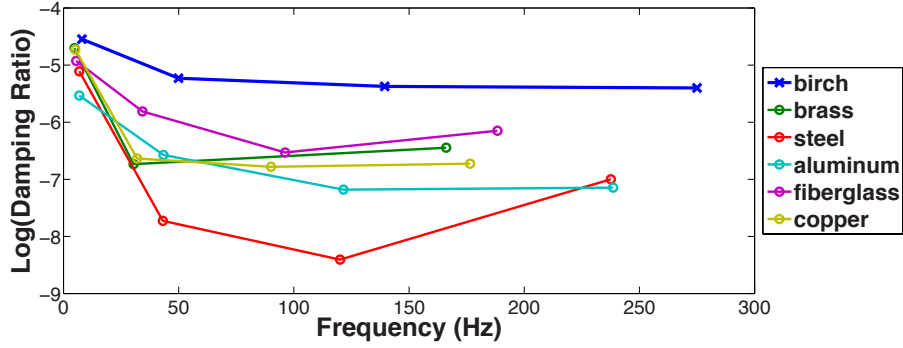
Figure 5-6: The damping ratio estimated from the recovered motion spectra for each automatically identified resonant frequency. While reported damping ratios for different materials vary greatly, general trends are recognized. Our recovered rod damping ratios show recognized trends of higher damping in wood than in metals [20], and higher damping in lower fundamental modes due to their high amplitude [4].

(a) Setup Diagram   (b) Setup Image   (c) Point Grey Camera (grayscale)   (d) Canon 6D SLR Camera
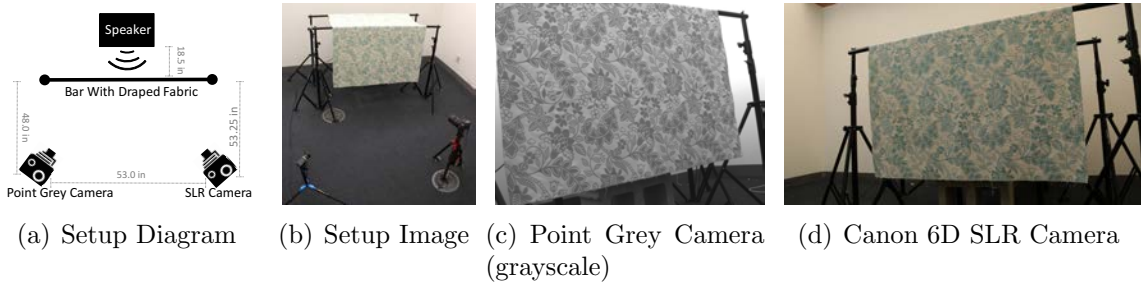
Figure 5-7: Videos were recorded of the fabric moving from (c) a grayscale Point Grey camera (800×600 pixel resolution) at 60 fps and (d) an RGB SLR Camera (Canon 6D, 1920×1080 pixel resolution) at 30 fps. The experimental layout (a,b) consisted of the two cameras observing the fabric from different points of view.

Excitation

Ambient Forces: Even without an explicit excitation force applied, hanging fabric is almost always moving. Ambient forces, such as air currents in the room or small vibrations in the building induce small motions in fabric. Figure 5-8a shows a space-time slice of a fabric moving due to ambient forces in the room.

Sound : As an alternative, we also tested sound as a source of excitation. Sound was used to provide a small, controlled "kick" to the hanging fabric. We excited each fabric with a one second, logarithmic frequency ramp from 15 to 100 Hz. Figure 5-8b shows a space-time slice of a fabric moving due to this "kick."

Video Capture

Hanging fabrics tend to have dominant vibration modes at relatively low frequencies. For this reason, we can use standard, commercial cameras operating at much lower framerates than were required for analyzing rods. Each combination of fabric and excitation force was captured simultaneously by two

(A) Location of Slice from Video Frame   (B) Ambient Force Excitation

(C) Acoustic Wave (Sound) Excitation
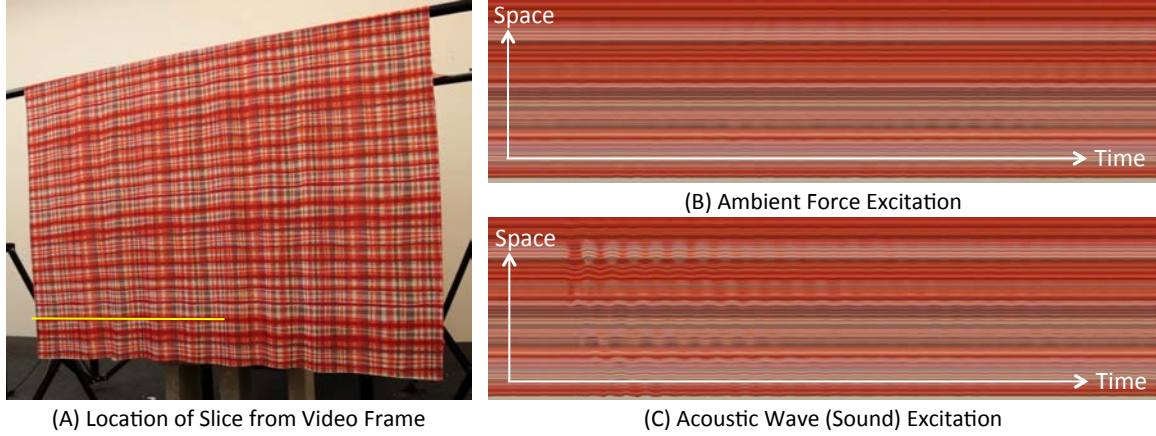
Figure 5-8: Videos of fabric excited by two different types of force were recorded. Here we see space × time slices from minute long videos of a fabric responding ambient forces (b) and sound (c). The motion is especially subtle in (b), but still encodes predictive information about the fabric's material properties.

cameras: an RGB SLR camera (Canon 6D, 1920×1080 pixel resolution) at 30 fps and a grayscale Point Grey camera (800×600 pixel resolution) at 60 fps. The cameras recorded different viewpoints (see Figure 5-7), which we use to test the invariance of our trained models to changes in perspective. Each video is approximately one-minute long and can be found, along with the corresponding fabric measurements (width and height), on our project website.

5.5.1 Property Estimation

Feature Extraction

Due to their comparatively high damping, fabric motion spectra do not contain the same clean, narrow peaks seen in rods. Damping causes the bandwidth around resonant frequencies to overlap, making it difficult to identify individual modes (see Figure 5-1). As a result, the inference strategies we used for rods will not work. However, the distribution of energy in the motion spectrum is still predictive of the fabric's material properties. For example, note how in Figure 5-1 the location of a fabric's resonant band shifts to the right with increasing area weight. Our approach is to use the motion spectra directly as features, and learn a regression model that maps these features to material properties.

As feature vectors we chose $N = 150$ uniform samples of the normalized motion spectra from 0 to 15 Hz. To reduce the effect of noise, we smooth the recovered motion spectra using a Gaussian with standard deviation $\frac{15}{2(N-1)} Hz$.

Inference

We learn regression models that map the motion spectra to the log of ground truth stiffness or area weight measurements provided in [7]. Models are fit to the log of measurements in order to directly compare with results presented in [7]. Fitting a regression model directly to the processed motion spectra results in overfitting. Instead, we have explored two standard regression methods

that reduce the dimensionality of the data: Principal Components Regression (PCR) and Partial Least Squares Regression (PLSR). Both methods perform comparably, suggesting that the power of our algorithm is in the features, the recovered motion spectra, rather than the regression model. In this chapter, we show results of the trained PLSR model. Additional results from PCR can be found in the supplemental material.

Cross Validation

Due to the small number of fabrics in the dataset, we use a leave-one-out method for training and testing. Precisely, all data corresponding to a fabric are removed from training of the regression parameters when predicting the material properties of that fabric. Using this method, we estimate the performance of our model on predicting the material properties of a previously unseen fabric. Performance was evaluated using a varying number of PLSR components. From this evaluation we chose a reduced number of PLSR dimensions, M, that is both robust and results in high accuracy for both material properties. For results presented in this chapter, we used $M = 2$ and $M = 5$ for the ambient force model and acoustic model respectively. Refer to Figure 5-9.

Testing Invariance

We saw in Chapter 3 that our motion power spectra should be invariant to changes in viewpoint. Here we test this invariance by training and testing on videos captured under different conditions. In total we have four conditions for fabrics: ambient (A) and acoustic (S) excitations, each captured from two different viewpoints (the left point grey (L) and right SLR (R) cameras). We used the same leave-one-out validation strategy when training and testing data were taken from different conditions.

5.5.2   Results

Our estimates of material properties are well correlated with the log of ground truth measurements (refer to Table 5.4). In all cases, even when testing under conditions with different viewpoints and excitation forces from the training data, our estimates outperform previous video-based fabric measurements [7] in predicting both stiffness and area weight.

Figure 5-10 contains correlation plots corresponding to the conditions presented in Table 5.4. These plots compare our algorithm's predicted measurements of stiffness and area weight to the log of ground truth measurements when models were trained and tested on videos of fabrics excited by ambient forces and acoustic waves separately.

We test the invariance of an object's extracted motion spectra to excitation and viewpoint change by training the regression model on the extracted features from one excitation/viewpoint combination and testing on the extracted features from another combination. Table 5.3 shows that correlation results across all combinations of training and testing data are comparable to training and testing on the same viewpoint and excitation. Figure 5-11 visually shows our estimates are still highly correlated with ground truth measurements when the training and testing is performed using different cameras, viewpoints, and excitation forces.
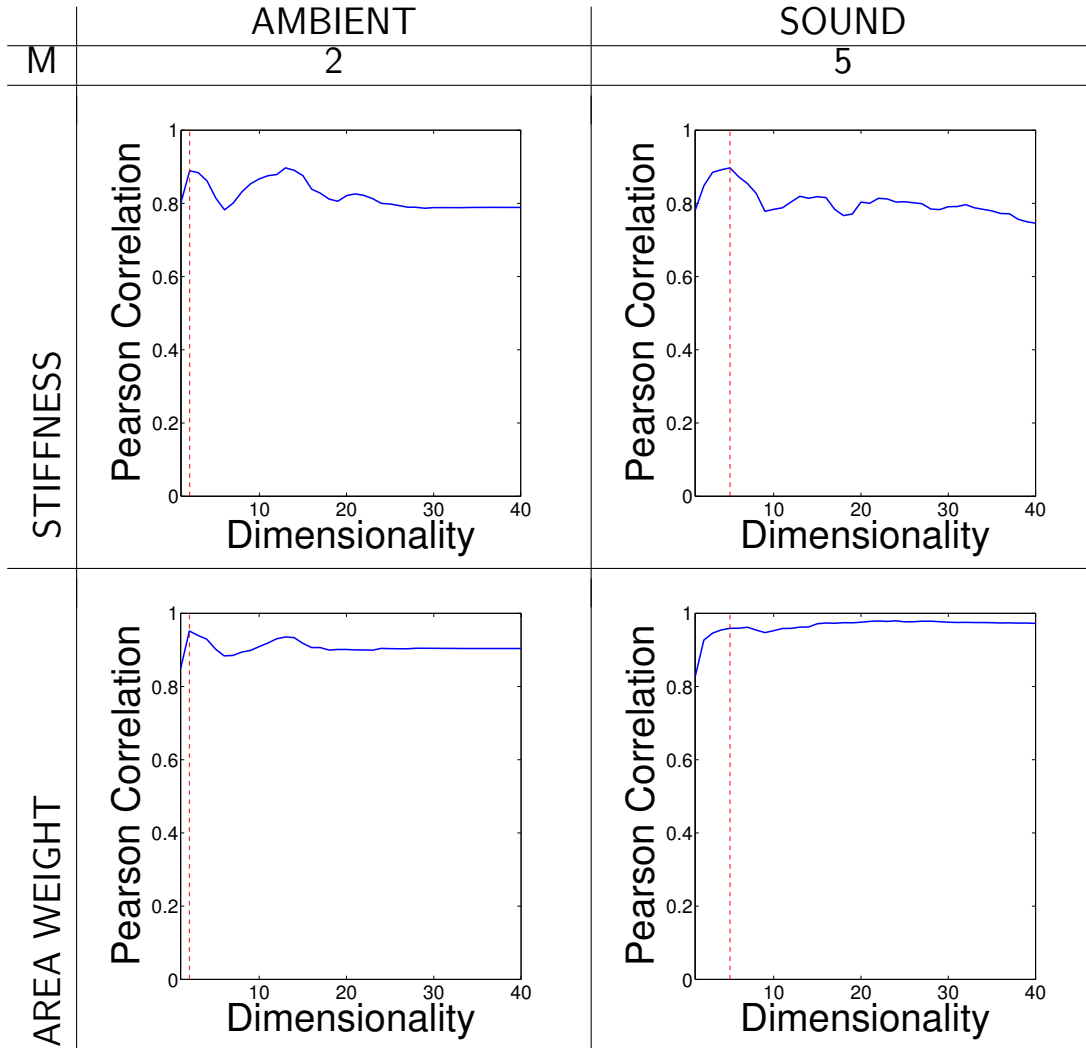
Figure 5-9: The Pearson product correlation value between predicted results and the ground truth measured properties when fitting a model with a varying number of components (dimensionality). The number of components, M, was chosen for each model that resulted in good accuracy for both material properties (stiffness and area weight). These selected M values are specified above and are indicated on the plots as a vertical red line.

Stiffness - Testing

	A/LR	A/L	A/R	S/LR	S/L	S/R
A/LR	0.89	-	-	0.80	-	-
A/L	-	0.89	0.89	-	0.73	0.80
A/R	-	0.87	0.89	-	0.74	0.77
S/LR	0.88	-	-	0.90	-	-
S/L	-	0.87	0.88	-	0.87	0.89
S/R	-	0.86	0.87	-	0.88	0.88

(Training)

Area Weight - Testing

	A/LR	A/L	A/R	S/LR	S/L	S/R
A/LR	0.95	-	-	0.90	-	-
A/L	-	0.94	0.95	-	0.87	0.90
A/R	-	0.94	0.95	-	0.87	0.87
S/LR	0.93	-	-	0.96	-	-
S/L	-	0.92	0.93	-	0.96	0.96
S/R	-	0.91	0.92	-	0.96	0.95

(Training)

Table 5.3: The Pearson correlation R value obtained when training and testing a PLSR model on videos captured under different excitation and viewpoint conditions. The testing and training shorthand notation specifies excitation/viewpoint using abeviations for the four possible conditions: ambient excitation (A), acoustic excitation (S), left camera viewpoint (L) and right camera viewpoint (R). Results are comparable to training and testing on the same viewpoint, suggesting that our features are somewhat invariant to the direction in which the material is observed. Note that all combinations of excitation and viewpoint perform better than results reported in [7].

	[7]	Ambient	Sound
Stiffness	$R = 0.71$	$R = 0.89$	$R = 0.90$
	$\% = 17.2$	$\% = 12.3$	$\% = 12.5$
		$\tau = 0.70$	$\tau = 0.74$
Area Weight	$R = 0.86$	$R = 0.95$	$R = 0.96$
	$\% = 13.8$	$\% = 15.7$	$\% = 13.3$
		$\tau = 0.86$	$\tau = 0.85$

Table 5.4: The Pearson correlation value (R), Percentage Error (%), and Kendall Tau (τ) measures of performance for our PLSR model compared to the performance of a previous video-based fabric property estimation method [7]. The model was trained and tested separately on videos of fabric excited by acoustic waves (Sound) and ambient forces (Ambient).

Frequency Sensitivity and Modes

The theory in Section 3 describes a predictable relationship between resonant frequencies and material properties. However, our regression model has no explicit notion of resonant frequencies; it simply looks for predictive patterns in the spectra of training data. By analyzing the sensitivity of our recovered regression models we can see which frequencies are most predictive of material properties in our fabrics. From the estimated regression coefficients (β_m) and dimensionality reducing basis vectors (E_m), the sensitivity (S) is computed as:

$$S = \sqrt{\left(\sum_{m=1}^{M} \beta_m E_m \right)^2} \tag{5.6}$$

Since the regression model for each of our fabrics is recovered using leave-one-out cross validation, we average the computed sensitivities across models to obtain a single measure of sensitivity for each material property.

Figure 9 shows that frequencies in the 0-5 Hz range were most predictive of material properties in our fabrics. By visualizing the pattern of relative pixel motion recovered for a specific frequency, we see that the fabrics' dominant vibration modes often appear in this frequency range of 0-5 Hz(see Figure 10). This suggests that our models use the same relationship between resonant frequencies and material properties predicted by modal analysis.

5.6   Detecting Changes in Resonance: Glasses of Water

There are many cases where *changes* in an object's resonant frequencies may be useful even when the contributions of material and geometry are left ambiguous. For example, the resonant frequencies of a leaking container will change over time as the container empties. In such a case, the changing resonance indicates a leak, regardless of specific structural or material properties. Similarly, a change in the resonance of a load-bearing structure may call for close attention, regardless of whether the change is caused by material weakening or an unseen change in geometry. One advantage of using
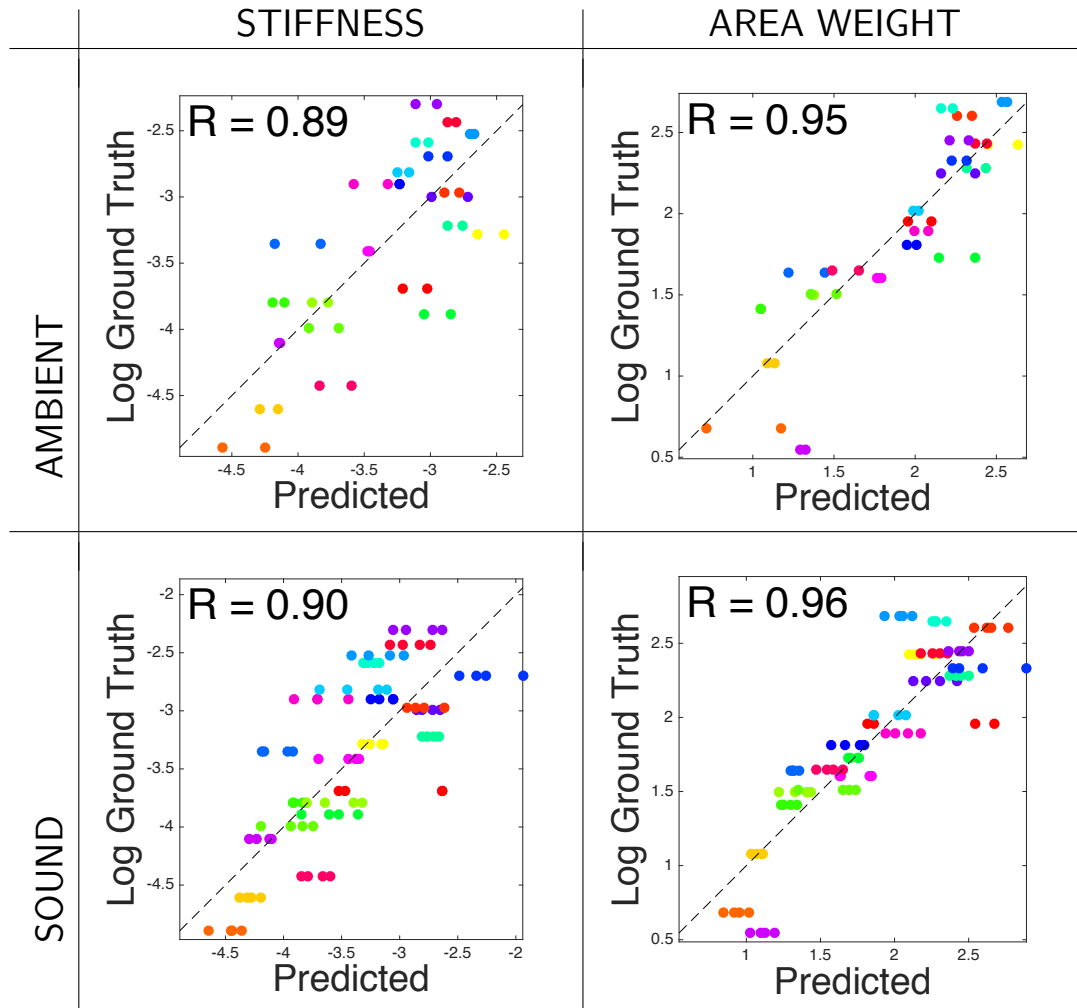
Figure 5-10: Comparisons between ground truth and PLSR model predictions on material properties estimated from videos of fabric excited by ambient forces and acoustic waves. Each circle in the plots represents the estimated properties from a single video. Identical colors correspond to the same fabric. The Pearson product-moment correlation coefficient (R-value) averaged across video samples containing the same fabric is displayed.

Stiffness

Area Weight
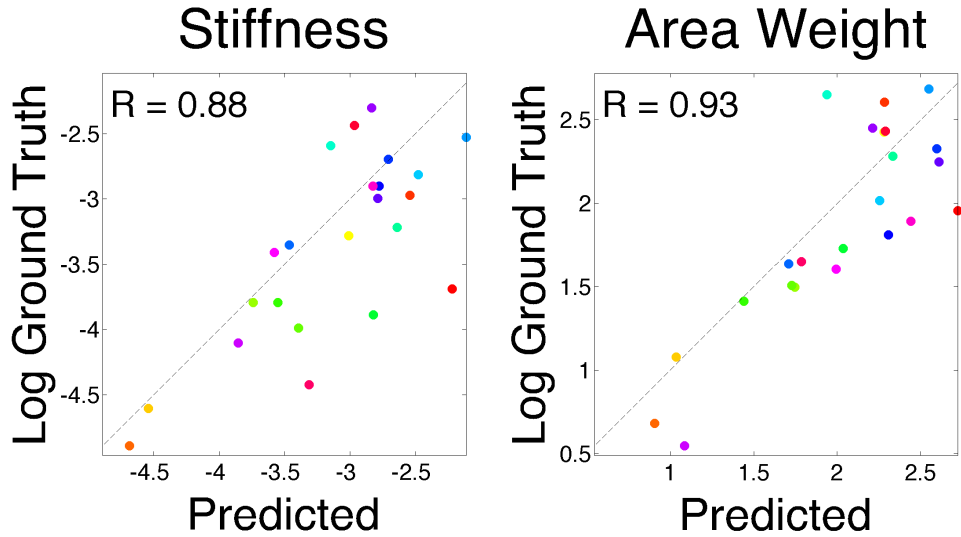
Figure 5-11: The features we use to estimate material properties are somewhat invariant to changes in excitation force and viewpoint. Here we show a comparison between ground truth material properties and PLSR model predictions when using models trained on Point Grey (left viewpoint) videos of fabric exposed to acoustic waves, but tested on SLR videos (right viewpoint) of fabric exposed to ambient forces. Although the training and testing conditions are different, we still perform well.

Stiffness
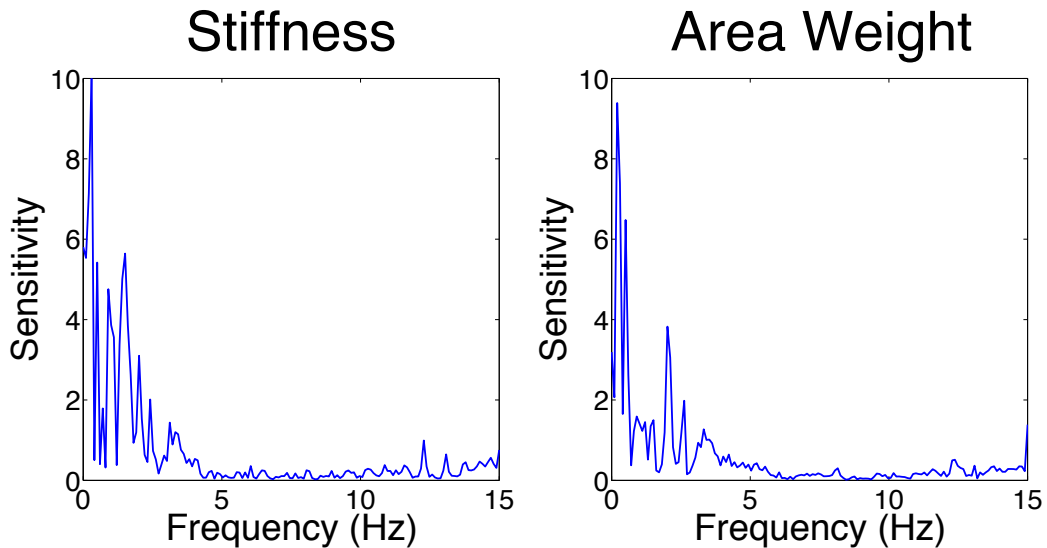
Area Weight

Figure 5-12: The sensitivity of each acoustically trained model to frequency regions in the motion spectrum. These sensitivity plots suggest that energy in the low frequencies is most predictive of a fabric's area weight and stiffness.
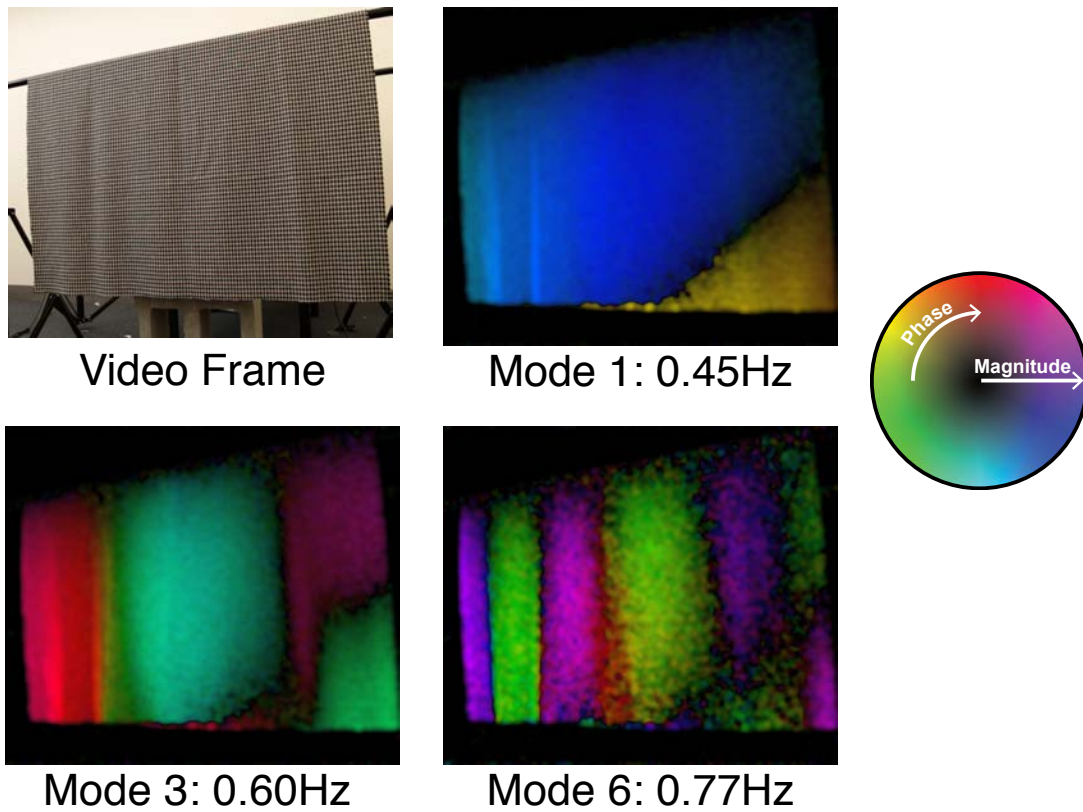
Figure 5-13: A sample of the recovered motion patterns for predictive frequencies identified by the regression models. These recovered motion patterns often resemble a fabric's mode shapes. Phase specifies the relative direction of the motion signal. Pixels moving in opposite directions are colored with hue from opposite sides of the color wheel.

Glasses Seen From Above

Camera View
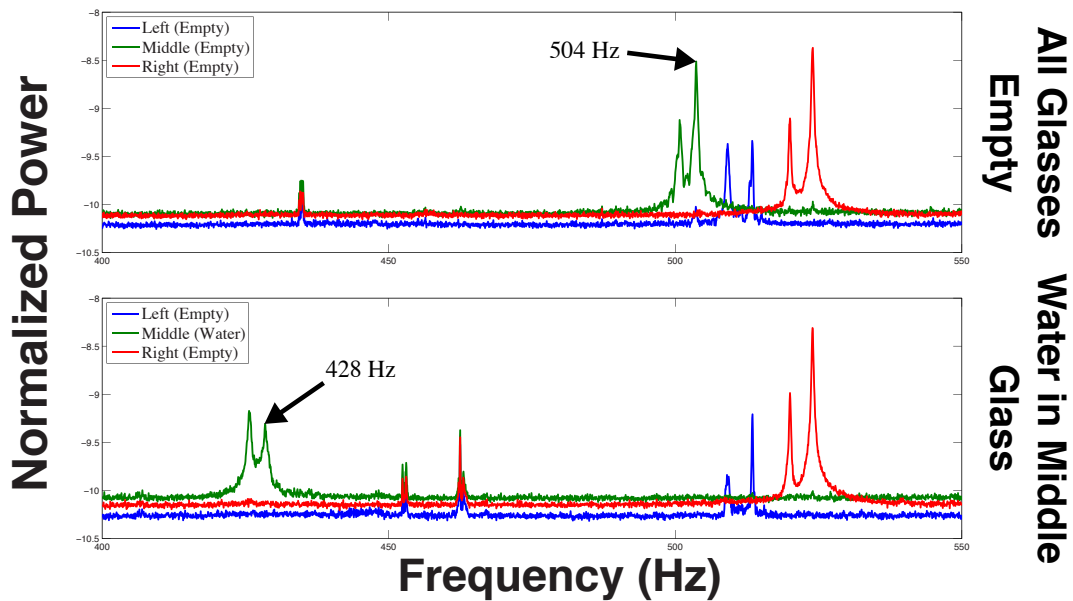
Figure 5-14: (Left) Three wine glasses are set on a table. They are filmed twice - once with all three empty and once with the middle glass partially filled with water (shown left). (Middle above) The glasses are partially occluded to that their contents are not visible, and a nearby loudspeaker plays a 15 second linear chirp of frequencies ranging from 200Hz to 800Hz. (Middle below) The rims of the glasses are filmed at 2.5kHz. (Right) Masks are used to extract the motion spectra of each glass from each video seperately. (Right above) When all glasses are empty, they show resonant peaks within the range of 500-530Hz. (Right bottom) When only the middle glass is filled with water, resonant frequencies of the empty glasses remain unchanged, while the resonant peak of the glass containing water shifts by 76Hz, to 428Hz.

resonance in such a scenario is that the souce of the problem, or change, does not have to be visible - shifting frequencies at visible parts of the object may reveal hidden or occluded changes. In this section we show a simple experiment, analogous to the example of a leaking container, to demonstrate how our recovered motion spectra could be used to detect hidden changes to an object.

The following experiment demonstrates that we can infer when a wine glass is empty or full by observing the vibrations of its rim. For this to be the case, the changes in resonant frequencies resulting from adding liquid to a glass have to be significant compared to natural variations over time, or between the glasses. We compare motion spectra extracted from two videos and show that the addition of water results in a shift of the spectra's peaks. In the first video, all three glasses were left empty. In the second, the middle glass was filled with water.

5.6.1 Setup

Three wine glasses were placed on a table (Figure 5-14 left) next to a loudspeaker and partially occluded so that their contents were hidden from view (Figure 5-14 middle, top). The tops of these wine glasses were filmed to recover vibrations caused by a loudspeaker - once with all three glasses empty and once with only the center glass filled approximately $\frac{2}{3}$ with water. Our goal was to see whether the hidden addition of water to the center glass could be easily detected in our recovered motion spectra.

5.6.2 Excitation

We played a 15 second linear chirp of frequencies ranging from 200Hz to 800Hz through the loudspeaker.

5.6.3 Video Capture

The tops of the glasses were filmed with a Phantom high-speed camera at 2500 fps for approximately 17.3 seconds. The video was captured at a resolution of 1248x153 pixels (an example frame is given in Figure 5-14 middle, bottom). To evaluate the motion spectrum for each glass seperately, a mask that segmented a single glass from the video frame was applied to the local, pixel motion spectra before averaging down to a single spectrum.

5.6.4 Results

Figure 5-14 (right) shows the motion spectra recovered from each glass in each of the two videos. In the spectra recovered from the first video, we see that the empty glasses have resonant peaks within 30Hz of one another. In the spectra recovered from the second video, we see no noticable change in the resonant frequencies of the empty glasses, but the water has shifted the resonant frequencies of the middle glass by approximately 76Hz.

5.7 Comparison With Traditional Vibrometry

The motion spectra we recover from video are analogous to spectra derived from laser vibrometers and accelerometers for traditional vibration analysis. To compare these different types of sensors we conducted an experiment where a steel cantilever beam was measured simultaneously with a high-speed camera, a laser vibrometer, and a piezoelectric accelerometer. A shaker was mounted to the top of the beam, and driven with a sum of sinusoids at resonant modes of the beam. The accelerometer was mounted directly to the beam, the laser vibrometer measured the motion of the accelerometer, and a high-speed camera recorded a video of the accelerometer and beam motion. All three measurement methods were used concurrently in time, measuring the same vibrations of the beam at the same location. The laser vibrometer and accelerometers sampled at 9kHz, while the video captured 2000 fps. each sensor recorded for approximately 15 seconds.
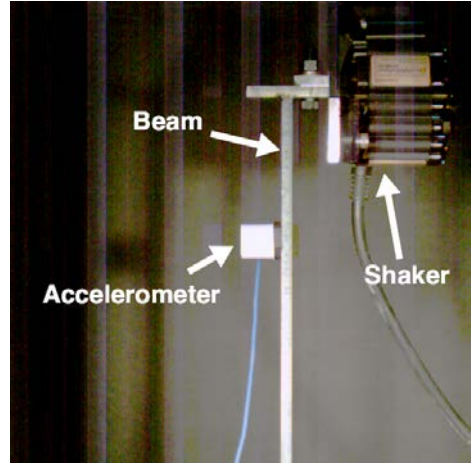
Figure 5-15: Example frame from our video of a forced beam, captured simultaneously with a video, laser vibrometer, and accelerometer.

[76] also compared accelerations measured with a laser vibromater with video measurements, focusing on a time domain analysis. Here, we study differences in the specra of recovered motions. It is natural for each sensor to produce slightly different spectra, as each tests a different derivative of position (the accelerometer measures acceleration, the vibrometer measures velocity, and our method measures position). However, we focus specifically on comparing the resonant frequencies and damping estimated in each case.

5.7.1 Frequency and Damping Estimates

Spectra recovered using each of the three techniques can be seen in Figure 5-16. Mode frequencies for each of these spectra were detected as the local maximum around each resonant peak, and are shown in Table 5.5. As all three sensors were recording the same object, we used the same range of frequencies to fit damping around each peak (± 3Hz). Recovered damping values can be found in Table 5.6

Figure 5-16 shows that the overall shape of spectra recovered using each of the three methods is very similar, though some harmonic artifacts are present in the spectra recovered using our technique. Table 5.5 shows that all three methods agree on the locations of resonant frequencies to within quantization errors. Table 5.6 shows that our method disagrees with the accelerometer and vibrometer on two out of three of the modes, with our strongest disagreement in the fundamental, where our estimate is approximately 39% higher. This amount of error is large relative to the differences in damping ratios for similar metals, but small compared to the differences between metals and materials like wood or rubber.

5.8 Discussion

We have shown that it is possible to learn about the material properties of visible objects by analyzing subtle, often imperceptible, vibrations in video. This can be done in an active manner by recording
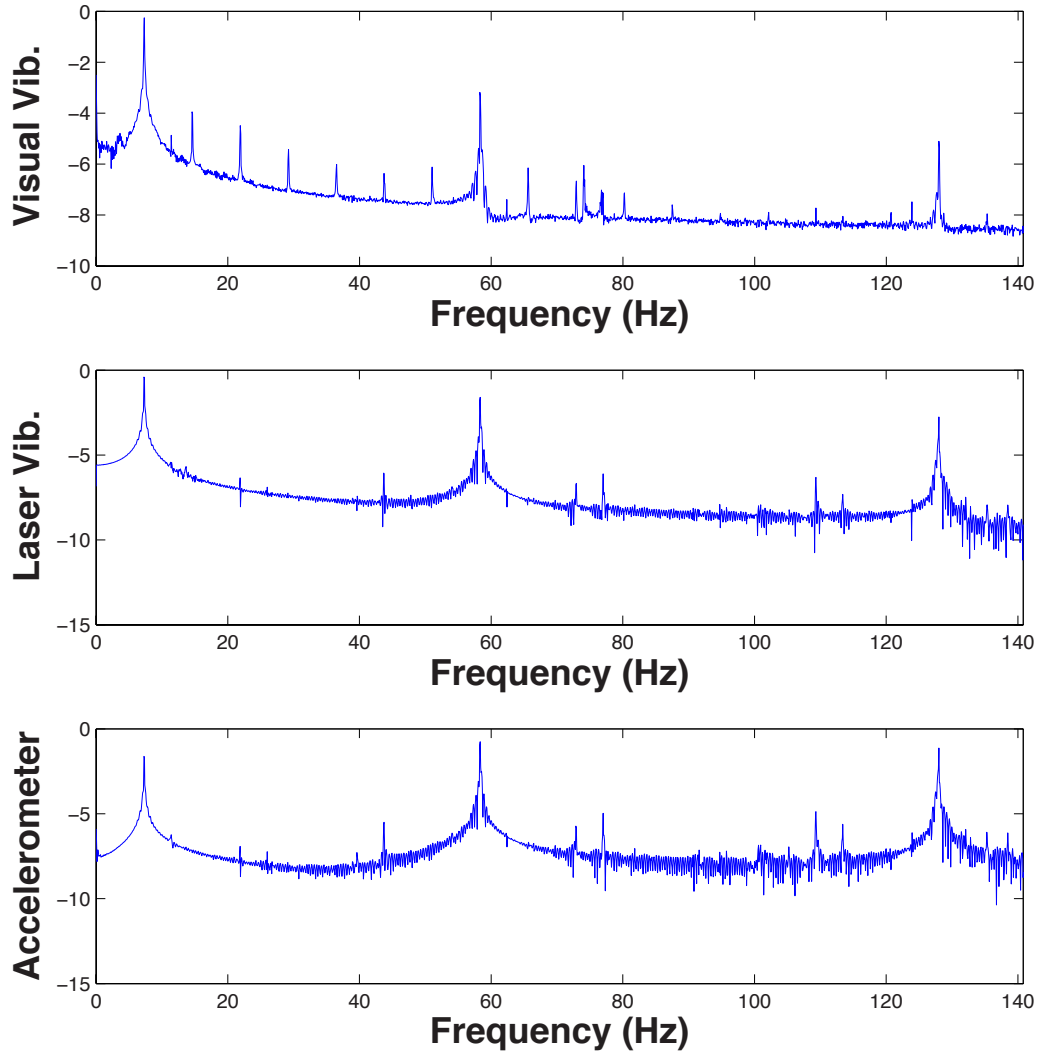
Figure 5-16: Recovered motion spectra from our beam experiment using visual vibrometry (top), a laser vibrometer (middle), and an accelerometer (bottom).

Estimated Frequency	Mode 1	Mode 2	Mode 3
Visual Vibrometry	7.3Hz	58.3Hz	128Hz
Laser Vibrometer	7.3Hz	58.3Hz	128Hz
Accelerometer	7.3Hz	58.3Hz	128Hz

Table 5.5: Recovered beam mode frequencies using our technique, a laser Doppler vibrometer, and an accelerometer. All mode frequencies agree to within the quantization error of our sampling.

Damping Ratio	Mode 1	Mode 2	Mode 3
Visual Vibrometry	6.1×10^{-3}	6.5×10^{-4}	3.9×10^{-4}
Laser Vibrometer	4.4×10^{-3}	6.5×10^{-4}	2.9×10^{-4}
Accelerometer	4.4×10^{-3}	6.5×10^{-4}	2.9×10^{-4}

Table 5.6: Damping ratios computed using spectra derived from the three different sensors. Each damping ratio was computed by fitting a Lorentzian to a 6Hz region around each identified mode frequency.

video of an object responding to sound, or, in some cases, even passively by observing an object move naturally within its environment.

The rod experiments in Section 5.4 demonstrate how our technique can be used as a low cost alternative to laser vibrometers in settings that are typical for testing manufactured parts (aircraft, automobiles, etc). Our technique also offers an affordable way to apply established methods from structural engineering to applications that require more than single point measurements.

The fabric experiments in Section 5.5 address a relatively unexplored area of potential for vibration analysis. While traditional applications of vibrometry are often limited by the need for detailed measurements and analysis of geometry, the ubiquity and passive nature of video offers unique potential as a way to enable data-driven alternative approaches. Our results on fabrics demonstrate that the relationship between motion spectra and material properties can be learned, and suggests that traditional vibration analysis may be extended to applications where geometry is unknown and only loosely controlled.

The simple wine glass experiment in section Section 5.6 highlights a use case that could be applicable to structural health monitoring and quality control in manufacturing. In these scenarios, precise geometry and material properties are not necessary; by directly comparing the motion spectra of similar objects, or of one object over time, it may be possible to detect failures or defects.

Our results suggest that the motion spectra we extract from video can be powerful features for scene understanding. The theory in Chapter 3 suggests that even when geometry is ambiguous, these spectra constrain the physical properties of visible objects. These constraints could be useful for many tasks in computer vision - just as color is often useful despite being an ambiguous product of reflectance and illumination. We believe that video motion spectra can be a powerful tool for reasoning about the physical properties of objects in the wild.

Our work offers cameras as a promising alternative to the specialized, laser-based equipment that is traditionally used in many applications in civil engineering and manufacturing.

6

Interactive Dynamic Video

6.1 Introduction

Computational photography seeks to capture richer information about the world, and provide new visual experiences. One of the most important ways that we experience our environment is by manipulating it: we push, pull, poke, and prod to test hypotheses about our surroundings. By observing how objects respond to forces that we control, we learn about their dynamics. Unfortunately, video does not afford this type of manipulation - it limits us to observing the dynamics that were recorded. However, in this chapter we show that many videos contain enough information to locally predict how recorded objects will respond to new, unseen forces. We use this information to build image-space models of object dynamics around a rest state, letting us turn short video clips into physically-plausible, interactive animations.

Most techniques for physically-based animation derive the properties that govern object dynamics from known virtual models. However, measuring these properties for objects in the real world can be extremely difficult, and estimating them from video alone is severely underconstrained. A key observation of our work is that there is often enough information in video to create a physically plausible model of object dynamics around a rest state in which the object is filmed, even when fundamental ambiguities make recovering a general or fully-accurate model impossible. We show how to extract these physically plausible models from short video clips, and demonstrate their use in two applications.

Most of this chapter was originally published in our paper [23] in collaboration with Justin G. Chen and Frédo Durand. (URL)

Interactive Animation: Video makes it easy to capture the appearance of our surroundings, but offers no means of physical interaction with recorded objects. In the real world, such interactions are a crucial part of how we understand the physical properties of objects. By building a model of dynamics around the state in which an object is filmed, we turn videos into interactive animations that users can explore with virtual forces that they control.

Special Effects: In film special effects, where objects often need to respond to virtual forces, it is common to avoid modeling the dynamics of real objects by compositing human performances into virtual environments. Performers act in front of a green screen, and their performance is later composited with computer-generated objects that are easy to simulate. This approach can produce compelling results, but requires considerable effort: virtual objects must be modeled, their lighting and appearance made consistent with any real footage being used, and their dynamics synchronized with a live performance. Our work addresses many of these challenges by making it possible to apply virtual forces directly to objects as they appear in video.

6.1.1   Overview

Our approach is based on the same linear modal analysis behind many techniques in physically-based animation. However, unlike most of these techniques, we do not assume any knowledge of object geometry or material properties, and therefore cannot rely on finite element model (FEM) methods to derive a modal basis for simulation. Instead, we observe non-orthogonal projections of an object's vibration modes directly in video. For this we derive a relationship between projected modes and the temporal spectra of optical flow. We then show that, while non-orthogonal, these projections can still be used as a basis to simulate image-space object dynamics.

Recovering accurate physical models of objects in video is severely underconstrained. To deal with this ambiguity, we make a few key assumptions, which we analyze in Section 6.3.1.

6.2   Related Work

Physically-based Animation: Many techniques in physically-based animation use modal analysis to reduce the degrees of freedom in deformable body simulations [54, 41, 42, 53, 40, 48]. These techniques work by first deriving orthogonal vibration modes from known geometry using FEM approaches. As high frequency modes generally contribute less to an object's deformation, they can often be discarded to obtain a lower-dimensional basis for faster simulation. We use a similar reduced modal basis to simulate objects in video, but assume no knowledge of scene geometry and cannot therefore use FEM approaches to compute vibration modes. Instead, we observe projections of these modes directly in video and show that, while non-orthogonal, these projections can still be used as a basis to simulate the dynamics of objects in image-space.

Observing Vibration Modes   The problem of directly observing vibration modes has been explored in several engineering disciplines, where the structure of objects must be carefully validated in the real world, even when a virtual model is available. The general approach is to relate the spectrum of surface motion, typically measured with accelerometers, to mode shapes. [38] applied

this analysis to motion estimated with a stereo rig, which they used to recover mode shapes for shell-like structures.

Recent work in graphics and vision has used narrow-band phase-based motion magnification to visualize the modal vibrations of objects in video [76, 77, 16]. [24] proposed an alternative visualization based on the temporal spectra of weighted optical flow. However, both approaches focus on providing a visualization tool, and neither has been used to recover a basis for simulation. We show that a similar algorithm, borrowing aspects of each of these visualization techniques, can be used to recover mode shapes that are suitable for simulation.

Motion Synthesis in Video: Several works in computer graphics and vision have focused on synthesizing plausible animations of quasi-periodic phenomena based on a video exemplar [27, 71, 18, 63, 55, 73]. In most of these applications, video synthesis is formulated as a stochastic process with parameters that can be fit to the exemplar. Such approaches work especially well for animating phenomena like rippling water or smoke, and with skeletal information provided by a user have even been extended to model the motion of structures caused by stochastic forces like wind [68, 70]. The applications we address are similar to many of these works in spirit, but, to our knowledge, we are the first to build image-space simulations based on a modal bases extracted directly from video.

Motion Magnification Like recent publications in motion magnification [76, 77, 16], our work can be used to magnify and visualize small vibrations of an object. However, our work is different from motion magnification in several key ways. First, while motion magnification is a time-varrying representation of motion, our technique extracts a static representation of each vibration mode, and can therefore average over the entire input video to reduce noise at each mode. Second, while phase-based methods for Eulerian motion magnification rely on expensive pyramid decompositions of video at render time, our approach to synthesis is Lagrangian and can be implemented efficiently on the GPU, allowing for real-time synthesis of motion composed of many vibration modes. Finally, while motion magnification only magnifies motion that was already present in a captured video, our technique can be used to synthesize responses to new combinations of forces that were never observed in the input.

6.3 Modal Images as a Basis

In this section we build on our derivations from Chapter 3 to show that modal images can be used as a basis for representing image-space dynamics. We first consider the dynamics of a single degree of freedom, which we later relate to the motion of a visible point in video.

An excitation force $\mathbf{f}$ given in modal coordinates can be decomposed into a set of impulses $\mathbf{f}_i = \alpha_i \delta(t)$ where α_i is the amplitude of the impulse at mode ϕ_i. Applying Equation 3.13, the response of the object at one degrees of freedom $\mathbf{x}_p(t)$ is given by

$$\mathbf{x}_p(t) = \sum_{i=1}^{N} \alpha_i h_i(t) \phi_i(p) \tag{6.1}$$

where $\phi_i(p)$ is the mode shape coefficient of the degree of freedom p of the object for mode i. Using Equations 3.14 and 6.1 we can construct the Fourier transform of Equation 6.1 as

$$\mathrm{X}_p(\omega) = \sum_{i=1}^{N} \alpha_i H_i(\omega)\phi_i(p) \tag{6.2}$$

Here we make an assumption that is common in engineering modal analysis [26, 8], but not necessary in FEM-based applications of modal analysis for simulation: that modes are well spaced, or non-overlapping in the frequency domain. Under this assumption, we can represent the frequency response of a single degree of freedom at ω_{di} as

$$\mathrm{X}_p(\omega_{di}) = \alpha_i H_i(\omega_{di})\phi_i(p). \tag{6.3}$$

Our next assumption is weak perspective - a common approximation in computer vision, but one that is also not necessary when modes are derived from known models. Using this approximation we align our object's coordinate system with the image plane of an input video, giving us observable degrees of freedom for each pixel's motion in the x and y dimensions of our image. For the purpose of derivation, we represent visibilty across all degrees of freedom with the unknown, binary, diagonal matrix $\mathbf{V}$, which multiplies the visible degrees of freedom in a mode by 1 and all other degrees of freedom by 0. The projection of a mode shape ϕ_i into the image plane is then $\mathbf{V}\phi_i$.

By taking Fourier transforms of all local motions $\mathbf{Vx}$ observed in video we obtain $\mathbf{VX}$, the Fourier spectra for visible degrees of freedom, which, evaluated at resonant frequencies ω_{di}, is

$$\mathbf{VX}(\omega_{di}) = \alpha_i H_i(\omega_{di})\mathbf{V}\phi_i. \tag{6.4}$$

Here, α_i and $H_i(\omega_{di})$ are constant across all degrees of freedom p, meaning that $\mathbf{VX}(\omega_{di}) \propto \mathbf{V}\phi_i$. Therefore we can treat the set of complex ϕ_i', the values of $\mathbf{VX}(\omega_{di})$ measured in video, as a basis for the motion of the object in the image plane.

6.3.1 Assumptions and Limitations

While linear motion is a standard assumption of linear modal analysis that usually applies to the type of small motion we are analyzing, our derivation makes a few key approximations that are not typical of modal analysis applied to simulation:

- *Weak Perspective* - Our analysis assumes that linear motion in 3D space projects to linear motion in the image plane. This can be violated by large motion in the z-plane.

- *Well-spaced modes* - We rely on separation in the frequency domain to decouple independent modes. This can fail in objects with strong symmetries, high damping, or independent moving parts.

- *Broad-Spectrum Forcing* - By using observed modes as a basis for the motion of an object in the image plane, we make an implicit assumption about the ratio of modal masses to observed modal forces. Allowing for an ambiguity of global scale, this assumption is still violated when observed forces are much stronger at some modes than others.

Because we deal with small motion around a rest state, weak perspective is almost guaranteed to be a safe approximation. However, there are many cases where our remaining two assumptions could fail. Fortunately, the consequences of these failures tend to affect the accuracy more than the plausability of simulation. Consider the failure cases of each approximation. Overlapping modes will

cause independent objects to appear coupled in simulation - in other words, the response of an object to one force will incorrectly be an otherwise appropriate response to multiple forces. Similarly, when broad-spectrum forcing is violated, the response of a object to one force will be the appropriate response to a differently scaled, but equally valid set of forces. In both cases, the failure results in inaccurate, but still plausible deformations of the object.

6.4 Algorithm

Our algorithms first extracts a volume of candidate vibration modes from an input video. We then provide a user interface for selecting a subset of these candidate modes to use as a basis for simulation.

6.4.1 Mode Selection:

Under ideal conditions, the observed candidate modes ϕ'_ω at each frequency ω would be zero everywhere but at real mode shapes. However, real video contains unintended motion from a variety of sources (e.g., camera shake, noise, moving background). To distinguish between object deformations and unintended motion from other sources, we first ask users to provide a rough mask of the content they are interested in. We then present them with our mode selection interface, as described in Chapter 3, to help select mode shapes. Using this interface users can select either an individual, or a range of candidate images to use as a basis for simulation.

6.4.2 Complex Mode Shapes:

Note that the set of mode shape solutions ϕ_i to Equation 3.7 are real-valued, i.e. they only have binary phase relationships. Similarly, the mode shapes derived using FEM in typical simulation applications are also real-valued. In contrast, the mode shapes we recover may have non-binary phases. This can happen for a number of reasons, including noise or a violation of one of our assumptions. We could force mode shapes to be real-valued by projecting them onto their dominant axis in the complex plane, however, we found that allowing non-binary phases actually improves results. Visually, such mode shapes allow for features like traveling waves and partial coupling that might otherwise require much higher-order modes to represent. By allowing these shapes, we effectively let our representation fit the motion in a video more closely. In this sense, our technique is allowed to behave a little more like methods for exemplar-based motion texture synthesis in situations where motion cannot be explained well with sparse, low-frequency modes.

To ensure that the behavior of our simulation reduces to one using only real mode shapes when observed modes contain only binary phase relationships, we calculate the dominant orientation of each selected mode shapes on the complex plane, and rotate all phases so that this orientation aligns with the real axis.

6.4.3   Simulation

Our simulation works on the state of an object in modal coordinates. The key components are a way to evolve the state of an object over time, and a way to translate user input into forces, displacements, and velocities.

Given Equation 3.11, we can define a state space model per modal coordinate to simulate the the object over time. We define the state vector $\mathbf{y_i}$ that describes the system for a single modal coordinate $\mathbf{y_i} = [\varrho_i \ \dot{\varrho}_i]^\intercal$, where ϱ_i and $\dot{\varrho}_i$ are the modal displacement and velocity vectors respectively which relate to the complex modal coordinate by $\mathbf{q_i} = \varrho_i - i\dot{\varrho}_i/\omega_i$. We evolve the state to $\mathbf{y}[n+1]$ given $\mathbf{y}[n]$ and a modal force $\mathbf{f_i}$ using the equation[1]:

$$\mathbf{y}[n+1] = \begin{bmatrix} 1 & h \\ -\omega_i^2 h & 1 - 2\xi_i\omega_i h \end{bmatrix} \mathbf{y}[n] + \begin{bmatrix} 0 \\ h/\mathbf{m}_i \end{bmatrix} \mathbf{f}_i[n], \tag{6.5}$$

and set h, the amount of time passed in the simulation, to be small enough to ensure that this equation is stable.

6.4.4   User Input

We provide users with modes of interaction that can be divided into two categories: forcing interactions and direct manipulations. Forcing interactions affect state indirectly by changing the force $\mathbf{f}_i$ applied to an object. Direct manipulations translate user input directly into intantaneous state $\mathbf{y}$.

Forcing Interactions: Forcing interactions translate user input into a force to be applied at a specified point. In the simplest forcing interaction, a user clicks at a point $\mathbf{p}$ on the object, and drags their mouse in a direction $\mathbf{d}$. We interpret this as specifying a force $\mathbf{f}$ to be applied at the point $\mathbf{p}$ in the direction $\mathbf{d}$. The scalar modal force $\mathbf{f_i}$ applied to each mode is computed by taking the magnitude of the dot product of $\mathbf{d}$ with the value of that mode shape ϕ_i' at point $\mathbf{p}$:

$$\mathbf{f_i} = \|\mathbf{d}^\intercal \phi_i'(\mathbf{p})\| \alpha \tag{6.6}$$

where α is used to control the strength of the force, and can be set by the user with a slider. Note that we take the magnitude here because the mode shape ϕ_i' is complex.

Direct Manipulation: Real objects are often found in configurations that are difficult or impossible to achieve through forces applied to one point at a time. However, fully specifying shaped forces is a difficult user interaction problem. We instead offer a mode of interaction that lets users directly manipulate the position or velocity of a single point. This lets users explore states with greater contributions from higher-order modes that are difficult to achieve without shaped forces. We accomplished this by explicitly setting the state of the object whenever the user's mouse is pressed, and only letting the state evolve once the mouse is released. As with forcing interactions, the user specifies a point $\mathbf{p}$ and direction $\mathbf{d}$ with a mouse. We then compute the magnitude of each modal coordinate in the same way that we computed the magnitude of modal forces before:

---

[1]A derivation of this equation can be found in [64]

$$\|\mathbf{q_i}\| = \|\mathbf{d^\intercal}\phi_i'(\mathbf{p})\|\alpha \tag{6.7}$$

where α is used to control the strength of the manipulation, and can be set by the user with a slider. However, in this case we set the phase of the modal coordinate to maximize either the displacement or velocity of $\mathbf{p}$ in the direction $\mathbf{d}$. This is accomplished by setting the phase $Arg(\mathbf{q_i})$ to

$$\text{Max Displacement:} \quad Arg(\mathbf{q_i}) = -Arg(\mathbf{d^\intercal}\phi_i'(\mathbf{p})) \tag{6.8}$$

$$\text{Max Velocity:} \quad Arg(\mathbf{q_i}) = -Arg(\mathbf{d^\intercal}\phi_i'(\mathbf{p})) + \frac{\pi}{2} \tag{6.9}$$

For objects with real mode shapes, velocity is maximized when displacements are zero, and displacement is maximized when velocities are zero. Intuitively, maximizing displacement lets users 'pull' a point around the screen and see how the the object deforms in response, while maximizing velocity specifies an impulse to be applied when the mouse is released.

6.4.5   Rendering Deformations

We render the object in a given state by warping a single color image, representing the object's rest state, by a displacement field $\mathbf{D}(t)$. $\mathbf{D}(t)$ is calculated as a superposition of mode shapes weighted by their respective modal coordinates:

$$\mathbf{D}(t) = \sum_i^N \mathbf{Re}\{\phi_i' q_i(t)\} \tag{6.10}$$

This can be evaluated efficiently on the GPU by representing each ϕ_i' as an RGBA texture storing two complex numbers per pixel, corresponding to the coupled image-space x and y displacements of ϕ_i'. Each $\phi_i' q_i(t)$ term is computed in a single rendering pass, accumulating $\mathbf{D}_t$ in a framebuffer that can be applied as a displacement map to the color image in a final pass. Our implementation uses depth culling and assigns pixels depth values that are inversely proportional to the magnitude of their displacement, causing parts of the image that move more to occlude parts that move less. This tends to work better than blending pixel values in practice, as objects closer to the camera usually exhibit larger screen space motion due to foreshortening.

6.4.6   Implementation Details

Our mode extraction and selection interface are written in MATLAB. Once modes have been selected, they are exported as 8-bit RGBA TIFF images, and loaded into our simulation software, which is written in C++ and uses Qt, OpenGL, and GLSL.

The slowest part of our algorithm is building a complex steerable pyramid on the input video. Using the MATLAB implementation from [67] this takes less than two minutes on shorter videos like the Wireman, but can take 2-3 hours on longer, or high-speed videos like the Ukulele. The only parameter we set for this is the standard deviation of the gaussian used for filtering local motion signals. Our strategy for setting this parameter is to effectively test out 4 values at once - we pick a standard deviation that is 5-10% of the larger image dimension, filter with this standard deviation at all scales, and use the highest-resolution scale that does not appear noisy. Mode selection can

then usually be done in less than a minute, but users may choose to spend more time exploring the recovered spectra with our selection interface.

In the Playground, YoutubeBridge, and ForceTree examples we use inpainting to fill disoccluded parts of the image.

6.5 Results

We tested our method on several different examples. Thumbnails showing the rest state of each example can be found in Table 6.6 along with additional details about the corresponding input video.

All of the input videos that we captured were recorded with a tripod. The input video for YoutubeBridge was downloaded from Youtube user KOCEDWindCenter (link).

Our simulations plausibly reproduce the behavior observed in most input videos. Our method works well with regular cameras operating at 30 frames per second. While higher-frequency modes exist in most objects, their fast temporal dynamics are not usually visible in output videos, just as they are not observable in the input. Our Ukulele example explores the use of a high speed camera to recover modes that are not visible at normal framerates.

Interactive Animations Video showing interactive sessions with our examples can be found in the supplemental material. In each interactive session, an arrow is rendered to indicate where users click and drag. The head of the arrow points to the current mouse location, and the tail of the arrow ends at the displaced point $\mathbf{p}$ where the user initially clicked.

For the most part, interactive animations are quite compelling. However, in some cases where our non-overlapping modes assumption is violated, independent parts of a scene appear coupled. This effect is subtle in most of our results, so we include an additional failure case designed to violate this assumption in our supplemental material (labeled 'dinos1'). The example shows two dinosaur toys with similar motion spectra resting on the same surface. When a user interacts with one of the toys, this causes some motion in the other toy as well. This problem could be addressed in the future by asking users to provide multiple masks, indicating independent parts of the scene.

We include another additional example in our supplemental material, labeled belly1, simulating the belly fat of a shirtless male. This example was designed as a real-world version of the main example used in [41]. It also shows the effect of a user changing damping during simulation.

Special Effects A variety of visual effects can be be achieved by specifying forces in different ways. We explore the possibility of using this to create low-cost special effects. For example, by using forcing interactions and setting $\mathbf{d}$ to be a vector pointing down, we can simulate the effect of increased weight at the point $\mathbf{p}$. In our supplemental video we use this to simulate a small robot rolling along the surface of different objects. When the robot 'lands' on a point $\mathbf{p}$ of the object, we fix the robot to $\mathbf{p}$ by applying the time-varying displacement at $\mathbf{p}$ to the image of the robot at each frame. By moving $\mathbf{p}$ along a trajectory specified in the object rest state, we cause the robot to 'roll' along the object's surface in a way that couples their dynamics.

In another example, ForceTree, we control the force $\mathbf{d}$ applied to branches of a tree so that the branches appear to be controlled by a moving hand elsewhere in the video. In this way, we make it appear as though the leaves of the tree are coupled (or controlled through some supernatural

force) by the hand. This is substantially simpler that modeling a synthetic tree and matching its appearance to the filmed scene.

6.6   Conclusion

We have shown that, with minimal user input, we can extract a modal basis for image-space deformations of an object from video and use this basis to synthesize animations with physically plausible dynamics. We believe that the techniques in this chapter can be a valuable tool for video analysis and synthesis. The interactive animations we create bring a sense of physical responsiveness to regular videos. Our work could also lead to low-cost methods for special effects by enabling the direct manipulation of objects in video.
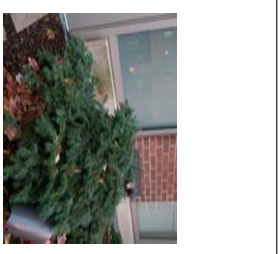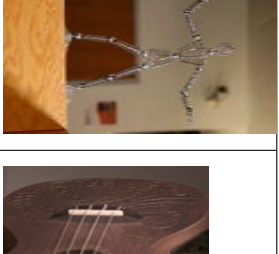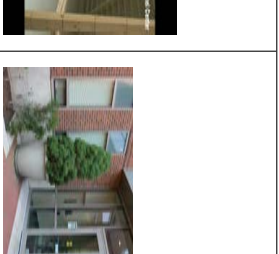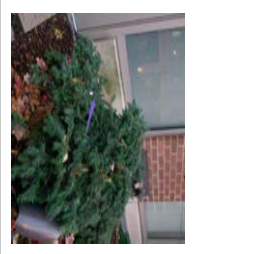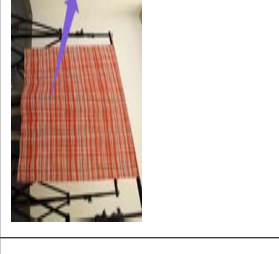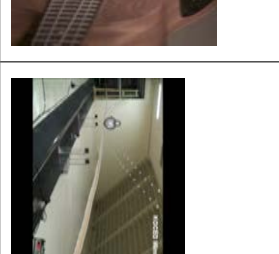
Example Name						
Bush	Playground	Cloth	Wireman	Ukulele	YoutubeBridge	ForceTree
Input Video Image						
Synthesized Deformation						

Table 6.1: This table gives a summary of the experimental results. The first row contains the names of all the examples. The middle row contains an image from the input video representing the rest state of the object, and the bottom row is an example of a synthesized deformation.

Example	Source	Source Length (s)	Framerate (fps)	Resolution	Excitation	Number of Modes	Frequency Range
Bush	SLR	80.18	60	640 × 480	Ambient/Wind	77†	1.3 - 4.2 Hz
Playground	SLR	53.85	60	1280 × 720	Impulse	34†	0.8 - 22 Hz
Cloth	SLR	59.77	30	1920 × 1080	Ambient/Wind	147†	0.3 - 0.8 Hz
Wireman	SLR	5.82	60	720 × 1280	Impulse	6	5 - 20 Hz
Ukulele	High-speed camera	8.87	1400	432 × 576	Sound	13	219 - 670 Hz
YoutubeBridge	Youtube (link)	50	30	640 × 480	Wind	18	0.25 - 11 Hz
ForceTree	SLR	35	60	1280 × 720	Impulse	13	0.6 - 9 Hz

† Range of frequencies selected

Table 6.2: This table gives a summary of the parameters of the experimental results. We give the source, length, framerate, and resolution of the source video. The excitation column describes the type of excitation used to excite the object in the input video where: ambient/wind means natural outdoor excitations mostly due to wind, impulse means that the object or its support was manually tapped, and sound means that a ramp of frequencies was played from 20 Hz to the Nyquist rate of the recorded video. We give the number of mode shapes identified from the input video local motion spectra that are used to simulate the object response and in the final column, the frequency range of these mode shapes.

7

Conclusion

Contributions and Applications:

In this dissertation we have shown that cameras and computation can be used to capture and analyze the vibrations of visible objects. In doing so, we have established powerful connections between computer vision, audio processing, and vibration analysis that impact a wide range of applications in a variety of fields

The Visual Microphone

We have shown that it is possible to recover sound from silent video of vibrating objects, turning those objects into visual microphones from a distance. Our work provides cameras as a way to locate, isolate, and even image sounds in an environment.

- **Surveillance:** Some of the most obvious applications of the visual microphone are in surveillance. We can use it to isolate specific sounds in otherwise noisy environments, listen to conversations happening behind sound-proof glass, and, with powerful optics, possibly even hear distant sounds that are too quiet for a regular microphone.

- **Acoustical Engineering:** The visual microphone also has potential in acoustical engineering, where the goal is often to find and reduce sources of unwanted noise. In many environments, locating sources of noise can be quite difficult, as sound may bounce off of many surfaces before reaching the listener. Our work makes it possible to image the vibrations that cause sound directly, making it easier to find their source.

- **Astronomy:** Our work may also be useful in astronomy. As there is no air in space, there is no sound. However, by analyzing visual vibrations, we may be able to "listen" to distant celestial bodies.

Visual Vibrometry

We have shown that the resonant frequencies of visible objects can be extracted from video and used to reason about those objects' physical properties. This work draws powerful connections between computer vision and vibration-based testing methods used in engineering. By offering cameras as a low-cost, ubiquitous alternative to laser vibrometers and accelerometers, we open up exciting new opportunities on both sides.

Engineering

- **Structural Health Monitoring:** Cameras offer exciting new opportunities in structural health monitoring. Large structures, like buildings and bridges, are especially difficult to instrument with accelerometers or laser vibrometers, making visual vibrometry a compelling alternative. These structures tend to have low resonant frequencies, making them easy to capture with regular framerate video.

- **Non-Destructive Testing:** The low cost and passive nature of cameras also makes them appealing for applications of non-destructive testing (e.g. of airplanes, automobiles, etc).

- **Data-driven Vibration Analysis:** Our work opens up exciting opportunities for data-driven vibration analysis by offering cameras as a ubiquitous alternative to the specialized devices used to measure vibrations in engineering.

Computer Vision

- **Material Property Estimation:** We have shown how to estimate the material properties of various objects by examining the spectra of their vibrations in video.

- **Scene Understanding:** The vibrations of visible objects may also reveal information about unseen, or occluded parts of a scene, such as whether an object is hollow, or a container is empty.

Interactive Dynamic Video

We have shown how to recover plausible image-space dynamic models of visible objects by analyzing the shapes and frequencies of their vibrations in video. Leveraging the spatial resolution of cameras, we are able to capture objects in a way that lets us predict how they will respond to new, unseen forces. Our work offers a new representation of objects that captures not just their appearance, but also their dynamics, and introduces interactive video-based simulation as an exciting new direction for computer vision and graphics.

- **Photography:** In the real world, we learn a lot about objects by interacting with them. Unfortunately, traditional images and video do not afford this kind of interaction. Interactive dynamic video is a compelling alternative that offers a richer representation of recorded objects.

- **Computer Graphcis:** The ability to quickly and easily digitize the dynamics of real-world objects has great potential in computer graphics. For example, we have demonstrated how

interactive dynamic video offers a low-cost way to synthesize compelling physical interactions between real objects and virtual characters.

- **Computer Vision:** The ability to model how objects in video move and bend could provide powerful new priors for computer vision.

Looking Forward

Most people understand that a red apple is probably ripe, and a brown apple is probably rotten. Some of us are taught this lesson, and some of us discover it for ourselves – perhaps by biting into an apple that is past its prime. Such experience is part of learning the deep connection between color, and other important properties of the objects we encounter.
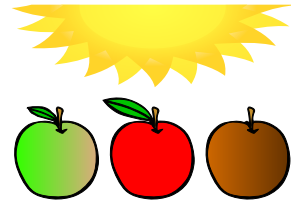
We have similar intuition for the sounds that some objects make: we can tell the difference between a ringing bell, and a beating drum; we know that men tend to have deeper voices than women — we may even recognize a colleague by the distinct jingle of office keys in a nearby hallway. But most of the objects we encounter are silent, and most of the sounds we hear come mixed in a stream of other noises.

In this dissertation we have shown that, with cameras and computation, we can image the vibrations of objects in much the same way that we image color. In doing so, we have established a powerful connection between computer vision and vibration analysis.
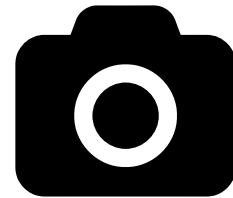
Our work has the potential to impact a wide range of existing applications in a variety of fields. But I believe some of the most exciting opportunities are yet to be discovered. By offering cameras as a way to capture the way that objects vibrate, we have added a new dimension to how we image the world.

Bibliography

[1] Omar Ait-Aider, Adrien Bartoli, and Nicolas Andreff. Kinematics from lines in a single rolling shutter image. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–6. IEEE, 2007.

[2] Vyacheslav Aranchuk, Amit Lal, James M Sabatier, and Cecil Hess. Multi-beam laser doppler vibrometer for landmine detection. *Optical Engineering*, 45(10):104302–104302, 2006.

[3] ASTM Standard E756 - 05. Standard Test Method for Measuring Vibration-Damping Properties of Materials. ASTM International, West Conshohocken, PA, 2010.

[4] Wilfred E Baker, William E Woolam, and Dana Young. Air and internal damping of thin cantilever beams. *International Journal of Mechanical Sciences*, 9(11):743–766, 1967.

[5] Kiran S. Bhat, Christopher D. Twigg, Jessica K. Hodgins, Pradeep K. Khosla, Zoran Popović, and Steven M. Seitz. Estimating cloth simulation parameters from video. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '03, pages 37–51, Aire-la-Ville, Switzerland, Switzerland, 2003. Eurographics Association.

[6] Steven Boll. Suppression of acoustic noise in speech using spectral subtraction. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 27(2):113–120, 1979.

[7] Katherine L. Bouman, Bei Xiao, Peter Battaglia, and William T. Freeman. Estimating the material properties of fabric from video. *Computer Vision, IEEE International Conference on*, 0:1984–1991, 2013.

[8] Rune Brincker, C Ventura, and Palle Andersen. Why output-only modal testing is a desirable tool for a wide range of practical applications. In *Proc. Of the International Modal Analysis Conference (IMAC) XXI, paper*, volume 265, 2003.

[9] A. BUCHARLES, H. CASSAN, and J. ROUBERTIER. *Advanced parameter identification techniques for near real time flight flutter test analysis*. American Institute of Aeronautics and Astronautics, 2016/08/31 1990.

[10] 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.

[11] Oral Buyukozturk, R Haupt, C Tuakta, and J Chen. Remote detection of debonding in frp-strengthened concrete structures using acoustic-laser technique. In *Nondestructive Testing of Materials and Structures*, pages 19–24. Springer, 2013.

[12] Paolo Castellini, Nicola Paone, and Enrico Primo Tomasini. The laser doppler vibrometer as an instrument for nonintrusive diagnostic of works of art: application to fresco paintings. *Optics and Lasers in Engineering*, 25(4):227–246, 1996.

[13] 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.

[14] Justin G Chen, Robert W Haupt, and Oral Buyukozturk. Acoustic-laser vibrometry technique for the noncontact detection of discontinuities in fiber reinforced polymer-retrofitted concrete. *Materials evaluation*, 72(10):1305–1313, 2014.

[15] Justin G Chen, Neal Wadhwa, Young-Jin Cha, Frédo Durand, William T Freeman, and Oral Buyukozturk. Structural modal identification through high speed camera video: Motion magnification. In *Topics in Modal Analysis I, Volume 7*, pages 191–197. Springer, 2014.

[16] Justin G Chen, Neal Wadhwa, Young-Jin Cha, Frédo Durand, William T Freeman, and Oral Buyukozturk. Modal identification of simple structures with high-speed video using motion magnification. *Journal of Sound and Vibration*, 345:58–71, 2015.

[17] 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.

[18] Yung-Yu Chuang, Dan B Goldman, Ke Colin Zheng, Brian Curless, David H. Salesin, and Richard Szeliski. Animating pictures with stochastic motion textures. *ACM Trans. Graph.*, 24(3):853–860, July 2005.

[19] L Collini, R Garziera, and F Mangiavacca. Development, experimental validation and tuning of a contact-less technique for the health monitoring of antique frescoes. *NDT & E International*, 44(2):152–157, 2011.

[20] Lothar Cremer and Manfred Heckl. *Structure-borne sound: structural vibrations and sound radiation at audio frequencies*. Springer Science & Business Media, 2013.

[21] J.D. Cutnell, K.W. Johnson, D. Young, and S. Stadler. *Physics, 10th Edition*. Wiley, 2015.

[22] 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.

[23] Abe Davis, Justin G. Chen, and Frédo Durand. Image-space modal bases for plausible manipulation of objects in video. *ACM Trans. Graph.*, 34(6):239:1–239:7, October 2015.

[24] 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 Trans. Graph.*, 33(4):79:1–79:10, July 2014.

[25] Alain De Cheveigné and Hideki Kawahara. Yin, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[26] Guido De Roeck, Bart Peeters, and Wei-Xin Ren. Benchmark study on system identification through ambient vibration measurements. In *Proceedings of IMAC-XVIII, the 18th International Modal Analysis Conference, San Antonio, Texas*, pages 1106–1112, 2000.

[27] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto. Dynamic textures. *International Journal of Computer Vision*, 51(2):91–109, 2003.

[28] Franz Durst, Adrian Melling, and James H Whitelaw. Principles and practice of laser-doppler anemometry. *NASA STI/Recon Technical Report A*, 76:47019, 1976.

[29] Timothy Emge and Oral Buyukozturk. Remote nondestructive testing of composite-steel interface by acoustic laser vibrometry. *Materials evaluation*, 70(12):1401–1410, 2012.

[30] William M Fisher, George R Doddington, and Kathleen M Goudie-Marshall. The darpa speech recognition research database: specifications and status. In *Proc. DARPA Workshop on speech recognition*, pages 93–99, 1986.

[31] David J. Fleet and Allan D. Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1):77–104, 1990.

[32] Roland W. Fleming, Ron O. Dror, and Edward H. Adelson. Real-world illumination and the perception of surface reflectance properties. *Journal of Vision*, 2003.

[33] T. Gautama and M.A. Van Hulle. A phase-based approach to the estimation of the optical flow field using spatial filtering. *Neural Networks, IEEE Transactions on*, 13(5):1127 – 1136, sep 2002.

[34] Matthias Grundmann, Vivek Kwatra, Daniel Castro, and Irfan Essa. Calibration-free rolling shutter removal. In *Computational Photography (ICCP), 2012 IEEE International Conference on*, pages 1–8. IEEE, 2012.

[35] John HL Hansen and Bryan L Pellom. An effective quality evaluation protocol for speech enhancement algorithms. In *ICSLP*, volume 7, pages 2819–2822, 1998.

[36] Robert W Haupt and Kenneth D Rolt. Standoff acoustic laser technique to locate buried land mines. *Lincoln Laboratory Journal*, 15(1):3–22, 2005.

[37] Selig Hecht and Simon Shlaer. Intermittent stimulation by light v. the relation between intensity and critical frequency for different parts of the spectrum. *The Journal of General Physiology*, July 1936.

[38] Mark N Helfrick, Christopher Niezrecki, Peter Avitabile, and Timothy Schmidt. 3d digital image correlation methods for full-field vibration measurement. *Mechanical Systems and Signal Processing*, 25(3):917–927, 2011.

[39] Yun-xian Ho, Michael S. Landy, and Laurence T. Maloney. How direction of illumination affects visually perceived surface roughness. *Journal of Vision*, 2006.

[40] Jin Huang, Yiying Tong, Kun Zhou, Hujun Bao, and Mathieu Desbrun. Interactive shape interpolation through controllable dynamic deformation. *Visualization and Computer Graphics, IEEE Transactions on*, 17(7):983–992, 2011.

[41] Doug L James and Dinesh K Pai. Dyrt: dynamic response textures for real time deformation simulation with graphics hardware. *ACM Transactions on Graphics (TOG)*, 21(3):582–585, 2002.

[42] Doug L James and Dinesh K Pai. Multiresolution green's function methods for interactive simulation of large-scale elastostatic objects. *ACM Transactions on Graphics (TOG)*, 22(1):47–82, 2003.

[43] AJEM Janssen, R Veldhuis, and L Vries. Adaptive interpolation of discrete-time signals that can be modeled as autoregressive processes. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 34(2):317–330, 1986.

[44] Erik Jansson, Nils-Erik Molin, and Harry Sundin. Resonances of a violin body studied by hologram interferometry and acoustical methods. *Physica scripta*, 2(6):243, 1970.

[45] Nebojsa Jojic and Thomas S. Huang. Estimating cloth draping parameters from range data. In *In International Workshop on Synthetic-Natural Hybrid Coding and 3-D Imaging*, pages 73–76, 1997.

[46] S. Kawabata and Masako Niwa. Fabric performance in clothing and clothing manufacture. *Journal of the Textile Institute*, 1989.

[47] HH Ku. Notes on the use of propagation of error formulas. *Journal of Research of the National Bureau of Standards*, 70(4), 1966.

[48] Siwang Li, Jin Huang, Fernando de Goes, Xiaogang Jin, Hujun Bao, and Mathieu Desbrun. Space-time editing of elastic motion through material optimization and reduction. *ACM Transactions on Graphics*, 33(4):Art–No, 2014.

[49] Ce Liu, Lavanya Sharan, Edward Adelson, and Ruth Rosenholtz. Exploring features in a bayesian framework for material recognition. 2010.

[50] Philipos C Loizou. Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum. *Speech and Audio Processing, IEEE Transactions on*, 13(5):857–869, 2005.

[51] Marci Meingast, Christopher Geyer, and Shankar Sastry. Geometric models of rolling-shutter cameras. *arXiv preprint cs/0503076*, 2005.

[52] Junichi Nakamura. *Image sensors and signal processing for digital still cameras*. CRC Press, 2005.

[53] Dinesh K. Pai, Kees van den Doel, Doug L. James, Jochen Lang, John E. Lloyd, Joshua L. Richmond, and Som H. Yau. Scanning physical interaction behavior of 3d objects. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 87–96, New York, NY, USA, 2001. ACM.

[54] Alex Pentland and John Williams. *Good vibrations: Modal dynamics for graphics and animation*, volume 23. ACM, 1989.

[55] A.P. Pentland and S. Sclaroff. Closed-form solutions for physically based shape modeling and recognition. 13(7):715–729, July 1991.

[56] Edgar Allan Poe. The Raven. 1845.

[57] Javier Portilla and Eero P. Simoncelli. A parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vision*, 40(1):49–70, October 2000.

[58] Robert L Powell and Karl A Stetson. Interferometric vibration analysis by wavefront reconstruction. *JOSA*, 55(12):1593–1597, 1965.

[59] Schuyler R Quackenbush, Thomas Pinkney Barnwell, and Mark A Clements. *Objective measures of speech quality*. Prentice Hall Englewood Cliffs, NJ, 1988.

[60] SJ Rothberg, JR Baker, and Neil A Halliwell. Laser vibrometry: pseudo-vibrations. *Journal of Sound and Vibration*, 135(3):516–522, 1989.

[61] Michael Rubinstein. *Analysis and Visualization of Temporal Variations in Video*. PhD thesis, Massachusetts Institute of Technology, Feb 2014.

[62] C Santulli and G Jeronimidis. Development of a method for nondestructive testing of fruits using scanning laser vibrometry (SLV). *NDT. net*, 11(10), 2006.

[63] Arno Schödl, Richard Szeliski, David H. Salesin, and Irfan Essa. Video textures. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 489–498, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[64] Ahmed A Shabana. *Theory of vibration*, volume 2. Springer, 1991.

[65] Lavanya Sharan, Yuanzhen Li, Isamu Motoyoshi, Shin'ya Nishida, and Edward H Adelson. Image statistics for surface reflectance perception. *Journal of the Optical Society of America. A, Optics, image science, and vision*, April 2008.

[66] P.J. Shull. *Nondestructive evaluation: theory, techniques, and applications*, volume 142. CRC, 2002.

[67] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger. Shiftable multi-scale transforms. *IEEE Trans. Info. Theory*, 2(38):587–607, 1992.

[68] Jos Stam. Stochastic dynamics: Simulating the effects of turbulence on flexible structures, 1996.

[69] AB Stanbridge and DJ Ewins. Modal testing using a scanning laser doppler vibrometer. *Mechanical Systems and Signal Processing*, 13(2):255–270, 1999.

[70] Meng Sun, Allan D. Jepson, and Eugene Fiume. Video input driven animation (vida). In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ICCV '03, pages 96–, Washington, DC, USA, 2003. IEEE Computer Society.

[71] Martin Szummer and Rosalind W. Picard. Temporal texture modeling. In *IEEE Intl. Conf. Image Processing*, volume 3, pages 823–826, September 1996.

[72] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(7):2125–2136, 2011.

[73] Hai Tao and Thomas S. Huang. Connected vibrations: A modal analysis approach for non-rigid motion tracking. In *CVPR*, pages 735–740. IEEE Computer Society, 1998.

[74] Neal Wadhwa. *Revealing and Analyzing Imperceptible Deviations in Images and Videos*. PhD thesis, Massachusetts Institute of Technology, Feb 2016.

[75] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32(4):80, 2013.

[76] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Phase-based video motion processing. *ACM Trans. Graph. (Proceedings SIGGRAPH 2013)*, 32(4), 2013.

[77] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Riesz pyramid for fast phase-based video magnification. In *Computational Photography (ICCP), 2014 IEEE International Conference on*. IEEE, 2014.

[78] Huamin Wang, James F. O'Brien, and Ravi Ramamoorthi. Data-driven elastic models for cloth: modeling and measurement. SIGGRAPH, 2011.

[79] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG)*, 31(4):65, 2012.

[80] Zeev Zalevsky, Yevgeny Beiderman, Israel Margalit, Shimshon Gingold, Mina Teicher, Vicente Mico, and Javier Garcia. Simultaneous remote extraction of multiple speech sources and heart beats from secondary speckles pattern. *Opt. Express*, 17(24):21566–21580, 2009.