James Tompkin | Research Statement

jamestompkin.com

Please watch the complementary video overview of my research:
www.jamestompkin.com/research.mp4

Video best represents how we visually perceive, and with ubiquitous cameras everyone can capture video of our world. However, our tools to edit, explore, and interact with video often require laborious work or expert skill, which makes video merely a consumption medium for most people. With advances in video understanding through computer graphics, vision, and interaction, my goal is *to reinvent how we think about video*: First, to transform video from a rigid and inaccessible medium into a malleable and creative one, so that non-experts can make sophisticated content-based edits. Second, to transform video from a linear medium viewed sequentially into novel interactive content-based and context-aware visual explorations. Third, to transform video from passive observation into an essential tool for real-time scene understanding to help us interact with the world.

Advanced Video Editing

Video is a rigid medium because it is hard to change the content within. Even for experts, content edits often require days or weeks of work. I aim to change this—video should be a creative medium for everyone.

Content Appearance Editing the appearance of objects within video is often hard due to complex illumination. I dramatically simplified this process by decomposing video into illumination and object color, which enables light-aware coloring, texturing, and compositing of objects. However, naively applying this kind of image filtering to videos frame by frame often results in flickering, so temporal consistency is key. To remove the need to write video versions of every new image filter that is invented, I generalized temporal consistency to treat image filters as 'black boxes', allowing the removal of flickering as a post process. This significantly expands the available number of filters for video. [SIGGRAPH Asia 2014, 2015]

Content Removal Another common problem is unwanted objects in the frame. I remove them automatically by filling in the hole with pieces of any shape or size from elsewhere in the video. This content correspondence allows us to recreate both the static background and the *dynamic* motions of occluded objects behind, e.g., to produce an unoccluded view of street musicians playing by removing a passerby. [Eurographics 2012; ECCV 2012]

Future Research These lightweight models for 'pixel shuffling' can remove some of the rigidity from video, but they are often insufficient for complex scenes. One option is heavyweight models of geometry, material, lighting, and motion, but these are even less approachable for novices.

Intrinsic Video Decomposition.. Separating a video *(top)* into illumination *(mid-left)* and object color *(mid-right)* allows us to realistically retexture the brick walls of the house *(bottom)*.

Video inpainting of dynamic objects. An occluding passerby is removed, and the unknown motions of the musicians behind are created by seamlessly recombining other parts of the video.

How might we create simple but powerful editing representations? Take interactive segmentation: raster models are imprecise with edge artifacts, so professional VFX artists use laborious hand-created vector models for editing. I aim to develop automatic vector region segmentations to reduce the barriers to high-quality video editing, and so democratize the skill.

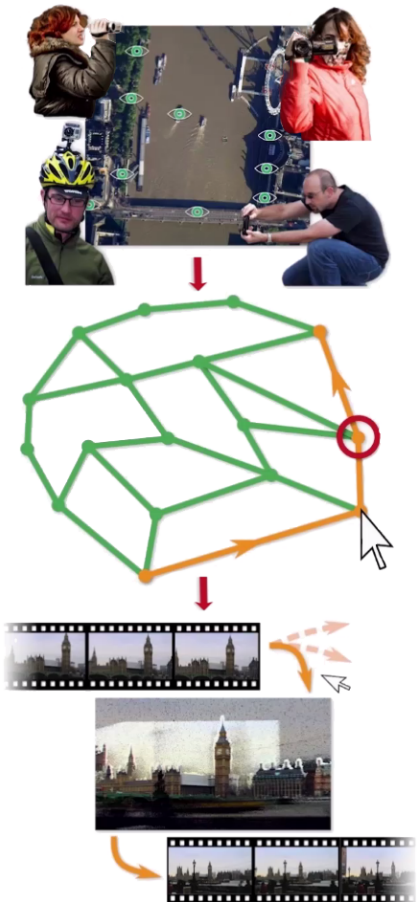
Exploring Video Collections

On websites like YouTube, we explore video collections through text and metadata connections, because it is hard to discover and expose the visual connections between content within videos. I develop systems to change video into a visual hypermedia with novel intuitive exploration interfaces.

Visual Connections With *Videoscapes*, I showed how to build a graph of connections from a large video collection, such as between places or events within crowd-sourced tourist footage of a city. This requires efficiently analyzing millions of video frames, and developing graph clustering and prediction algorithms to find high-accuracy visual connections. For seamless graph exploration, I synthesized smooth transitions between videos from vision-reconstructed world geometry. As partial reconstruction failures cause visual artifacts, I evaluated artifact perceptual preference to be able to show the most suitable transition. For intuitive exploration of spatial connections, I exploited the graph to resolve the geographical locations of video contents for interactive maps and tours. This work shows how applying robust methods to large-scale heterogeneous video can create new tools to explore our world. [SIGGRAPH 2012, ECCV 2012; TAP 2013]

Spatial Contexts Directly comparing events within videos is hard when videos are viewed one after another; however, providing direct spatial context changes the perspective to make this easy. I developed a system to allow users to collectively embed smartphone videos into panoramic image contexts like Google Streetview. This allows detailed comparison of places over time, for instance, seasonal changes, or viewing a protest in a plaza. I devised new interactions for spatial and temporal searches of video collections within contexts, and created tools to analyze events across the collection. Viewing and interacting is flexible across desktop, spherical, head mounted, and tablet displays: Taking a tablet to a real-world place creates a *window into the past*, where rotating the tablet at arm's length discovers events previously captured by other people. [UIST 2013]

Hollywood Augmented To show the power of context, film-maker Jeff Desom and I reinterpreted Alfred Hitchcock's 1954 movie *Rear Window*. Hitchcock never shows the wide view from the window, but we are able to spatially reconstruct the murder mystery within a panoramic context. Then, we turn the passive viewer into an active voyeur by building pan and zoom control into an antique camera—figuratively and literally changing perspectives on art! [Museum of the Moving Image, New York City, 2015]



Videoscapes: Crowd-sourced videos are mined to build a visual connection graph, which can be walked with seamless geometry-based transitions.

Top: Four videos (white outline) are embedded into a panoramic context, showing four seasons of changes and dynamic events at once. *Bottom:* A tablet used as a time machine window to see events from the past in a real world context.

Future Research As more and more of our media is connected within collections, contexts will become the critical element which helps us derive information from massive content. Imagine a historical and live spatio-temporal registration of *all* imagery—the so-called 'one world model' idea. How might we efficiently build this representation of the world? How might we summarize and intuitively explore the millions of events within? Videos exist within other contexts, too, such as motion, social, narrative, or even artistic. I aim to build adaptive interfaces which redefine typical linear presentations into powerful fluid, non-linear exploration tools.

Video for Interaction

Real-time scene analysis from video can be used directly to develop new fundamental techniques for interacting with the digital and physical world.

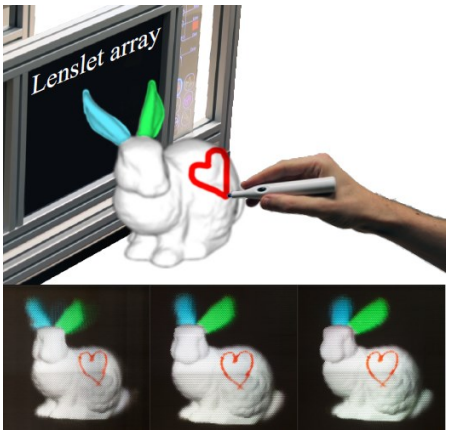
Light Field Interaction We can work with 3D content in flat perspective, but how might we work in 3D directly? Light field displays give the illusion of a real 3D scene—glasses-free binocular stereo and motion parallax—but light field interaction is nascent. I developed a pen-based light field system with a joint display and sensing optical path. This deduces 3D position and 2D orientation of an infrared pen to human hand accuracy by analyzing 40,000 tiny camera images through a lenslet array at 150 Hz. This allows artists to draw directly in free space with light—to *sculpt* with light.

For the lenslet array, I developed a technique to maximize the joint spatio-angular resolution for a given pixel budget by optimizing lens shape, arrangement, and content. To produce these custom optics, I 3D printed spatially-varying lenslet sheets. This allows locally-curved image surfaces and embedded baffles to reduce cross-talk. [SIGGRAPH 2013, UIST 2015]

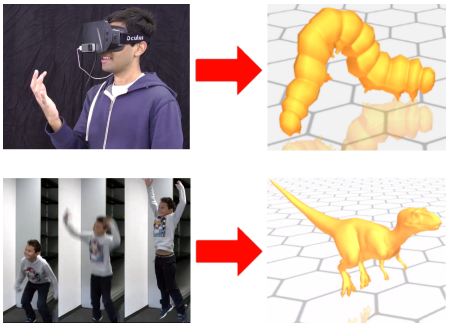
Wave Gestures Recognizing hand and body gestures from tracked video allows natural interaction in environments like virtual reality. However, it is hard to reliably estimate continuous motion parameters, e.g., instantaneous frequency, amplitude, and phase, especially for multiple simultaneous gestures. I developed a method which, given single examples of control gestures, can generalize robustly to different motions, simultaneous motions, and different body shapes and sizes. This simplifies real-time control of virtual characters, such as hard-to-control caterpillars, and simplifies control of real-world robots. [Eurographics 2014; SIGGRAPH Asia 2015]

Future Research Natural interaction will increasingly rely on vision. Take the control of *transparent cameras*, e.g., in the environment, perhaps hidden from view, and without displays. How might we control a set of distributed cameras, and then synthesize the view from the user's eyes? I wish to couple advanced gestural sensing with scene analysis and visual post-processing to ensure that what is captured is exactly what the user intended.

Rear Window Augmented, with Jeff Desom. Panning and zooming the antique camera turns the user into Hitchcock's most famous voyeur.



Light Field Painting. Detecting an infrared pen through a lenslet array allows us to sculpt with light *(top)*, and view the output naturally with binocular stereo and parallax effects *(bottom)*.



Wave Gestures. Robustly estimating the frequency, amplitude, and phase of multiple simultaneous gestures is important for the widespread adoption of virtual reality, such as for the control of virtual characters like caterpillars.