Interactive Video for Handheld Camera Captures

Brad Huang
Stanford University
brad0309@stanford.edu

Abstract

This project is an extension of interactive dynamic video using a Lagrangian flow field. We extract and track feature points of objects in the video, then perform modal analysis to simulate object responses to user interactions. We showed that the approach can correct for global motion from handheld camera movements, and with carefully curated feature points, the interpolation scheme can result in some parallax related depth cues. However, the approach also suffers from many artifacts, stemming from the interpolation and the feature point selection, and simulation of large motion still leads to undesired deformation given that vibration modes need to be manually selected.

1. Introduction

The interactive dynamic video is a research project from Abe Davis [4], where they extracted image space modal bases from videos of vibrating objects, and used the knowledge to simulate interaction with the objects. The method they used to extract modal bases is considered Eulerian - optical flow of each pixel in the first frame is computed against every subsequent frames, and the flow vectors are then used to perform modal analysis.

2. Related Works

The most related works would be in the fields of structural analysis and video magnification.

Observing Vibration Modes: Directly observing modal shapes from videos has been a topic discussed widely in the optical and mechanical engineering society [6, 11]. In addition, there has also been research in computer graphics and vision that seeks to identify [7, 2] vibrational modes for the purpose of motion magnification and visualization. Our approach, as extensively investigated in Abe’s thesis [3], focuses on modal shapes that gain knowledge of object properties in the scene and that are best for simulation of object responses.

Motion Magnification: Recent publications in motion magnification of object motion [7, 2] have demonstrated the ability to extract and magnify motion from observations. Interactive dynamic video, as in [4] and our project, is different since by learning about vibration modes from videos, we are able to synthesize motion in response to different forces that were not present in the videos to make the final result interactive. Elgharib et al. 2015 attempted to use dense Lagrangian approach for motion magnification, but their method did not include modal analysis or interaction, but simply magnified existing motion [5].

3. Methods

Figure 2: The analysis flowchart.

3.1. Video Capture

Examples in this project are all captured with a handheld iPhone 6S using slow motion with 240 FPS, except for the reference wireman example, which was captured with a high speed camera stabilized on a tripod.

3.2. Feature Tracking

We use the feature points extracted from Good Features (Shi and Tomasi 1994), and then use OpenCV’s implemen-
tation of Lucas-Kanada method for sparse optical flow to track the feature points through the video.

Before tracking the detected feature points, we manually add fixed points to the reference frame. The fixed points are not tracked. At simulation time, each fixed point moves the average amount of its neighboring detected feature points, and the points without neighboring detected feature points would stay fixed during the simulation.

Figure 3 shows how the feature points should be arranged - the blue points are the detected feature points, while the red points are the manually added points. The manual points surround the detected points, and together mark the fixed background of the reference frame.

![Figure 3: Triangulations of the example captured videos.](image)

### 3.3. Modal Analysis

The changes in x and y direction of each feature points are known as the optical flow vectors, and are used to extract modes of the feature points. Since the optical flow vectors are the deformations of the feature points from the reference image and are the sums of vibrational modes at different frequencies, performing a Discrete Fourier Transform (DFT) on the time series would allow separating modes of different frequencies. This introduced the assumption that the vibrational modes of different frequencies are independent, which usually holds for physically modeled vibrations, but may be violated since we operate only in image space.

The modal shapes are visualized in Figure 4. The phase of the modal shapes at each feature point is converted to hue in the HSV color space, and the magnitude is converted to the value. Since we separate X and Y optical flow vectors, each modal shape would have two images and two values, one for deformations in the X direction and one for Y.

![Figure 4: Representative modal shapes have color distribution correlated with structures of the feature point region.](image)

In order to use Lagrangian tracking to correct for global motion, we chose to compute the global motion as average of all flow vectors for a given frame, and subtract the global motion from the feature point flow vectors.

### 3.4. Simulation

Simulation of the image-space dynamics is based on the computed modal shapes from the modal analysis. For each feature point i and frequency f combination, there are two complex values, $x_{i,f}$ and $y_{i,f}$ for the modal shape. In addition, a global modal coordinate for each mode $g_f$, which damps exponentially through time, is used to keep track of how energetic the corresponding mode is. The displacement of each feature point at a certain point at time t is then given by:

$$dx_i(t) = C \sum_{f \in \mathcal{F}} mag(g_f) \times mag(x_{i,f}) \times \sin([sf + ph(g_f)+ ph(x_{i,f})])$$

where $mag(c)$ is the magnitude of the complex c, and $ph(c)$ is the phase of the complex c. Similar equation holds for $dy_i(t)$. Note that since the displacement is unitless, and the complex values’ magnitudes may vary widely between different scenes, we need to multiply a scalar $C$ to the displacement to fit the scene.

### 3.5. Interaction

Interacting with the objects is modeled through exciting certain modal shapes. Pressing and dragging the mouse in the scene would temporarily stop the simulation and displace the feature points to match to drag. First, we find the enclosing triangles of the clicked position, and use barycentric coordinates to average the modal shapes of the vertices at every frequency to get $cx_f$ and $cy_f$ as the modal shape at the point of the click. With the dragging displacements $d_x$ and $d_y$, we calculate the excitation of mode f as

$$E_f = cx_f * d_x + cy_f * d_y$$

This equation essentially projects the dragging displacements onto the modal shape, similar to how dot products between force and modal shapes are used in other vibration models. While the mouse is still pressed, simulation is paused and the global modal coordinate $g_f$ is set to be equal to $E_f$. When the mouse is released, the simulation restarts from the set modal coordinates.

### 4. Results and Discussion

The results are more easily observed from videos that I have attached with the project. Here I show several deformations captured in the simulation.

![Figure 5. Wireman deformation. Left is the reference and right is the deformed look.](image)
The wireman example is what Abe used in his previous re-
search [4], and is an well-formed example with more sta-
bilized camera, uniform background and a simple physical
model that is mostly flat. The resulting simulation is thus
much more plausible. The other examples are simulations
from handheld camera captured videos. Although the vi-
bration modes observed from these examples are promising
and do show how global motion does not affect the simula-
tion, there are still various artifacts.

4.1. Feature Point Distribution and False Coupling

For the indoor scene of flowers in Figure 5a, the feature de-
tector picks up the edges of the flowers well, and the track-
ing is accurate. As a result, the foreground flowers’ mo-
tion is simulated well, with how the flowers have indepen-
dent movements from each other, but also influences neigh-
boring flowers more than others. However, there are also
other feature points being detected in the background due
to the apparent and contrasting edges in parts of the back-
ground. This leads to undesired background deformation
when the flowers move. Most modes for the feature points
in the background are not as strong as those of the feature
points for the flowers as seen in Figure 4. However, since
the background feature points do not cover the background
objects as a whole (for example the painting on the wall),
background objects are distorted.

For the strings example in Figure 5b, the background de-
formation is reduced due to the more uniform background,
although it is still observable around the strings. A more
conspicuous artifact that can be observed in the video is
that all the hanging strings would move together when only
one of them is pulled. This is due to false coupling of the
modes - since we do not separate feature points belonging
to different parts of the object, all feature points move when
a mode is excited. Since the hanging strands would oscil-
late together when the external force stops and the motion
propagates through the network, the hanging strands are ob-
served to be mostly vibrating together, and thus leads to the
false coupling.

The above two problems can be solved by categorizing
feature points into different groups based on the motion pat-
tern. For example, background feature points can be iden-
tified as points that only move along with global motion,
so fixating background feature points is possible as long as
we can separate them from foreground object points. On
the other hand, with different components in an object, the
object would respond with quite different vibration modes
when forces act on different parts of the object. For ex-
ample, in the strings example, if separate videos of pulling
on different strands can be combined to analyze and realize
the fact that hanging strands do not vibrate together, sepa-
rating components of an object then becomes possible, and
how the vibration propagates through the network can be
deduced with more in-depth analysis.

4.2. Approximating Depth Cues

However, modal bases in image space still suffer from in-
formation loss, since the image space modal bases make the
assumption that linear motion in 3D is projected to be linear
in 2D. Since the analysis does not consider depth informa-
tion, large motion in the depth direction and structure span-
ned large depth ranges would lead to out of focus blurring,
depth resizing and parallax effects. Some of them can be ap-
proximated by interpreting the motion as vibration modes,
as the front flower petals get larger when pulling the flower
down, or the petals on the side are larger then pulled to the
other direction. However, the resizing effect undesirably re-
duces the size of the petal closest to the hand, which would
in reality appear larger when the flower is pulled towards
the camera. The difference is that the particular petal is
more parallel to the viewing direction and the image space
modal bases would not be able to correctly pick up its mo-
tion, while the other petal surfaces are relatively vertical to
the view direction, allowing their motion to be correctly ob-
served.

5. Conclusion

We explored the Lagrangian approach to interactive dy-
namic video and discovered that the method does extend the
original Eulerian approach to deal with global motion and
larger motion well, but the automatically detected features
points lead to different background artifacts. A natural follow up would then be categorizing and identifying objects based on the learned vibration modes in order to decrease these artifacts. Nevertheless, the loss of information from 3D to image space still makes the method suffer when the objects make large depth movements.

References