<JIAMIN: Motion editing is hard.>

Thank you, Jiamin. Jiamin just talked about how to manipulate segments of videos to create ‘moment images’. Now, sometimes to create these moments, we capture footage specifically. But, what if we had arbitrary footage and wanted to explore it to find moments in a video? How can we explore within a video, to find content, in creative and imaginative ways?

Now, we know that videos don’t exist in isolation of other videos or in isolation of the real world, so let’s expand our question: how can we explore video, or potentially multiple videos, through the contexts in which they were captured?

So, hello SIGGRAPH, I’m James Tompkin - a post-doc at Harvard – and I’m going to try and show you some of the interesting answers to these questions that our community has developed, which use graphics, vision, and interaction techniques. Now, while my slot isn’t quite a graveyard shift, it is the last slot of the conference, so I’m going to try and keep the conversation light and easy to follow.

[All right everyone – you made it this far – and I’m very grateful you chose me to be the last speaker of your SIGGRAPH.]

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Breadth of work

- Lots of work in this field.
- Picking a few of my favourites; apologies to authors if I've not included or overlooked some works – lots of good work out there.

- Unfortunately, must leave out all the work from ACM Multimedia
  - Sorry; too much!
  - Mostly a review from graphics and interaction communities.

2015-08-13 James Tompkin – User-Centric Computational Videography 2

So I am very excited to present works in this area – I think it's a very creative and interesting area of video processing, with potentially large impact. There are lots of works that I really like, but unfortunately I only have time to show a few. There are a few more in the notes that will go online.

I'll also mostly be covering work from the graphics and interaction communities, and must I'm afraid leave out great works from other areas.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Build upon explained techniques

- Many of these works build upon existing-mentioned methods
  - Pixel correspondence / optical flow / PatchMatch
  - Segmentation / matting
  - Seamless compositing with graph cuts
  - Structure from motion / geometry reconstruction
  - Lighting estimation
- Lots of these techniques are *not real time (i.e., slow; some are slooooow).*
  - Require significant pre-processing.
  - Always a trade-off between benefit and processing time.

Now, lots of the techniques build upon at least one of the previously-mentioned approaches, so forgive me if I skip over a few details and instead aim to give you the broadest view of this area.

<ADVANCE>

Finally, before we begin, I’m going to ignore computation time in the presentation of these techniques.

Some of them are slow, and some of them are very slow – and there’s always a trade-off between the benefit and the processing time – so see if you can guess which ones.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Exploring a video (singular)

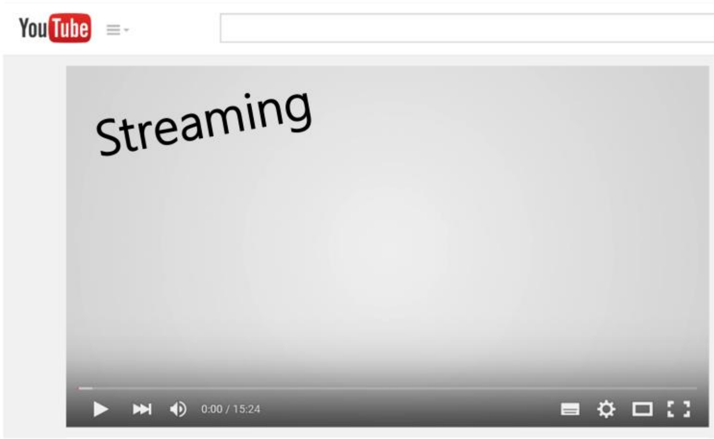
So, onwards. How do we explore videos?

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Typical video exploration - consumption

[VLC Media Player]

- Timeline / play / fast forward / fast back
- Shuffle / repeat

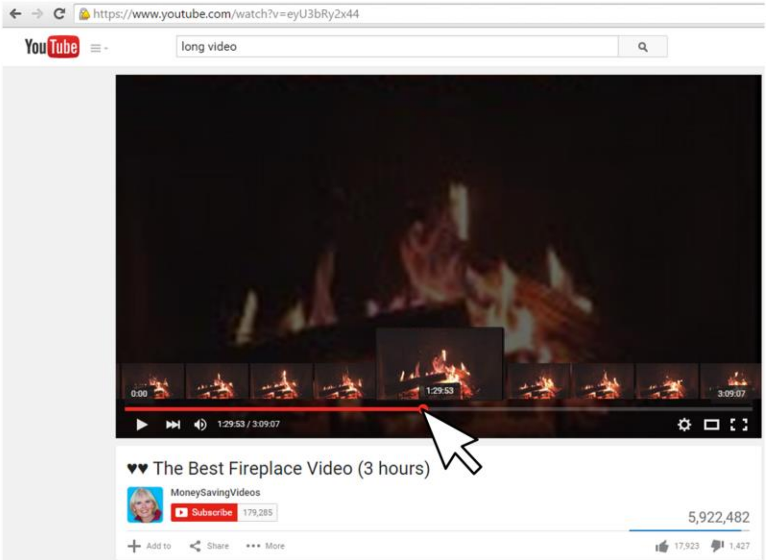


[YouTube.com]

- Timeline / play / skip
- Aspect ratio / fullscreen

YouTube video interface

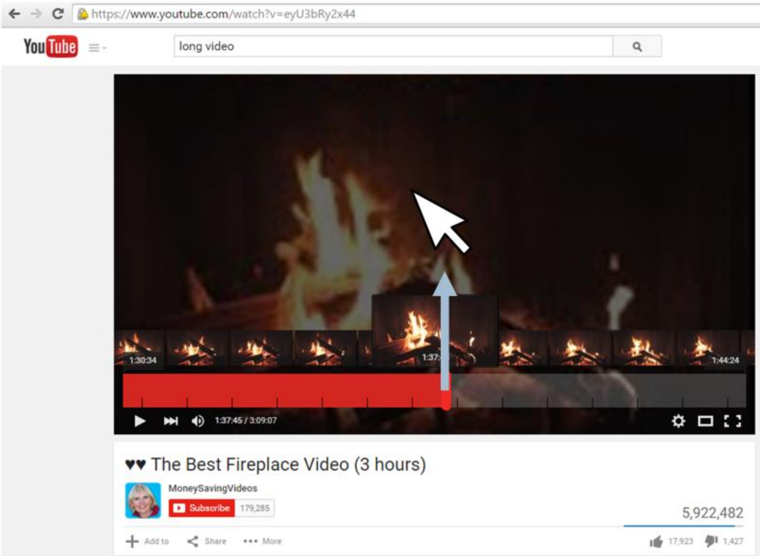
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences



- Many small thumbnails in a row on drag.

YouTube long video interface

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

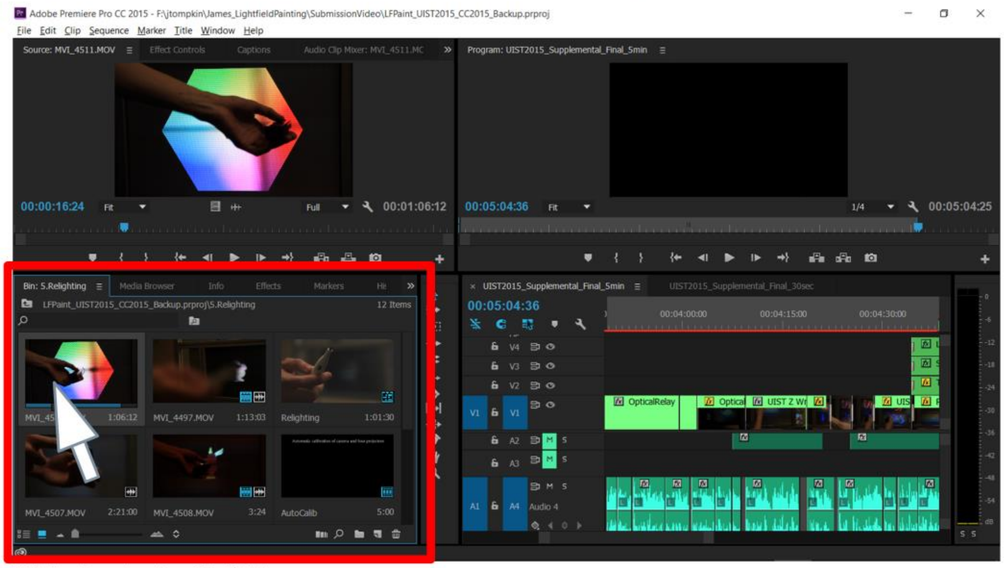


- Many small thumbnails in a row on drag.
- Precision available by dragging up.
- Bar expands, and we get per-second control.

…and for my favourite long video – 3 hours of fireplace – we can get extra precision my dragging up in y – so that I find just the right flame flick to woo my lover.

Typical video exploration - editing

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

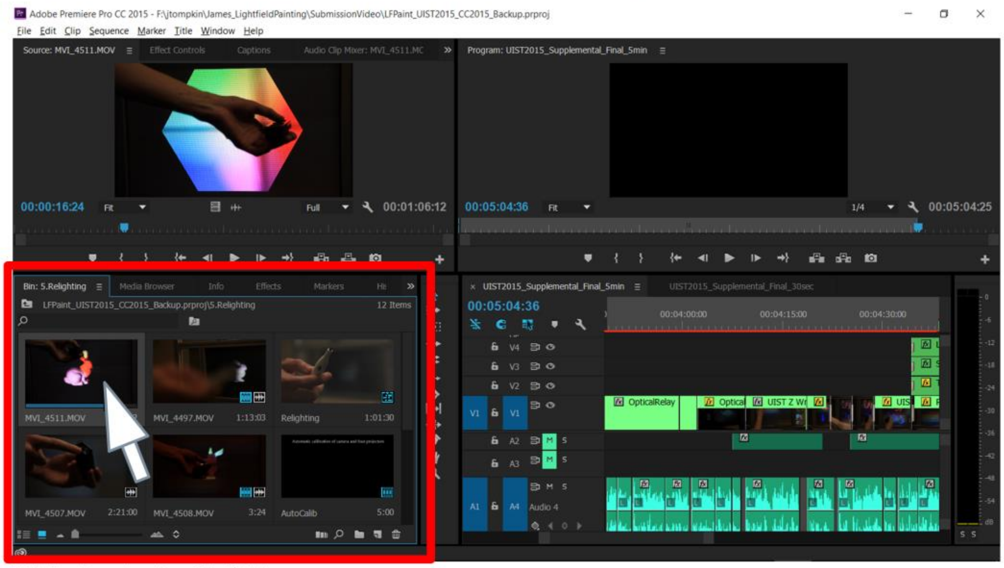


- Mark in / out
- Same thumbnails
  - Now live!
  - Scrubbing

[Adobe Premiere Pro CC 2015]

In the editing case, there’s a greater emphasis on exploring individual videos to find footage, so that we can trim and place the right shot on the timeline. As such, our thumbnails can now be scrubbed and will update as we mouse over them.

Typical video exploration - editing



[Adobe Premiere Pro CC 2015]

- Mark in / out
- Same thumbnails
  - Now live!
  - Scrubbing

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Typical video exploration – editing (timeline)

2015-08-13 [Apple iMovie] James Tompkin – User-Centric Computational Videography 10

This thumbnail idea extends down to the low end of editing, too, and here we see how iMovie lays out the thumbnails.

Temporal exploration modes

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

- Video skims / image summaries
  [Truong, B. T., and Venkatesh, S. 2007. Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. 3, 1, 3.]
- Video surrogates
  [Schoeffmann, K., Hopfgartner, F., Marques, O., Boeszoermenyi, L., & Jose, J. M. (2010). Video browsing interfaces and applications: a review. SPIE Reviews, 1.]
- Timeline (+scrub)
  - Fast seek for larger discrete jumps in time
- Fast forward/backward
  - Fast scan/review for continuous playback at speed

So, if we mark out these typical temporal exploration modes, we have

Timeline, for fast seeking, and we can think of scrubbing the thumbnails as the same kind of operation

…and we have..

Fast forward/backward, for fast review

In general, idea of video surrogates: objects/items which represent the video/sections of video, to allow the user to make a decision about viewing/not viewing the video.

Let's think about the timeline for now.

The iMovie thumbnails are a layout of frames from the whole video at once.

What if we just tried to lay out the whole video timeline as video, in the same block-like structure?

[Song, Y., & Marchionini, G. Effects of Audio and Visual Surrogates for Making Sense of Digital Video. ACM SIGCHI 2007.]

Swift Timelines

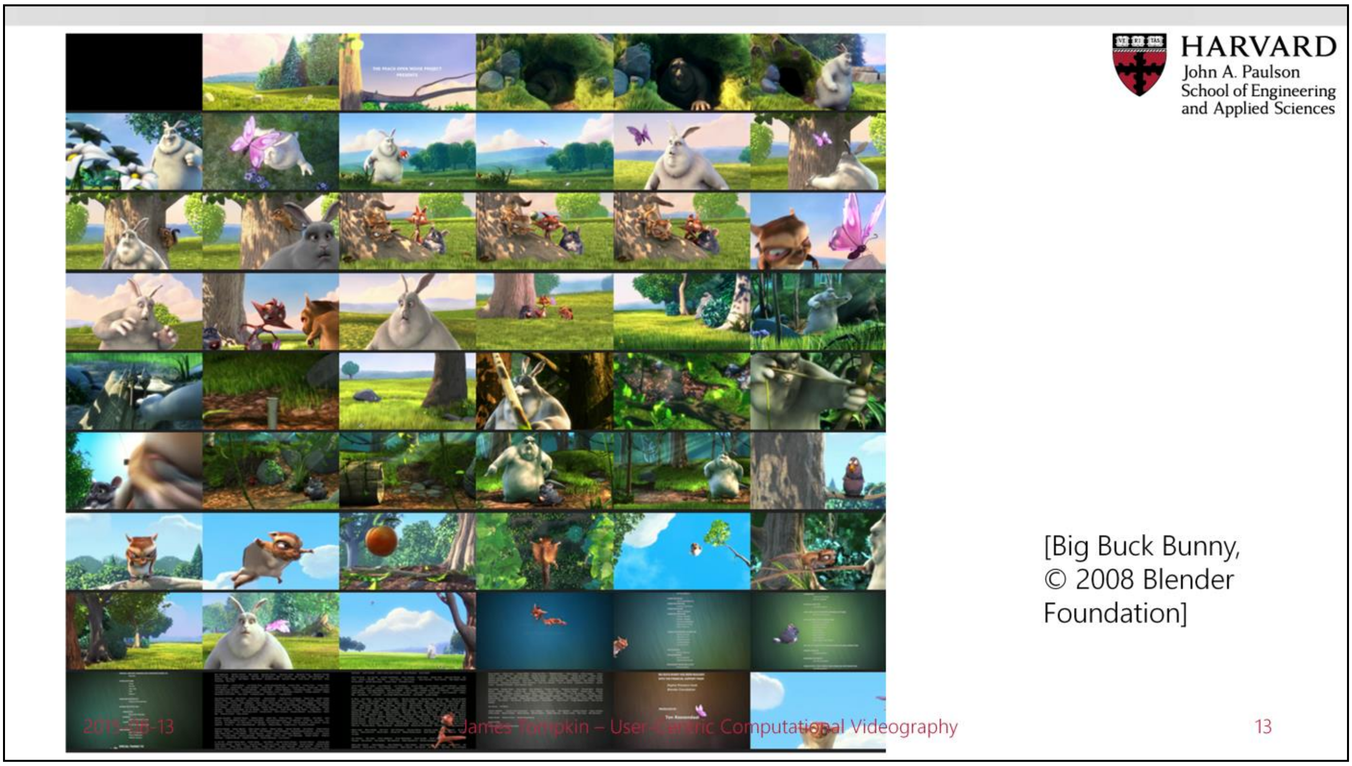
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

[Matejka, J., Grossman, T., & Fitzmaurice, G.

Swift: reducing the effects of latency in online video. ACM SIGCHI, 2012.

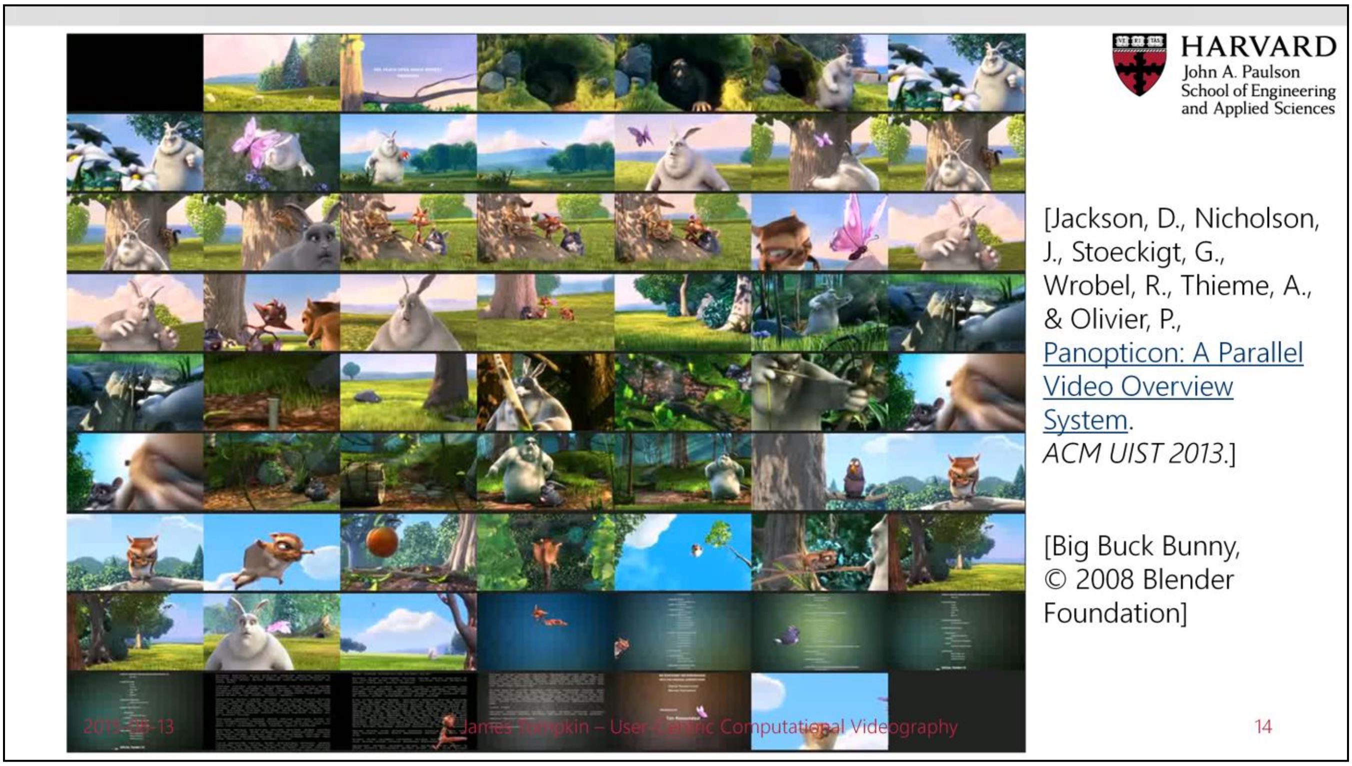
Swifter: Improved Online Video Scrubbing. ACM SIGCHI 2013.]

Looked at different layout options and movement options (page/row at a time, continuous scroll, but always vertical motion)



… and if we apply this to the whole video and try to fit it into one screen or slide, it looks something like this.

We could index into this with the mouse, but we still only see still frames. We could make these videos play, like a wall of videos, but is there a way to increase our ability to search in this format?



Each block of video is offset in time by a fixed amount, so that the whole video fits into a screen's size in little blocks.

Easy to fit an entire 10-minute video in 10-second chunks on the screen.

The blocks are not fixed in position; they move at a rate which ensures that, if you follow with your eyes from the top left block, by the end of the video it will be at the bottom right. There's one column of blocks duplicated at the edges to allow your eyes some leeway in changing rows.

Benefits: we can see the whole video at once, and we can scan the video like we scan the page of a book, looking for interesting parts or things we remembered from a past viewing or from a shoot.

We can jump to approximate times, because we know that each row represents a fixed amount of time.

For longer videos, either the blocks get smaller, or the length of time of each block increases. Both have drawbacks. However, this kind of visual index for video works pretty well for around 30 minutes of video or so, where each block would be approximately 30 seconds.

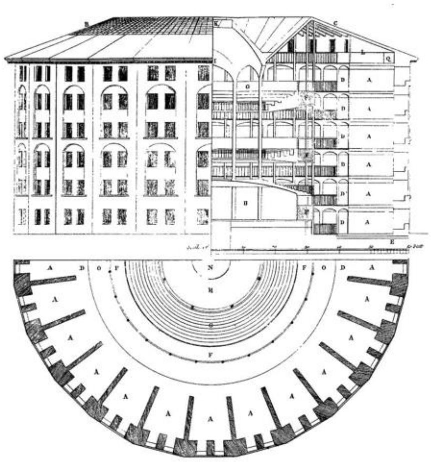
You can imagine extensions of this where cut detection is run on a movie, and each block is a scene. Some of the elegance is lost because timing is now not uniform, but you could potentially squeeze a longer video into the space.

[Jackson, D., Nicholson, J., Stoeckigt, G., Wrobel, R., Thieme, A., & Olivier, P. (2013). Panopticon: A Parallel Video Overview System. *In Proceedings of UIST 2013*.]

https://openlab.ncl.ac.uk/things/panopticon-2/

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Panopticon Historical Interlude



[The works of Jeremy Bentham vol. IV, 172-3]

"A mill for grinding rogues honest."

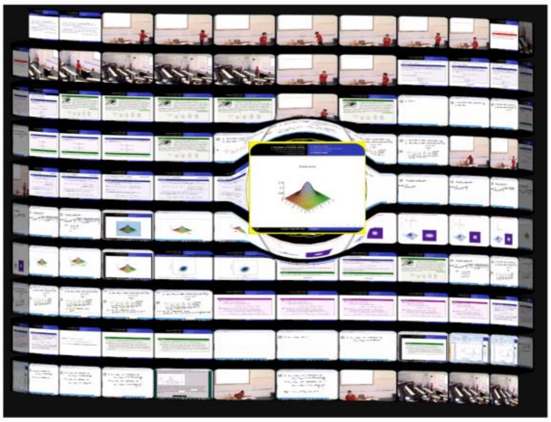
- J. Bentham

Now, the name Panopticon, meaning "to see all", comes from a prison designed by the 18th century English philosopher Jeremy Bentham, in which a single watchman could observe all prisoners - "a mill for grinding rogues honest".

Fortunately for us, Dan Jackson, James Nicholson and co-authors decided to use their technique not for evil, not to torture prisoners with 24-7 endlessly looping videos of Big Buck Bunny, but for good.

Fast seek alternative - Panopticon

- Follow-up CHI 2014 work
  - Application to eLearning as a video search tool



[Nicholson, J., Huber, M., Jackson, D., & Olivier, P. Panopticon as an eLearning Support Search Tool. ACM SIGCHI 2014.]

They showed with a follow-up study that it can be an effective tool for finding content in slide style lecture videos. [but not talk-style lecture videos, where the content is more audio]

In this case, they add a focus+context lens distortion effect to the video panopticon index.

[Significantly faster seeking with users than YouTube (4x faster) or than with static variant (1.15x faster)

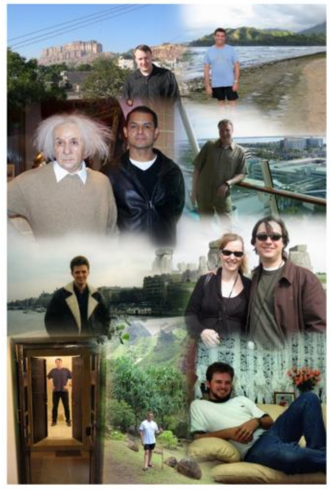
Narrative / unedited / surveillance footage tests

Eyetracking comparisons show that moving video easier to seek than scrubbing, as less eye seeking across board.]

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Collages and Tapestries

[Barnes, C., Goldman, D.B., Shechtman, E., Finkelstein, A., Video Tapestries with Continuous Temporal Zoom, ACM SIGGRAPH 2010]



[Rother, C., AutoCollage, ACM SIGGRAPH 2006]

What happens if we try to make this visual timeline more abstract?
Some works have been inspired by older art forms of narrative tapestries.

Also inspired by color scripts – used to demonstrate and be a reference for the color look of a film

Bi-directional similarity – must be completeness and coherence to input video.

[SIMAKOV, D., CASPI, Y., SHECHTMAN, E., AND IRANI, M. Summarizing visual data using bidirectional similarity. In CVPR 2008.]

In this case, for the tapestry, we pick keyframes.

Pick series of keyframes from input video (which maximizes bi-directional similarity at the frame level)

Retarget these using a patch-based bi-directional similarity to merge the images by 75% to remove redundant info.

Special compensation for faces to make sure these are kept correct.

For hierarchy for zoom, they identify and track islands of imagery common to both max and min zoom levels.

Synthesizing reasonable imagery to fill in gaps between these islands using PatchMatch – this again can be thought of as regargetting the max zoom level to a smaller size, with some constant island regions from the min zoom level already in place.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Video Tapestries – how it works

[Barnes, C., Goldman, D.B., Shechtman, E., Finkelstein, A., Video Tapestries with Continuous Temporal Zoom, ACM SIGGRAPH 2010]

2015-08-13 James Tompkin – User-Centric Computational Videography 19

Here’s how it works in display.

Keyframes are chosen by maximizing the coverage and minimizing the redundancy in the content of the keyframes chosen – called bi-directional similarity.

User study vs. thumbnail frames and timeline scroll bar

- Find important events in user-known films
- No significant difference in task performance
- Preference (~75%) for tapestries
- Used Matrix, Star Wars films

But….how to play videos in tapestry? Sequence under mouse could play. ??

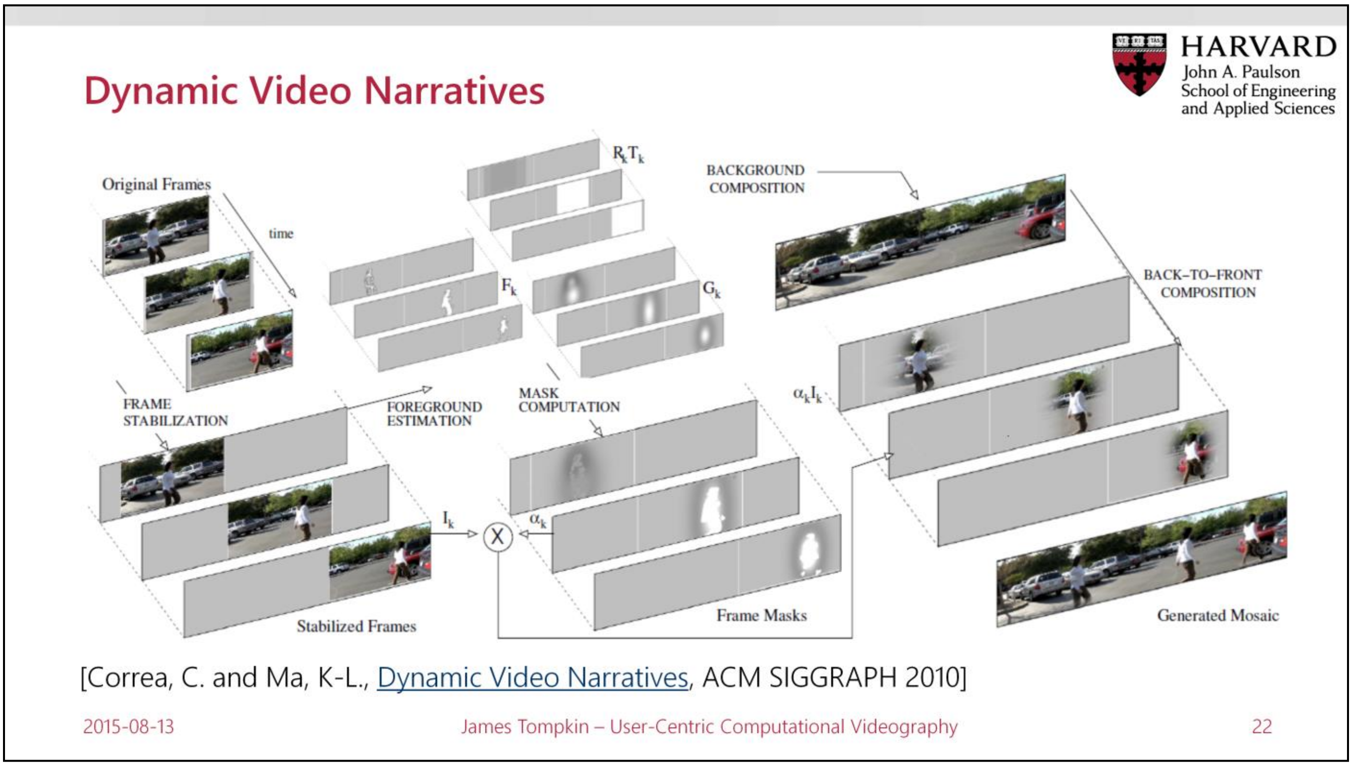
Dynamic Video Narratives

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

[Correa, C. and Ma, K-L., Dynamic Video Narratives, ACM SIGGRAPH 2010]

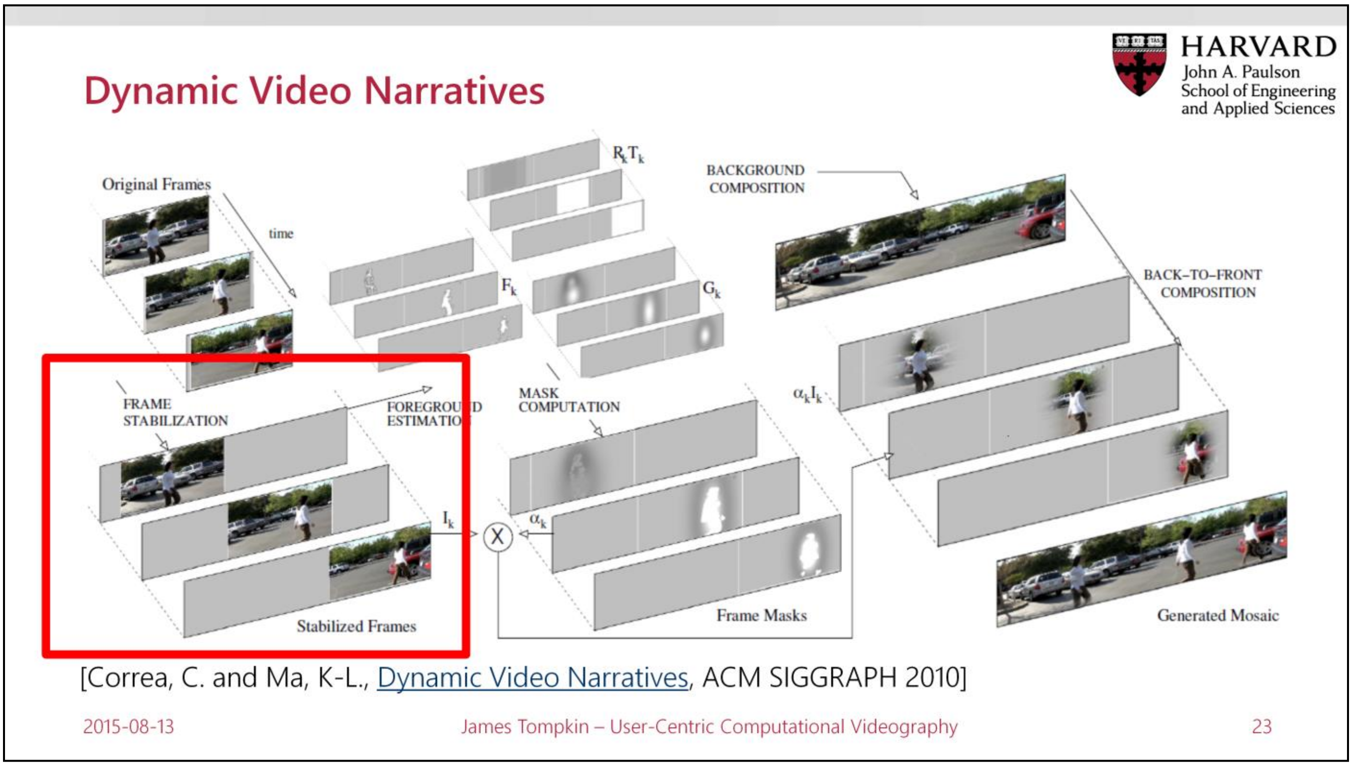
2015-08-13 James Tompkin – User-Centric Computational Videography 21

Making video mosaics dynamic and interactive.



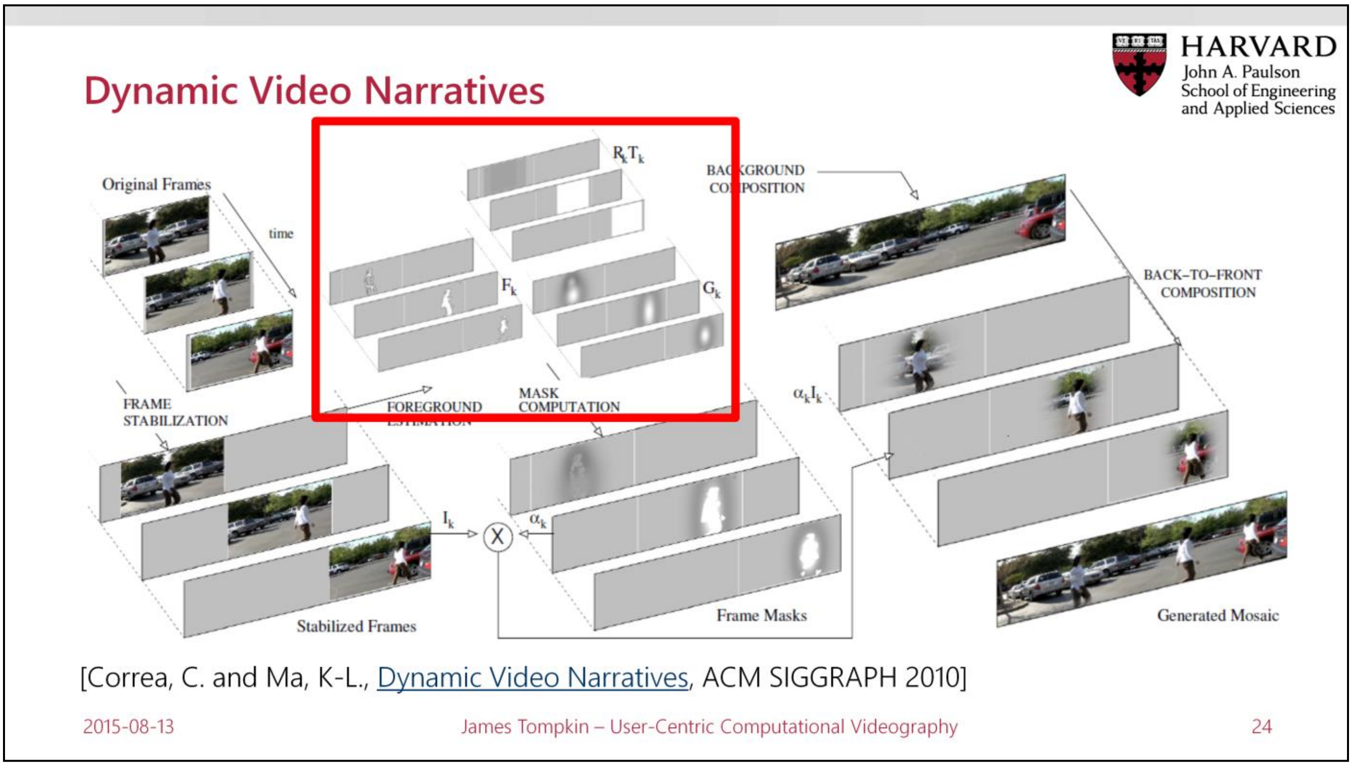
LIGHTWEIGHT MODEL

How it works:



How it works:

Stabilize the pan in a video shot using SIFT features

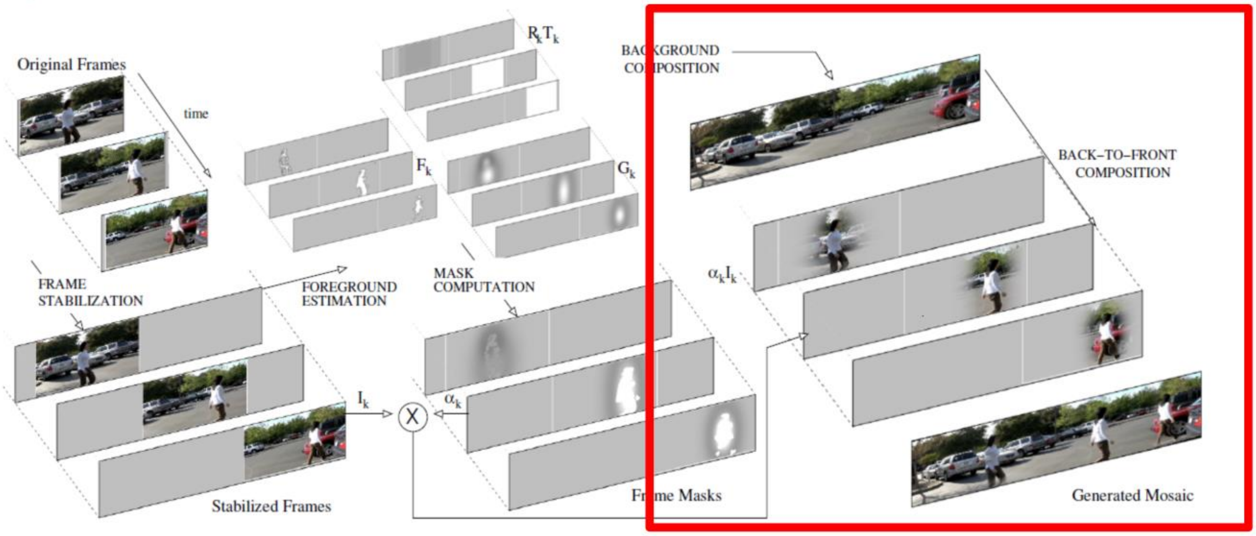


Try to extract the foreground regions in the shot by using Gaussian Mixture Models to estimate the background ([Kaewtrakulpong and Bowden 2001])

A little bit of smoothing and clean-up of the masks.

Dynamic Video Narratives

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences



[Correa, C. and Ma, K-L., Dynamic Video Narratives, ACM SIGGRAPH 2010]

Simple multiplication to recover the layers.

Dynamic Video Narratives

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

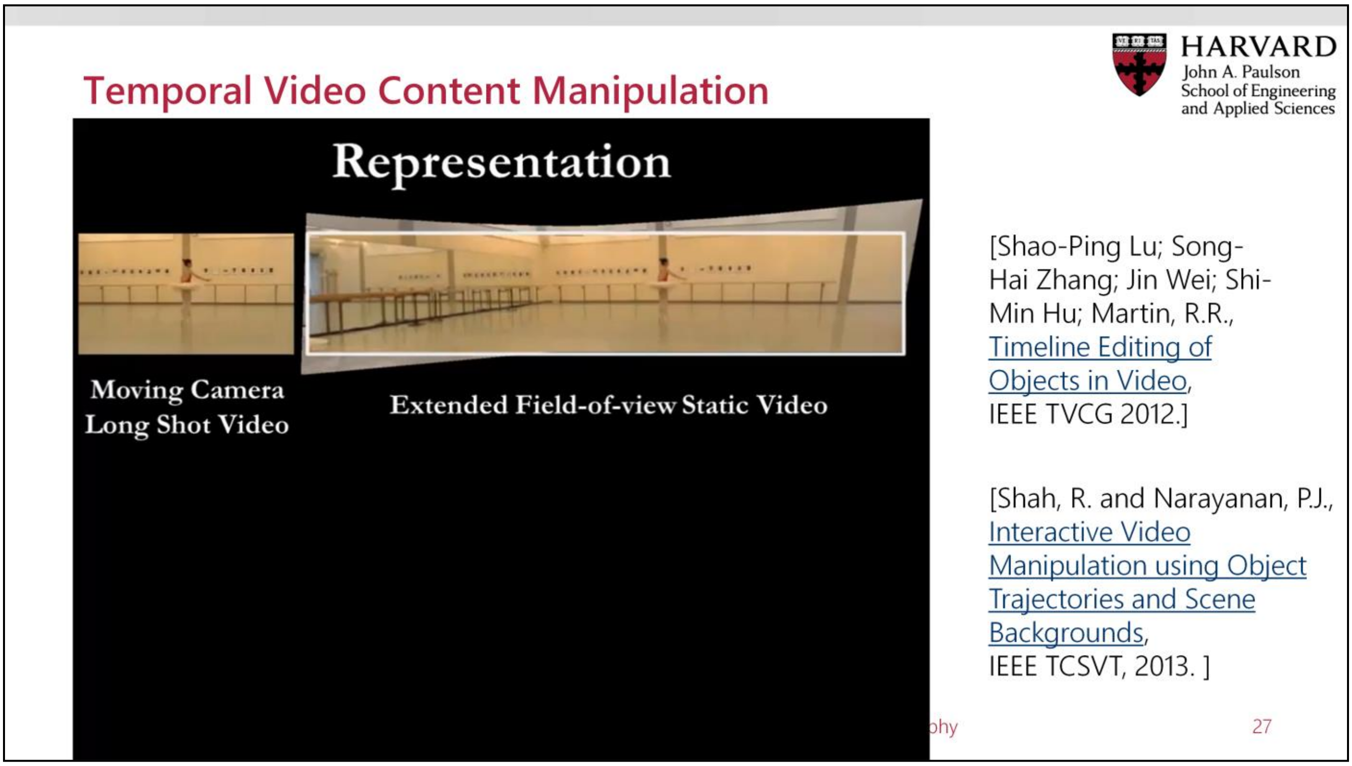
[Correa, C. and Ma, K-L., Dynamic Video Narratives, ACM SIGGRAPH 2010]

2015-08-13 James Tompkin – User-Centric Computational Videography 26

Then, a time-line-based narrative construction method where different scenes are placed together to form the tapestry from left to right.

This approach is much more expensive than the Video Tapestries work of Barnes et al., as it includes registration and segmentation.

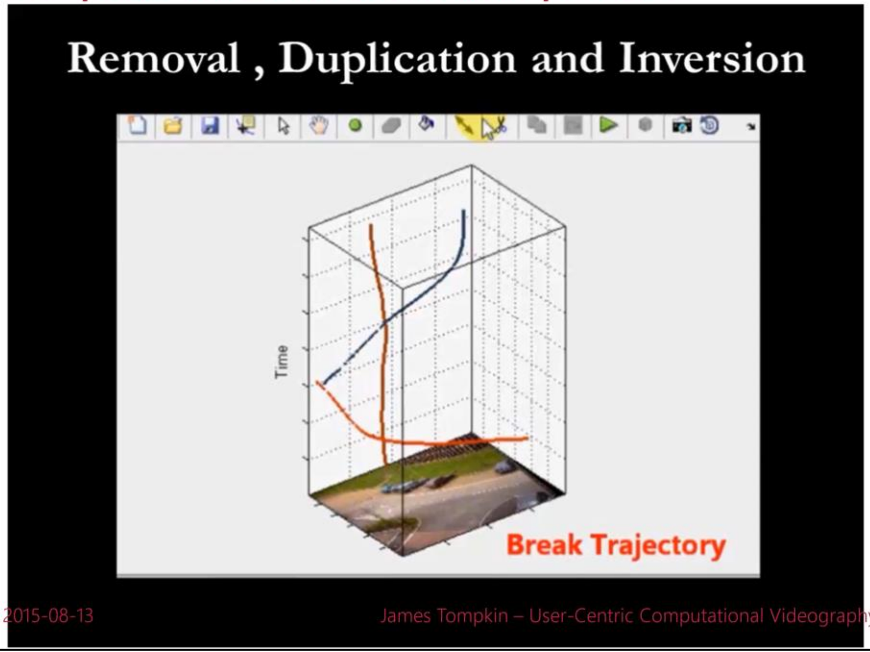
However, we can see dynamism of objects in scene because of this lightweight model of foreground/background.



And, of course, once you have a segmentation for objects and a background reconstruction, then, given some constraints, you can start manipulating them directly.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Temporal Video Content Manipulation



[Shao-Ping Lu; Song-Hai Zhang; Jin Wei; Shi-Min Hu; Martin, R.R., Timeline Editing of Objects in Video, IEEE TVCG 2012.]

[Shah, R. and Narayanan, P.J., Interactive Video Manipulation using Object Trajectories and Scene Backgrounds, IEEE TCSVT, 2013.]

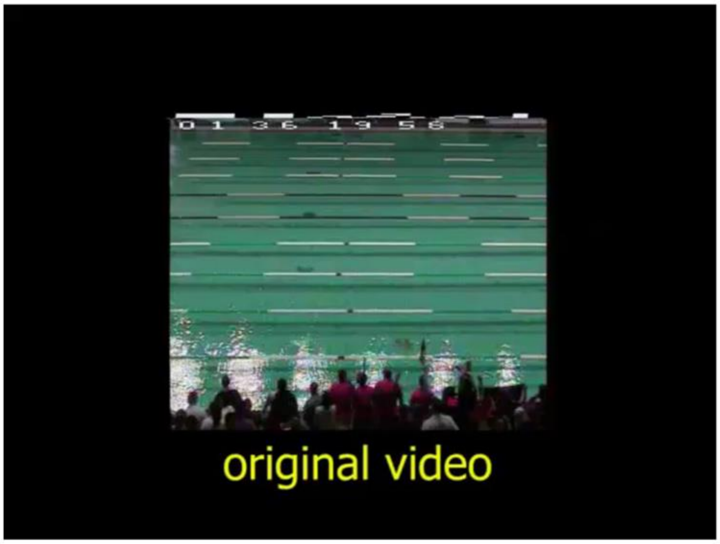
Another example of trajectory manipulation which completely reorders the events in the scene (so long as they don't overlap).

Cung Win and colleagues investigate showing stroboscopic events in a volume.

Volumetric works
[Cuong Nguyen, Yuzhen Niu, and Feng Liu. 2012. Video summagator: an interface for video summarization and navigation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12). ACM, New York, NY, USA, 647-650.]

Temporal Video Content Manipulation

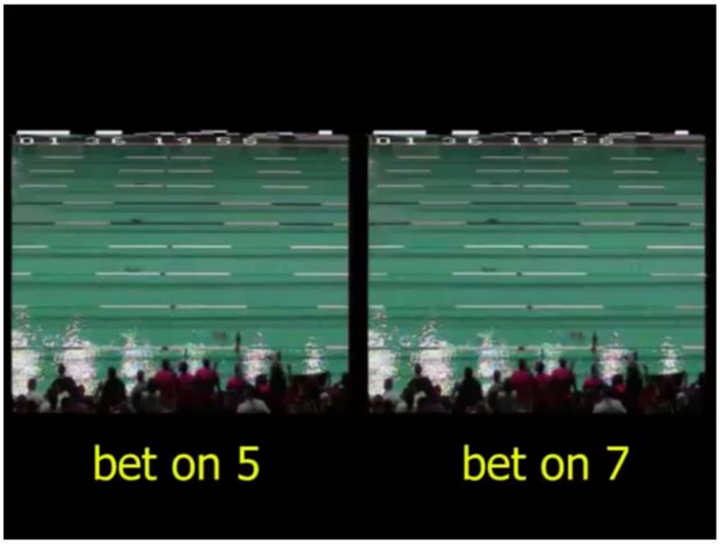
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences



[A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg, *Dynamosaicing: Mosaicing of Dynamic Scenes*, IEEE TPAMI 2007]

And, if you apply trajectories to the whole video at once, rather than just individual objects, we end up in effect with something approaching the Dynamosaicing work of Rav-Acha and colleagues.

Temporal Video Content Manipulation

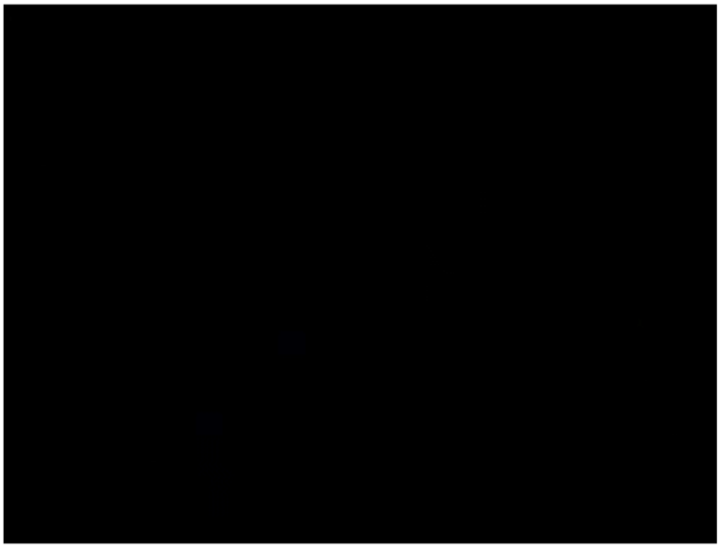


[A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg, *Dynamosaicing: Mosaicing of Dynamic Scenes*, IEEE TPAMI 2007]

But here's where we can start using these techniques for EVIL by rigging the bet.

Video Manipulation as Exploration

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences



[Dan B Goldman, Chris Gonterman, Brian Curless, David Salesin, Steven M. Seitz. Video Annotation, Navigation, and Composition. *UIST* 2008.]

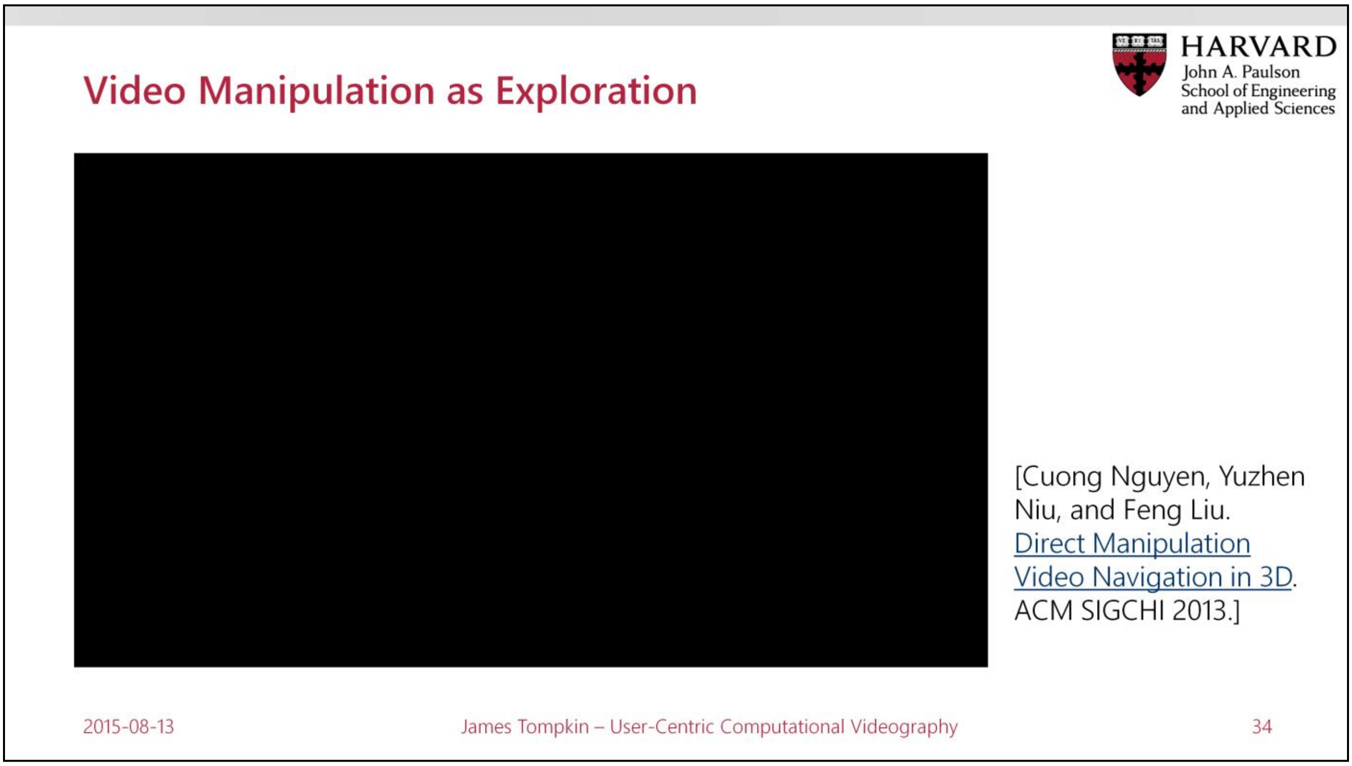
Uses point features and point tracking for more complex scenes, plus motion trajectory clustering for segmentation.

Video Manipulation as Exploration

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

[Dan B Goldman, Chris Gonterman, Brian Curless, David Salesin, Steven M. Seitz. Video Annotation, Navigation, and Composition. *UIST* 2008.]

Point tracking allows more sophisticated content-based exploration methods.



…and, from an interface improvement angle, again Cuong Nguyen (Cung Win) and colleagues apply 3D perspective and temporal corrections to 2D paths to make their trajectories easier to manipulate finely.

Pause/reflect:

At this point we are now very far from our typical idea of ‘video exploration’ that we started with.

We’ve allowed content-based interaction, and even started to create what are arguably new narratives with varying timefronts and trajectory manipulaition.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Fast scan

- Manipulate temporal elements
  - Timelapse (super fast forward)

- [Kalyan Sunkavalli, Wojciech Matusik, Hanspeter Pfister, and Szymon Rusinkiewicz. Factored Time-Lapse Video. ACM SIGGRAPH 2007.]
- [Eric P. Bennett and Leonard McMillan, Computational Time-lapse Video, ACM SIGGRAPH 2007.]
- Rubinstein, M., Liu, C., Sand, P., Durand, F., Freeman, W.T., Motion Denoising with Application to Time-lapse Photography, CVPR 2011]
- [Ricardo Martin-Brualla, David Gallup, Steve M. Seitz. Time-lapse Mining from Internet Photos, ACM SIGGRAPH 2015]

At this point, we've looked at a few different modes for fast seeking, and even some creative examples which explore seeking in videos in spatial ways or in ways which make new art forms in themselves with dynamic narratives.

Let's now look at a few techniques for the other identified navigation mode – the fast scan.

For the fast scan, the most extreme case we can think of is timelapse, which makes a very long video short by massively undersampling.

<ADVANCE>

There are some techniques published at SIGGRAPH for analysing and generating time lapse video, but not so many for exploring it in new ways.

Bennett work – virtual shutter for time-lapse effects

Fast scan

- Exploit knowledge of temporal elements
  - Timelapse (super fast forward)
- Saw Hyperlapse earlier (exploiting camera path smoothing too, either via vision or via metadata like gyro)
  - [Poleg, Y., Halperin, T., Arora, C., Peleg, S., EgoSampling: Fast-Forward and Stereo for Egocentric Videos, CVPR 2015]

We saw how there are new techniques to generate hyperlapse videos either from vision or from gyros, and even there was more work at CVPR recently on this issue.

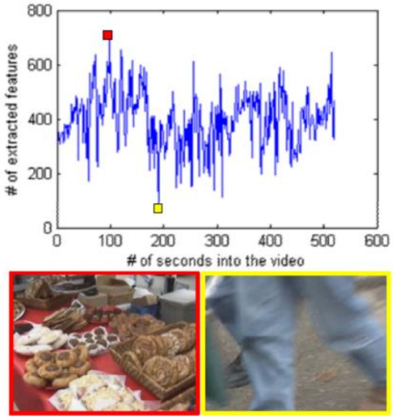
Fast scan

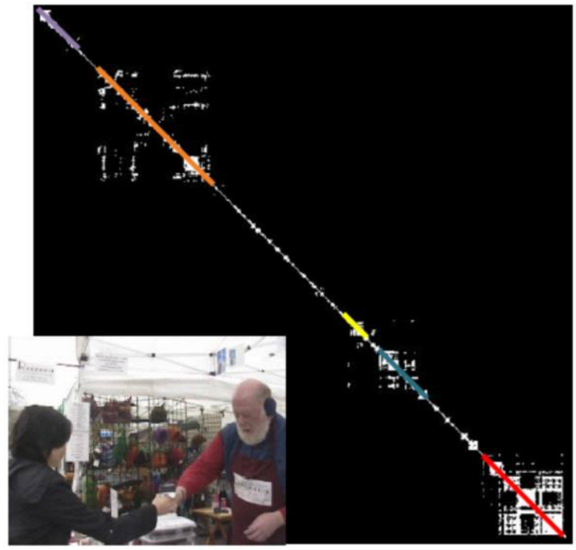
- Exploit knowledge of temporal elements
  - Timelapse (super fast forward)
- Saw Hyperlapse earlier (exploiting camera path smoothing too, either via vision or via metadata like gyro)
- Better fast forward for content, not just camera movement?

What we’re going to look at briefly is how we can look at more than just the camera path to create a better fast forward.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Intelligent fast forward





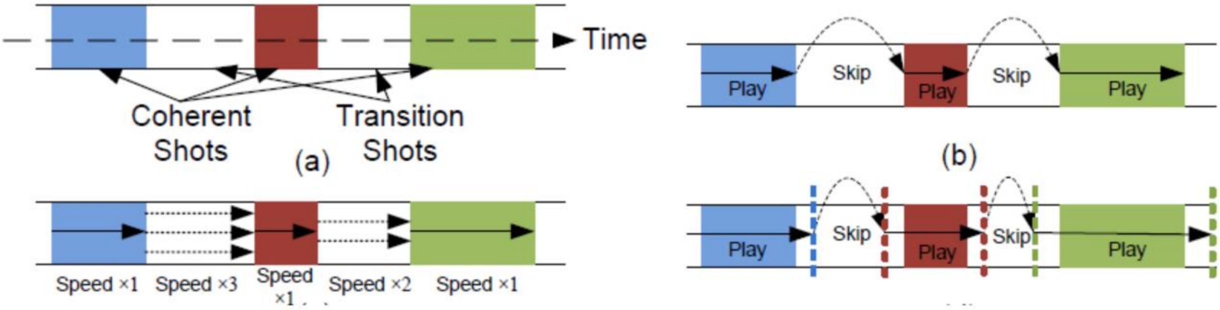
[Pongnumkul, S., Wang, J., Cohen, M., Creating map-based storyboards for browsing tour videos, ACM UIST 2008; with Ramos, G., Content-Aware Dynamic Timeline for Video Browsing, ACM UIST 2010.]

Suporn Pongnumkul

Based on the coherence matrix, we demonstrate a simple al-gorithm to automatically pick out the interesting shots which are represented as blocks in the matrix. We project all the off-diagonal values in the matrix onto the diagonal line, and accumulate them to create a 1-D coherence histogram. We then smooth the histogram using a Gaussian kernel, and ap-ply a threshold (which we set to be 0.3 times the maximum value in the histogram) to generate an initial set of shots. A postprocessing step is then applied to the initial set to remove shots that are too short in time, and to merge shots that are too close to each other into a longer one, to generate the final set of coherent shots.

Intelligent fast forward

HARVARD John A. Paulson School of Engineering and Applied Sciences



[Pongnumkul, S., Wang, J., Cohen, M., Creating map-based storyboards for browsing tour videos, ACM UIST 2008; with Ramos, G., Content-Aware Dynamic Timeline for Video Browsing, ACM UIST 2010.]

This allows more of the interesting content to appear in the fast forward.

And, once computed, we can also apply this to timeline manipulations too.

This is a form of summarization across time, but there are also approaches which apply summarization across space time.

How does this work?

Uses event detection and seamless compositing to shorten very long videos drastically. Used effectively in commercial products for surveillance operations – making very long videos manageable.

http://www.vision.huji.ac.il/video-synopsis/

Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, Clustered Synopsis of Surveillance Video, 6th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS'09), Genoa, Italy, Sept. 2-4, 2009. in press.

Y. Pritch, A. Rav-Acha, and S. Peleg, Nonchronological Video Synopsis and Indexing, IEEE Trans. PAMI, Vol 30, No 11, Nov. 2008, pp. 1971-1984.

Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, Webcam Synopsis: Peeking Around the World, ICCV'07, October 2007. 8p.

A. Rav-Acha, Y. Pritch, and S. Peleg, Making a Long Video Short: Dynamic Video Synopsis, CVPR'06, June 2006, pp. 435-441.

http://www.vision.huji.ac.il/video-synopsis/

Y. Pritch, S. Ratovitch, A. Hendel, and S. Peleg, Clustered Synopsis of Surveillance Video, 6th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance (AVSS'09), Genoa, Italy, Sept. 2-4, 2009. in press.

Y. Pritch, A. Rav-Acha, and S. Peleg, Nonchronological Video Synopsis and Indexing, IEEE Trans. PAMI, Vol 30, No 11, Nov. 2008, pp. 1971-1984.

Y. Pritch, A. Rav-Acha, A. Gutman, and S. Peleg, Webcam Synopsis: Peeking Around the World, ICCV'07, October 2007. 8p.

A. Rav-Acha, Y. Pritch, and S. Peleg, Making a Long Video Short: Dynamic Video Synopsis, CVPR'06, June 2006, pp. 435-441.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

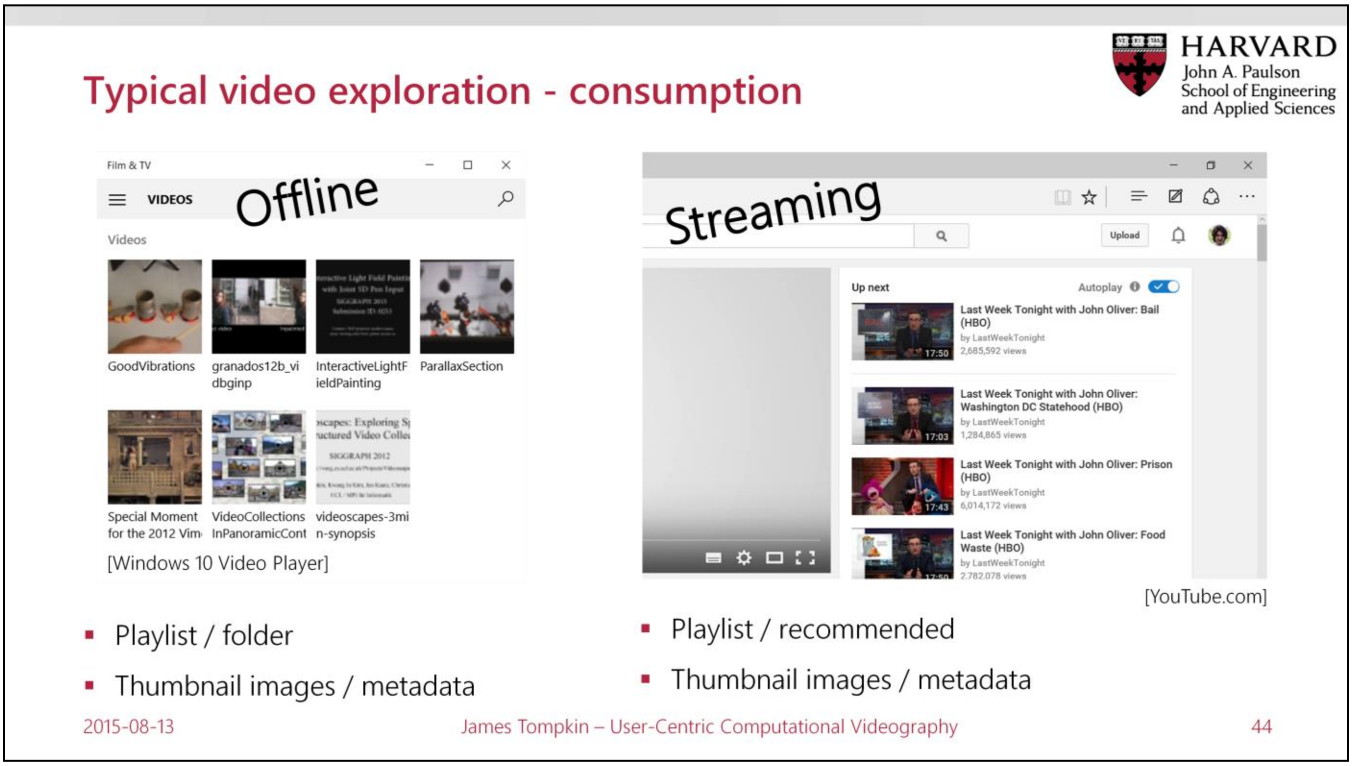
Quick pause – any questions at this stage?

19 minutes

One question max.

Exploring videos (plural)

Let's now move over to looking at new video tools for exploring multiple videos or video collections.



YouTube - manual metadata labelling for content

More complex proprietary systems underneath

Fast automatic content labelling from multiple sources of info

Recommender systems for type

Often too much data to do significant vision-based processing

Automatic labeling

- Vision and information retrieval communities
  - TRECVid workshops and competitions
- Learning about video content
  - Semantic indexing, search
- Labeling videos automatically
- Labeling content within videos automatically

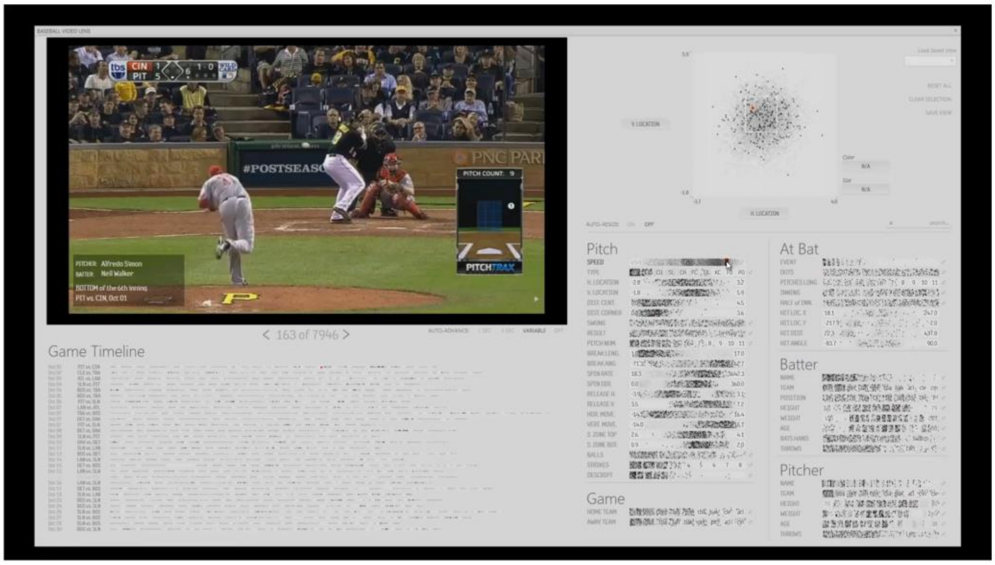
Can’t mention summarization without mentioning the very large and very competent TRECVid community, which largely considers video summarization almost from the vision perspective of trying to deduce what is going on in the scene.

[semantic indexing, interactive surveillance event detection, instance search, event detection, event counting]

However, I’m just going to highlight one work that is more from the visualization side, of what you can do with really exquisitely labelled data.

Faceted metadata search

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

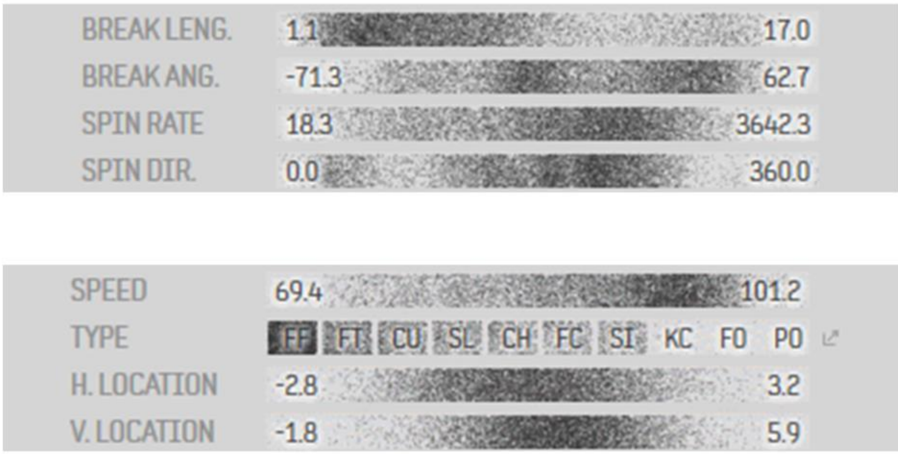


[Justin Matejka, Tovi Grossman & George Fitzmaurice. Video Lens: Rapid Playback and Exploration of Large Video Collections and Associated Metadata. ACM UIST 2014.]

Each dot is a video.

Videos are synchronized by audio of the ball striking the bat.

Faceted metadata search



[Justin Matejka, Tovi Grossman & George Fitzmaurice. Video Lens: Rapid Playback and Exploration of Large Video Collections and Associated Metadata. ACM UIST 2014.]

Very sophisticated query system.
Brushing and linking.

Still, all only enabled by metadata.

Exploring videos (plural) by content

Now, so far, these multiple videos have been connected by their metadata

What kind of content?

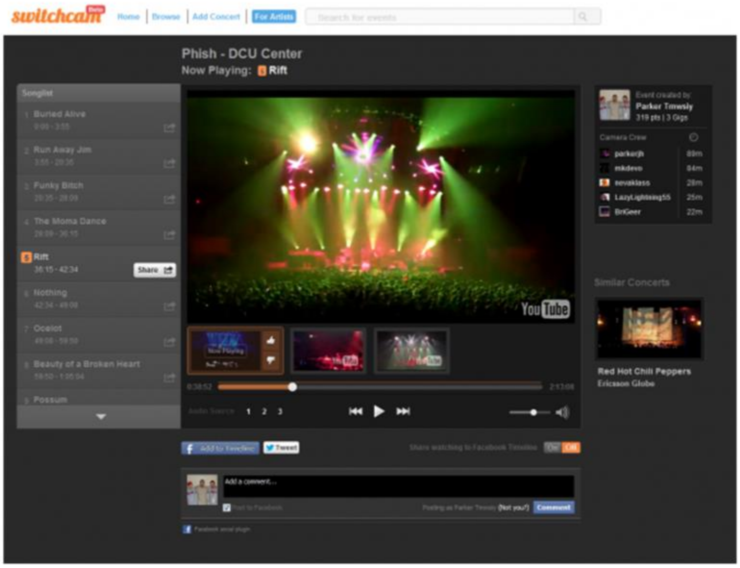
- Metadata

- Temporal and spatial content
  - Image frame content
  - Camera pose
  - Audio content

Metadata + audio contents

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

- Vyclone.com



Switchcam

[Kennedy, L., and Naaman, M. Less Talk, More Rock: Automated Organization of Community-contributed Collections of Concert Videos. WWW, 2009.]

Jiamin VideoSnapping earlier

SwitchCam shutdown; hired by Facebook

Both techniques allow people to take video ‘together’ and then switch camera angles.

In Vyclone, ‘together’ is created manually, with a map-based display showing you nearby concurrent users. The product at the end is a slick edited video combining all angles on to the action.

For Switchcam, the videos were crowdsourced and synchronized by their audio tracks. For concerts. Allowed viewers to explore concerts and switch streams to get a different view.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Metadata + image frame contents

[McCurdy NJ, Griswold WG. A systems architecture for ubiquitous video. MobiSys 2006.]

McCurdy – work with disaster responders. Using early wireless data connections.

Commander view (left) integrated past views of scene to provide transitions (think PhotoTourism before PhotoTourism, but in real time and for disaster response).

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

In-frame spatial alignment

- CamSwarm / CrowdCam

- Exploit scene content connections
  - Unstructured VBR, Ballan et al., SIGGRAPH 2010
  - Videoscapes, Tompkin et al., SIGGRAPH 2012
  - Social camera editing, Arev, Park, et al., SIGGRAPH 2014

- Heavy-weight models – pose, scene geometry

Unstructured VBR

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

[L. Ballan, G. J. Brostow, J. Puwein and M. Pollefeys. Unstructured Video-Based Rendering: Interactive Exploration of Casually Captured Videos. SIGGRAPH 2010.]

Unstructured VBR

[L. Ballan, G. J. Brostow, J. Puwein and M. Pollefeys. Unstructured Video-Based Rendering: Interactive Exploration of Casually Captured Videos. SIGGRAPH 2010.]

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Unstructured VBR

[L. Ballan, G. J. Brostow, J. Puwein and M. Pollefeys. Unstructured Video-Based Rendering: Interactive Exploration of Casually Captured Videos. SIGGRAPH 2010.]

And lets just look at those results again because they are very effective.

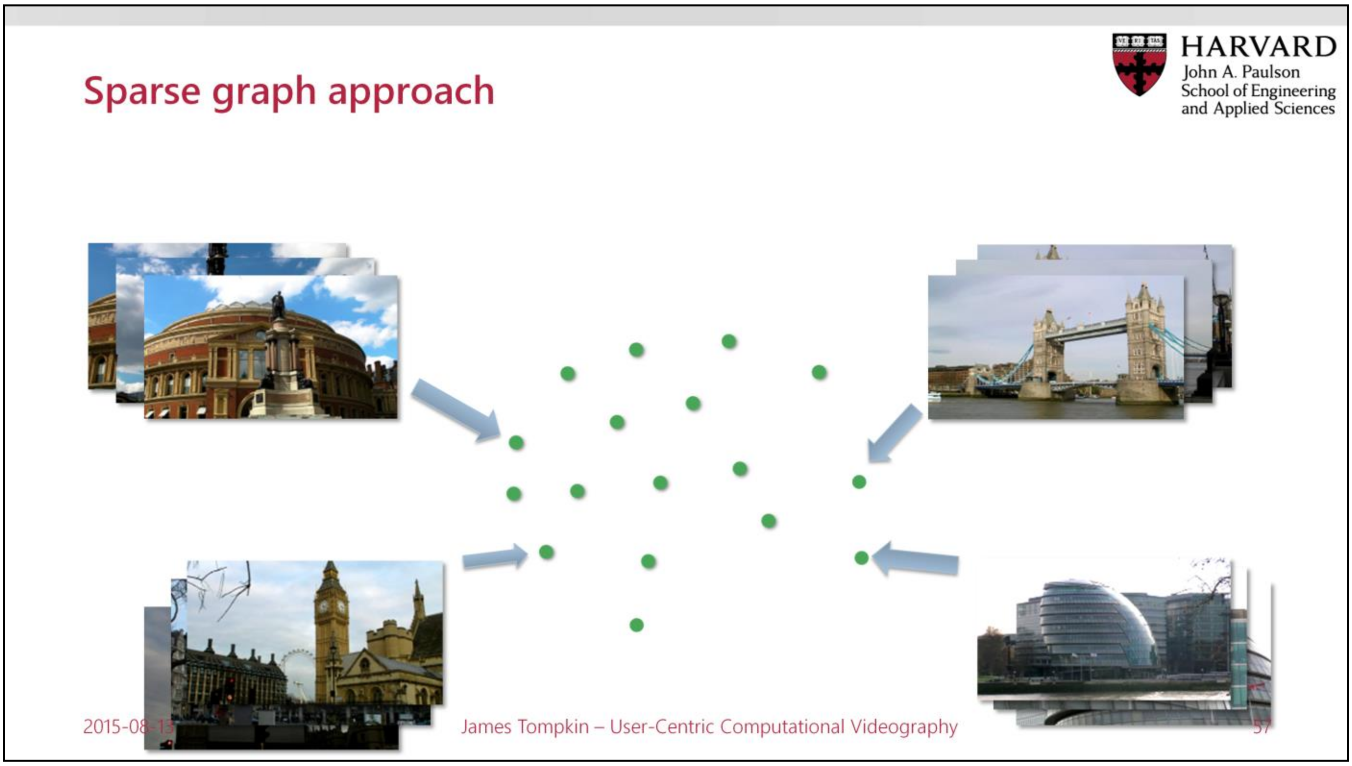
Now, the previous work was narrow baseline, but what if our videos are sparsely located geographically? Say, of a city?

How might these videos be structured and explored?

This is some work that myself and colleagues at UCL and Max-Planck attempted, which we called Videoscapes.

So what's the major problem here?

Well, unlike the cases where we know all the videos are looking at the same target, in our case we have no idea what the videos are looking at,….



<ADVANCE>

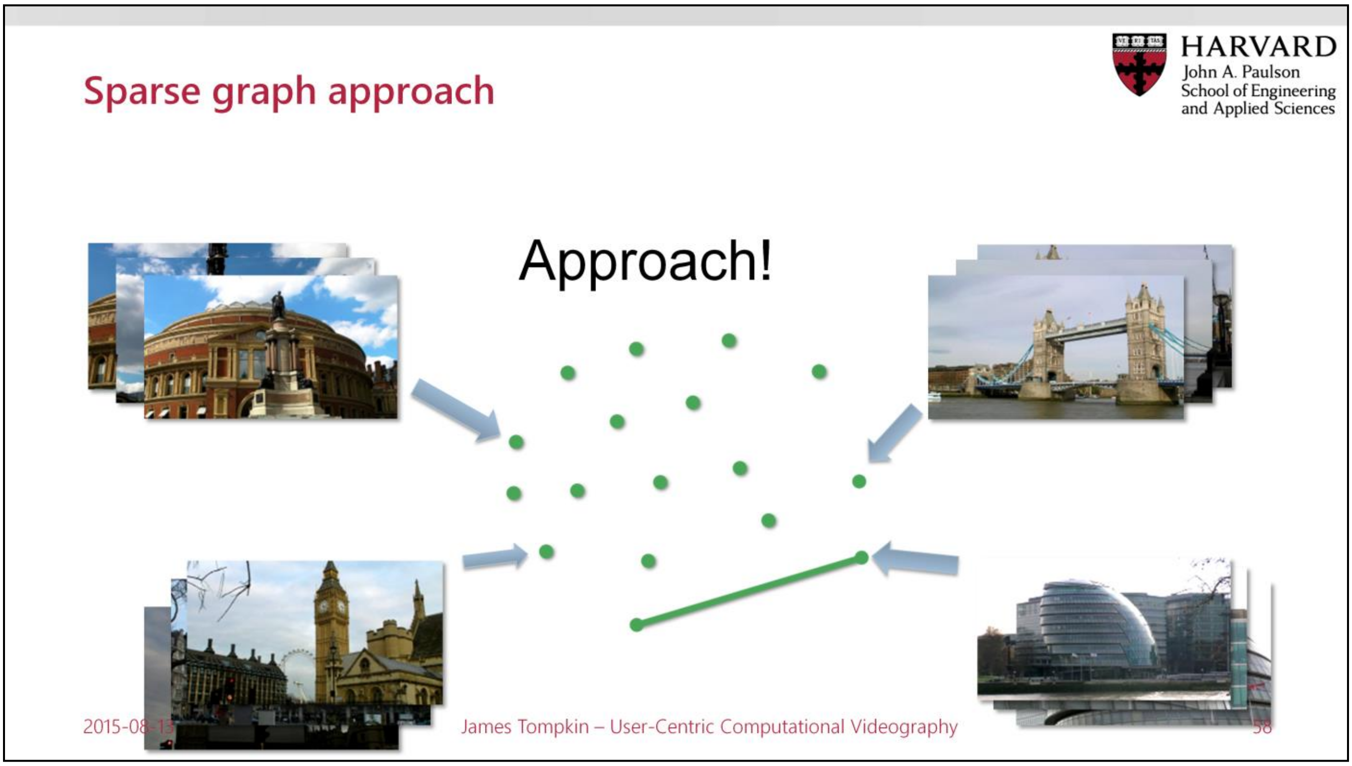
Our idea was to try and find landmarks or buildings within the video collection of a place will likely be captured many times, for example Big Ben.

These provide static, unchanging, easily identifiable points of content similarity within the collection.

At these points, which we call PORTALS,

<ADVANCE>

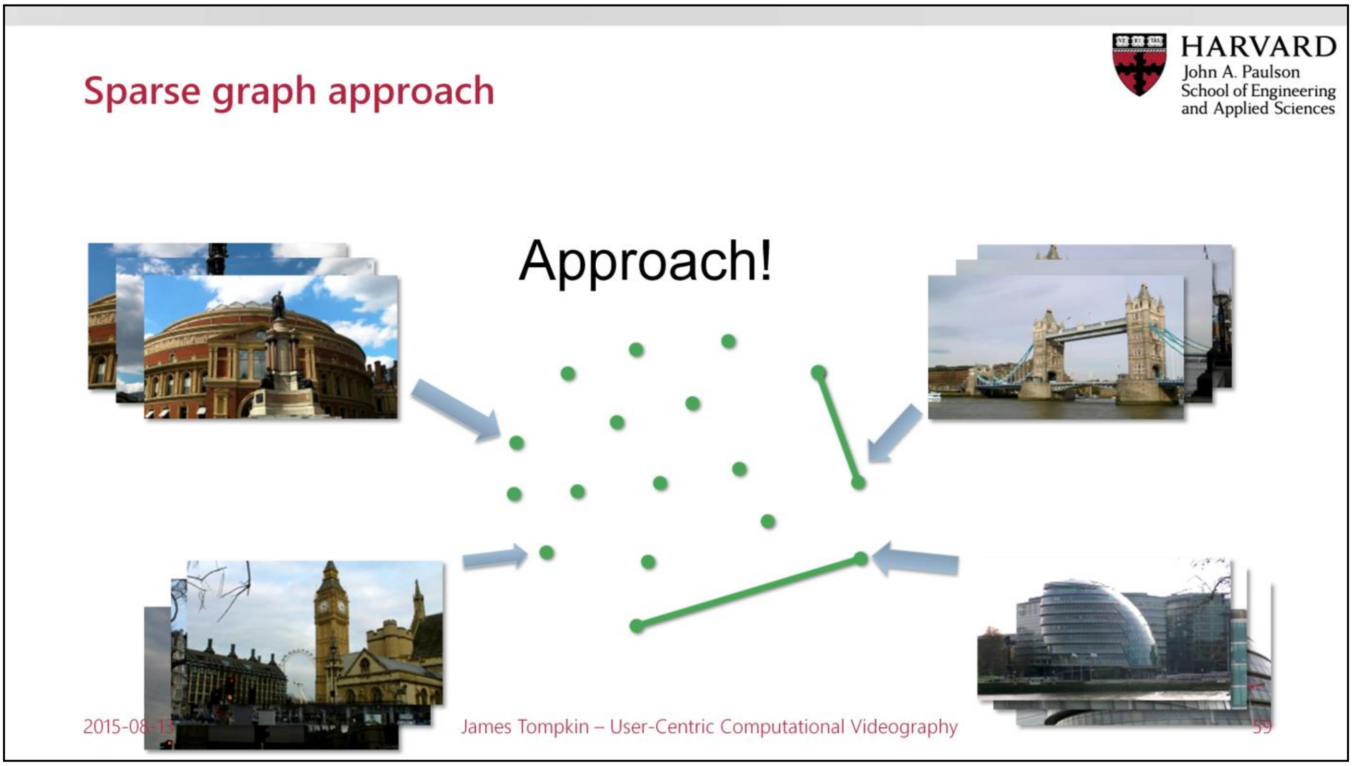
there will probably be enough views of the buildings throughout the whole video collection to perform camera localization and geometry reconstruction. This should then allow us to smoothly transition between different videos at these portals. However, in between portals, we don’t expect to be able to reconstruct anything.



<ADVANCE>

Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

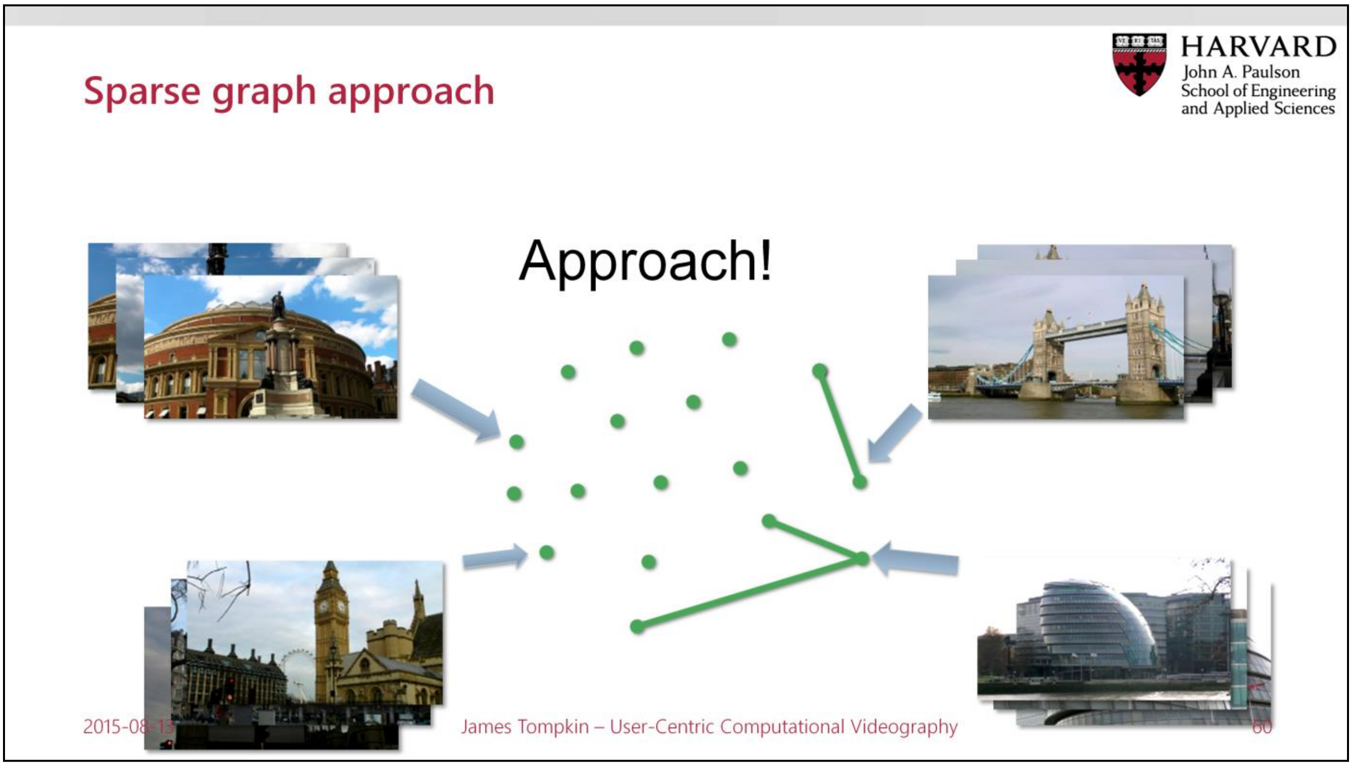
As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.



<ADVANCE>

Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

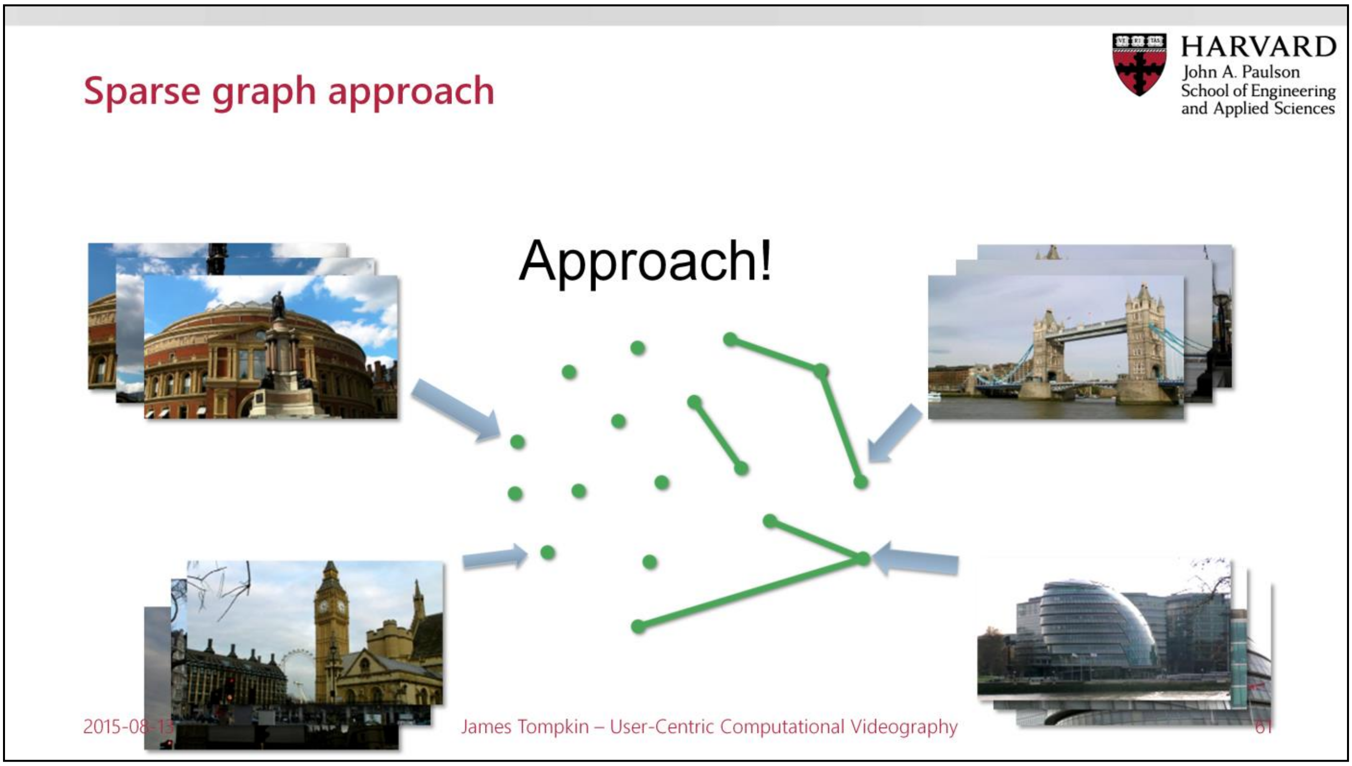
As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.



<ADVANCE>

Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

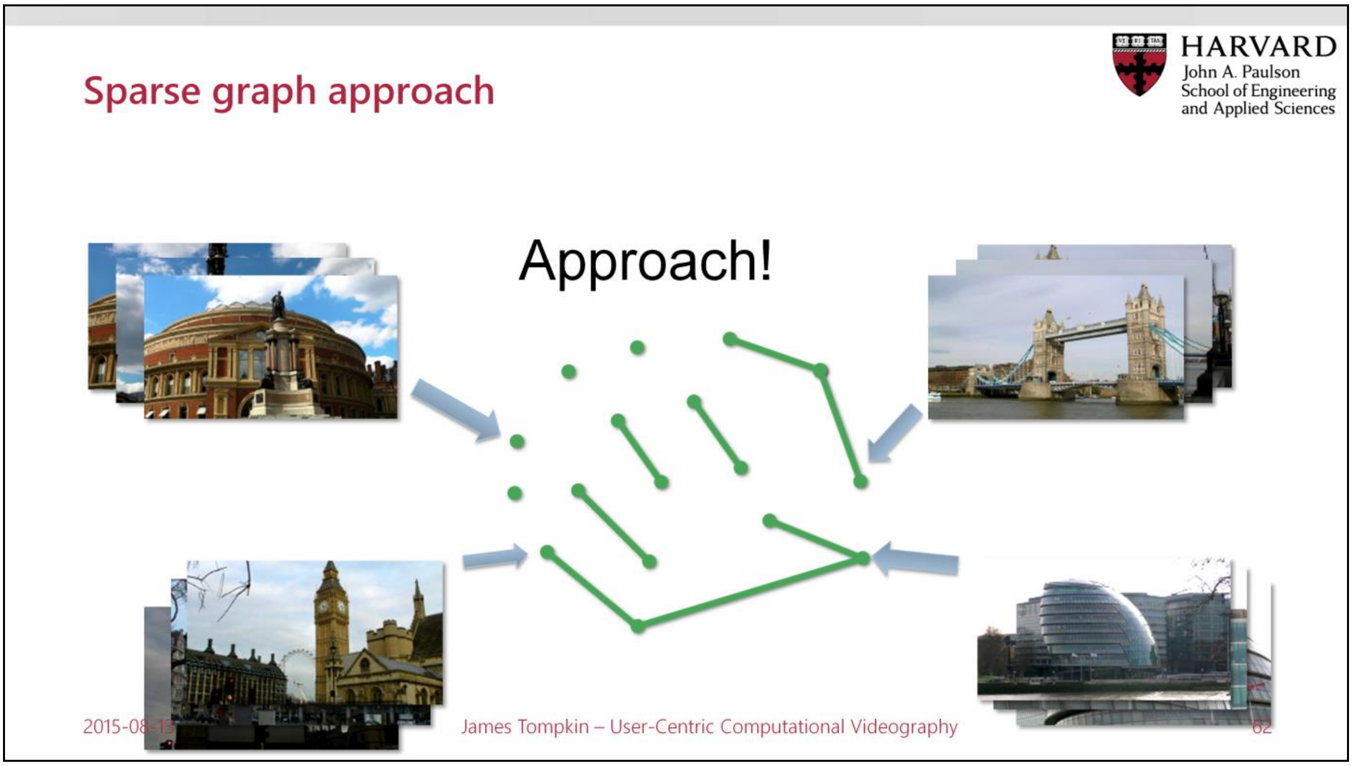
As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.



<ADVANCE>

Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

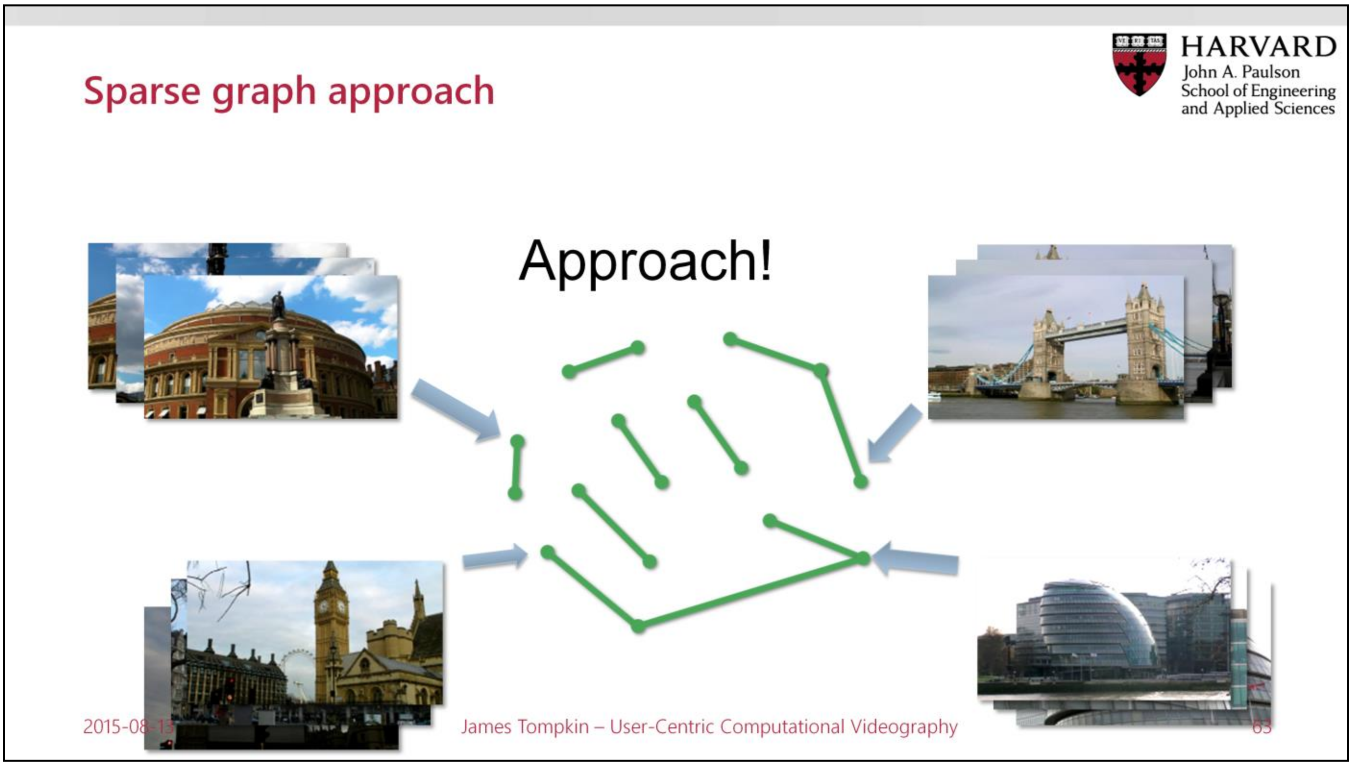
As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.



<ADVANCE>

Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

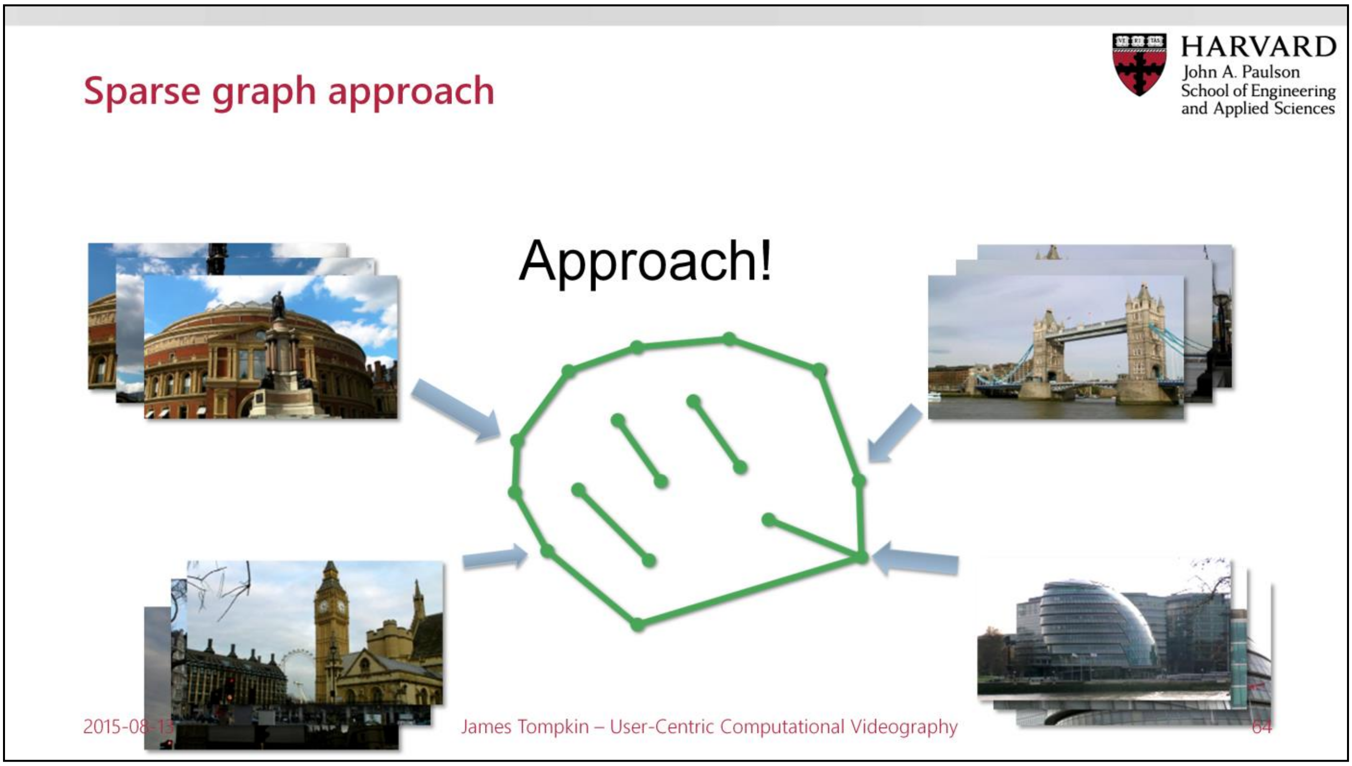
As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.



<ADVANCE>

Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

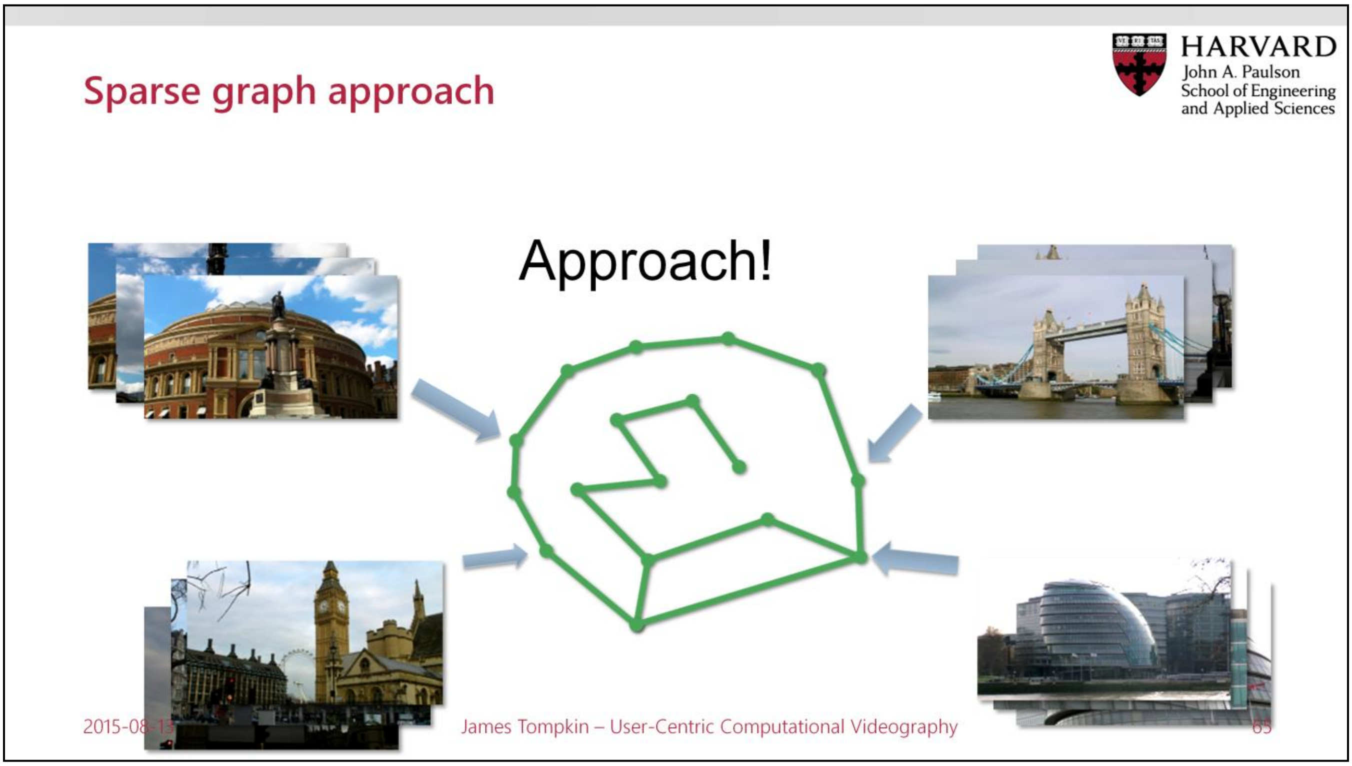
As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.



<ADVANCE>

Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

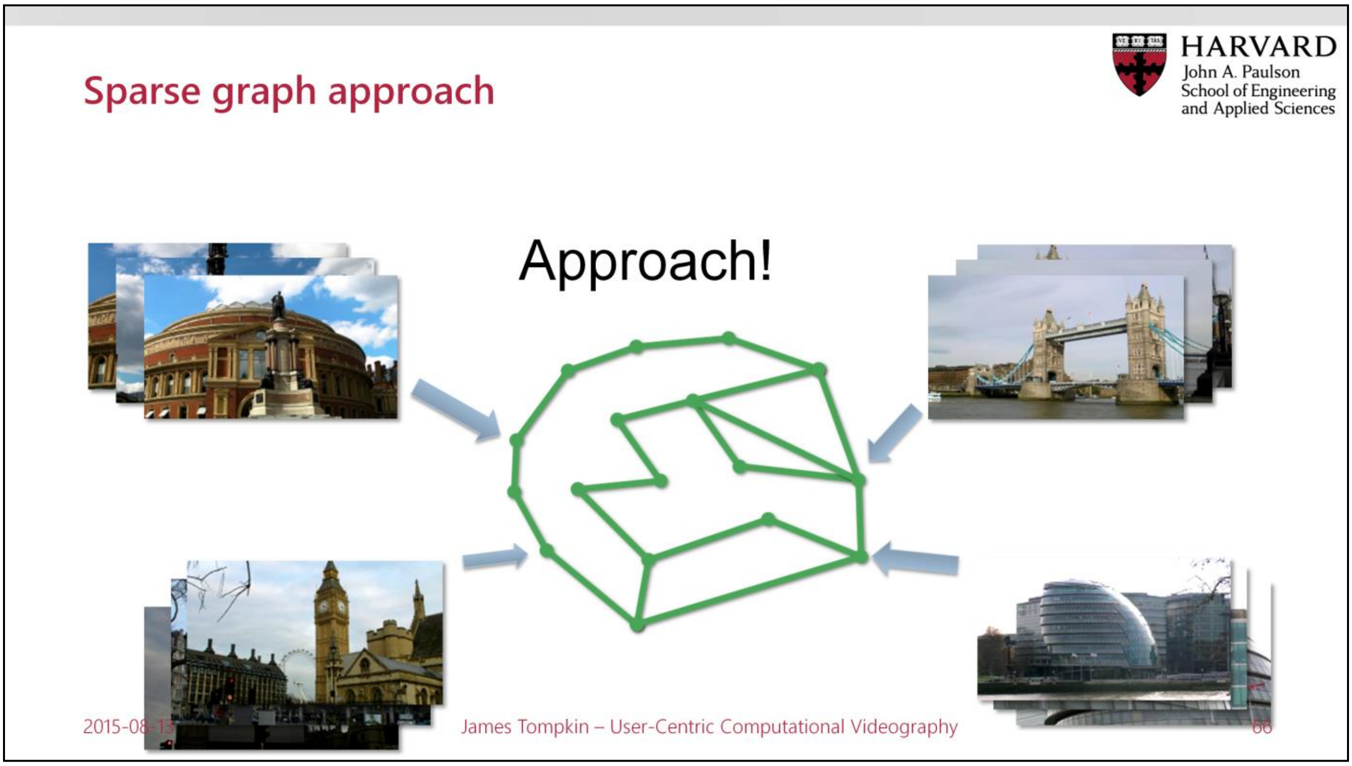
As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.



<ADVANCE>

Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

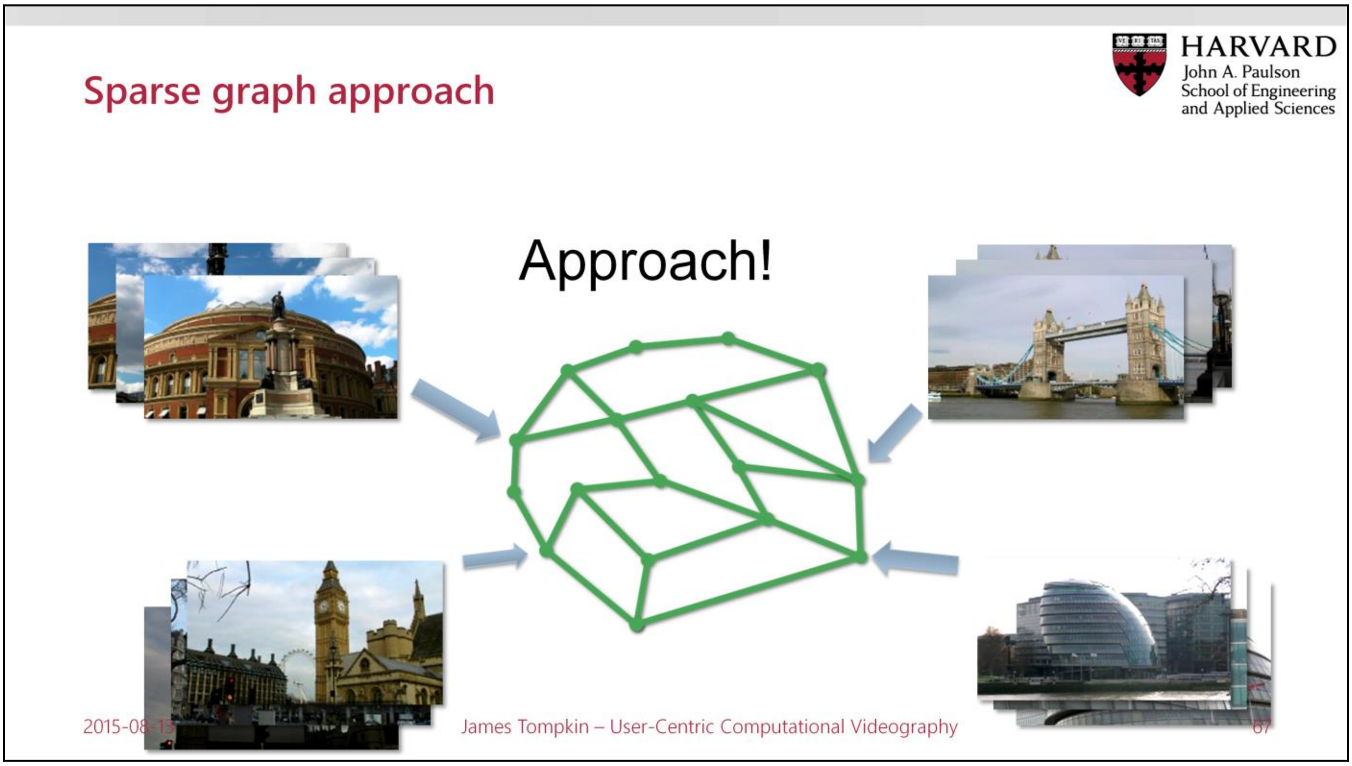
As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.



<ADVANCE>

Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.



<ADVANCE>

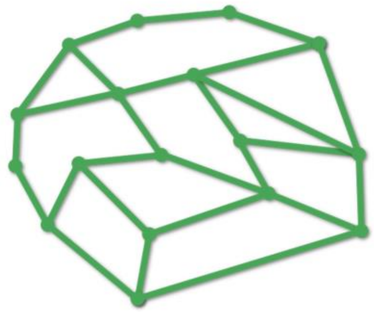
Here we have video to link together the portals. In this way. we recover the local _linkage structure_ of the video collection, rather than a complete reconstruction.

As we will show, for various interfaces and use cases, this linkage structure is enough information to provide novel video collection exploration.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

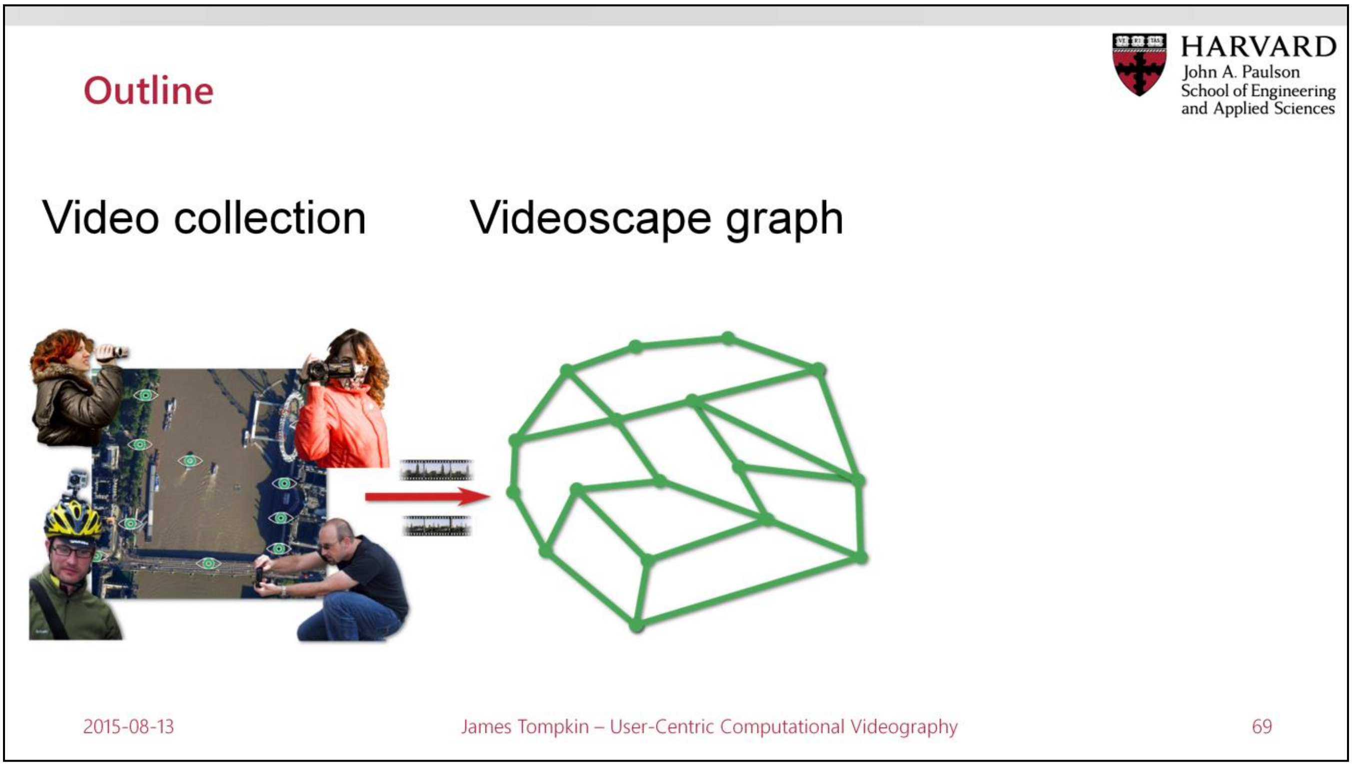
Outline

Videoscape graph

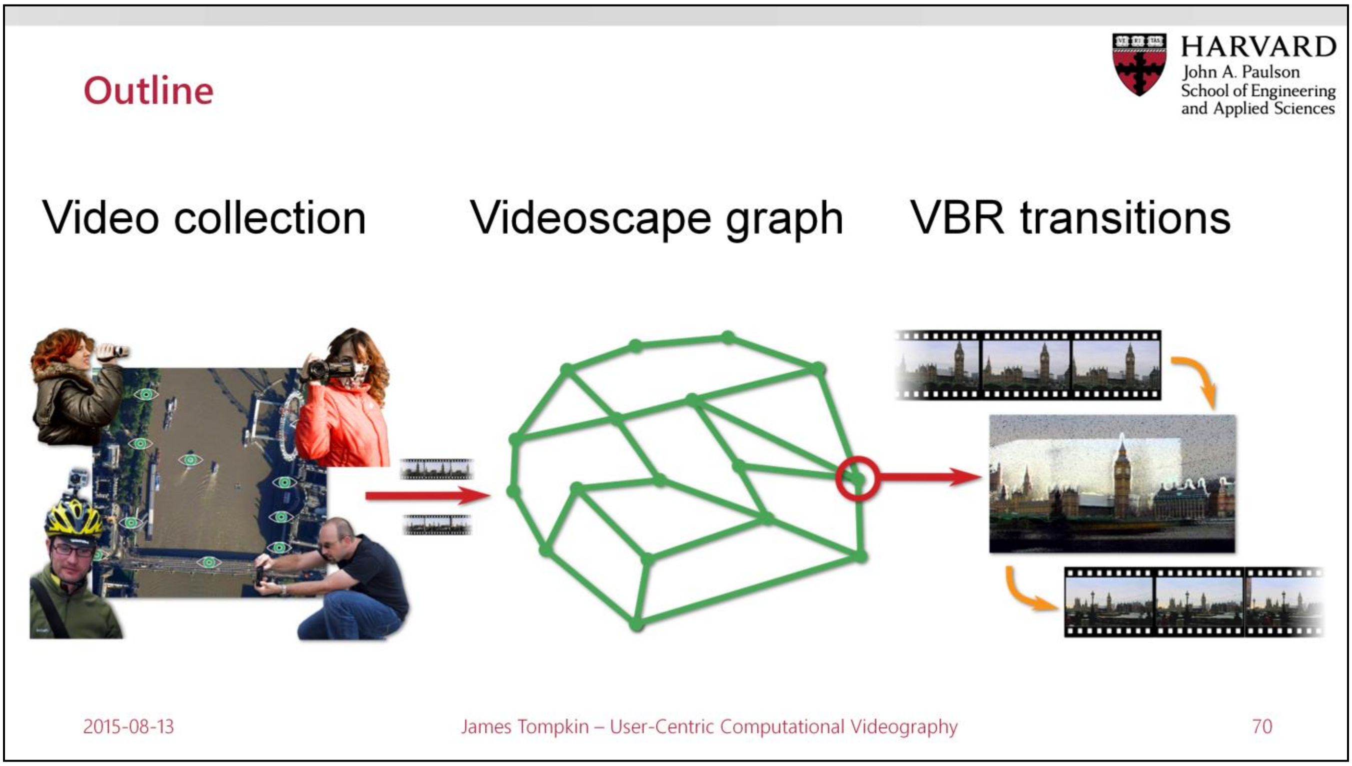


Our talk proceed as follows.

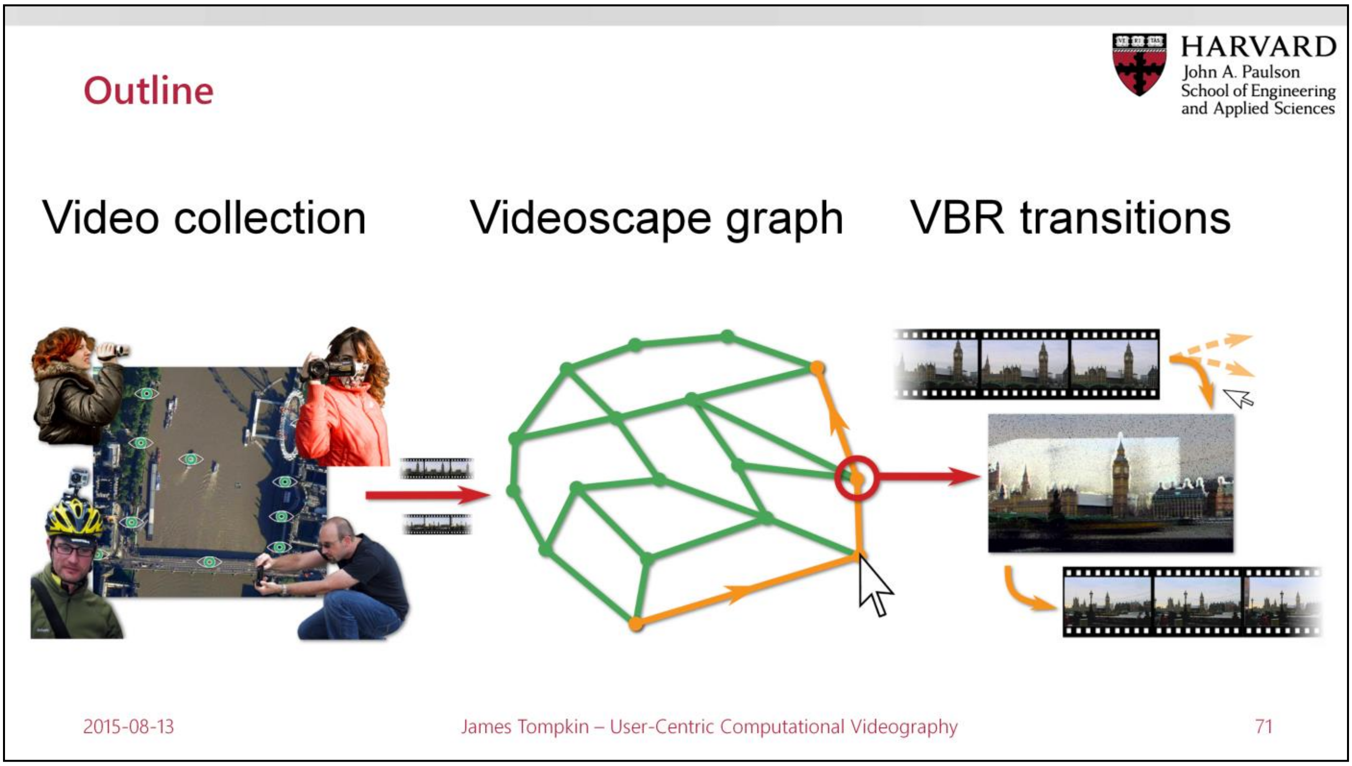
<ADVANCE>



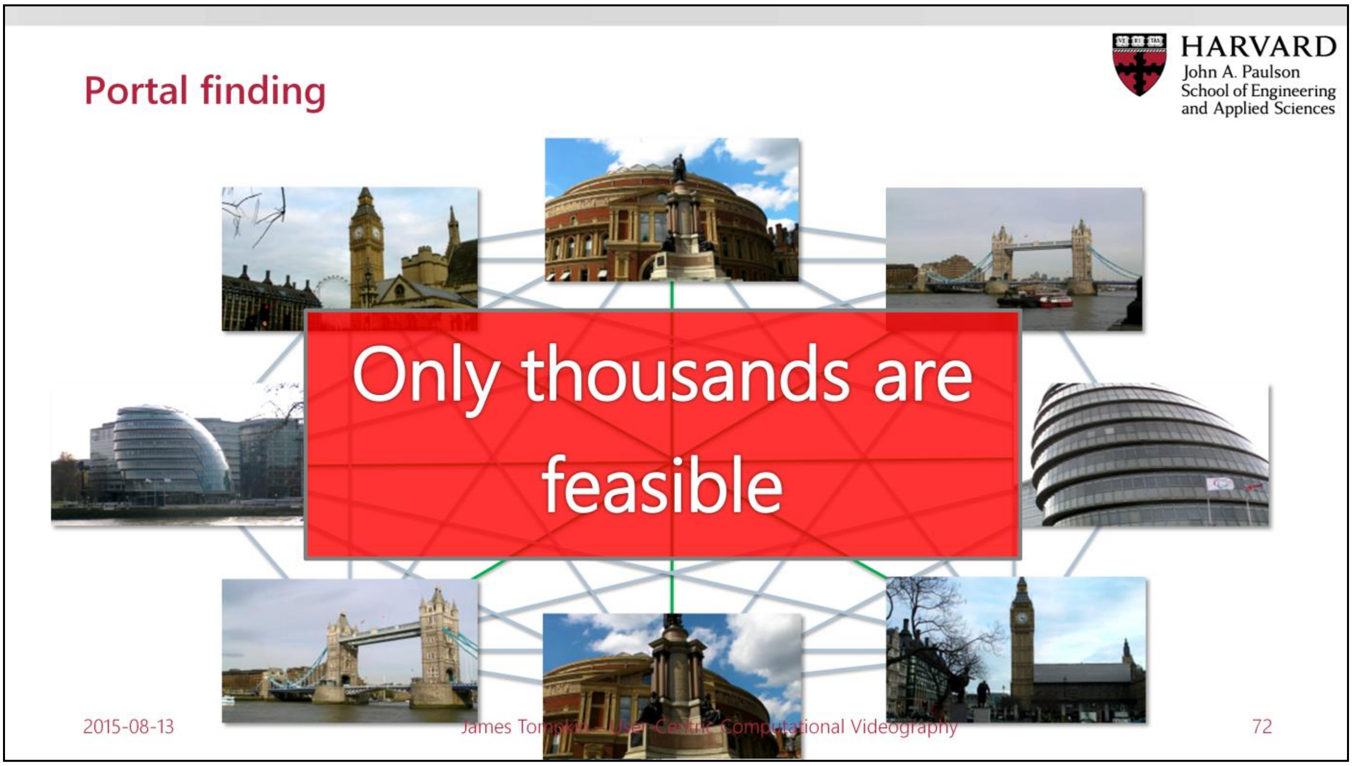
First, we will explain how to turn a sparse, unstructured video collection of a place into a graph that we call a Videoscape.<ADVANCE>



Second, we will explain how to exploit this graph to produce transitions at portals.
<ADVANCE>



Third, we will describe our interfaces which exploit the graph
<ADVANCE>
to provide novel explorations.



In a collection, as a reasonable minimum we might have a million video frames to begin.

If we wished to exhaustively find all portals, we would need to visually match all frames in each video to all frames in every other video in the collection.

Now, obviously this is a big task, as our video collections have in the order of millions of video frames. Realistically, we can only robustly n-to-n match thousands of frames in a day.

As such, we need to throw away most of our frames.

Portal finding

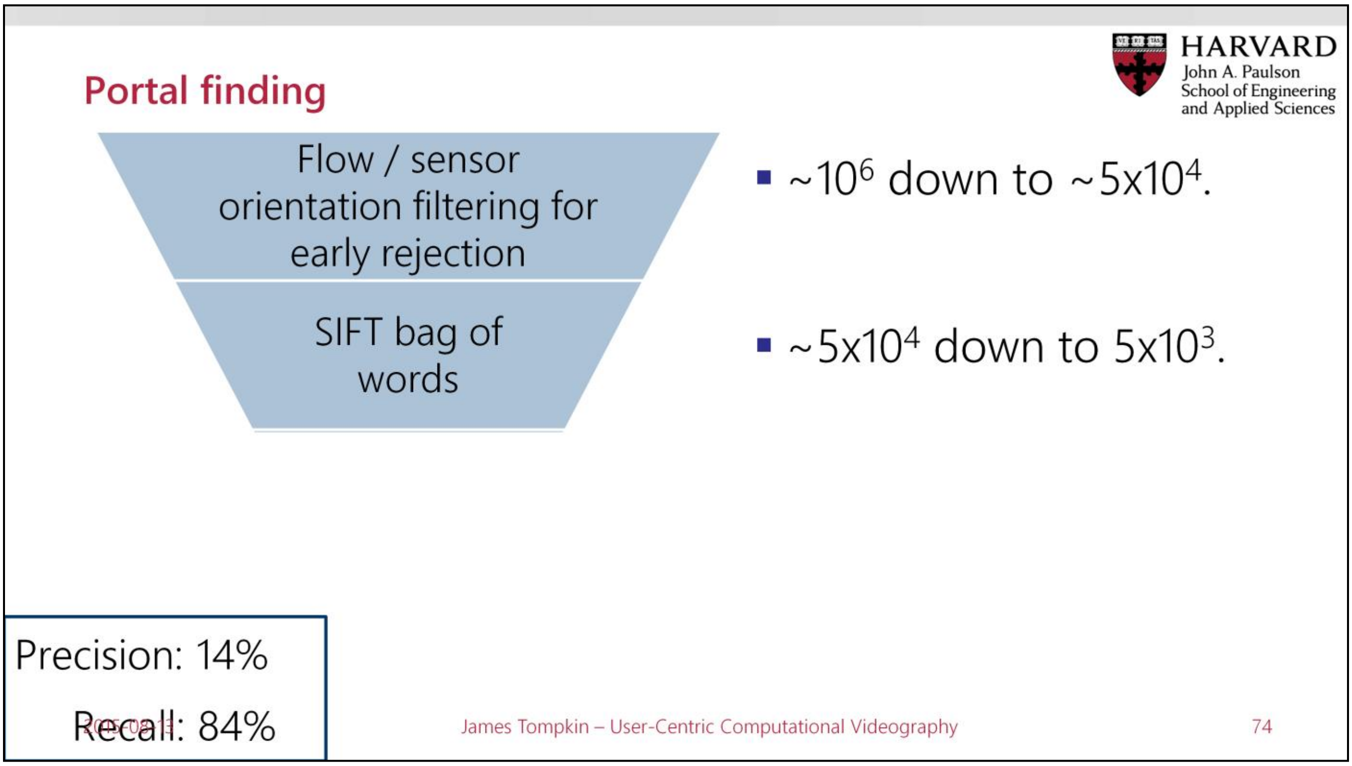
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Flow / sensor orientation filtering for early rejection

- ~10^6 down to ~$5x10^4$.

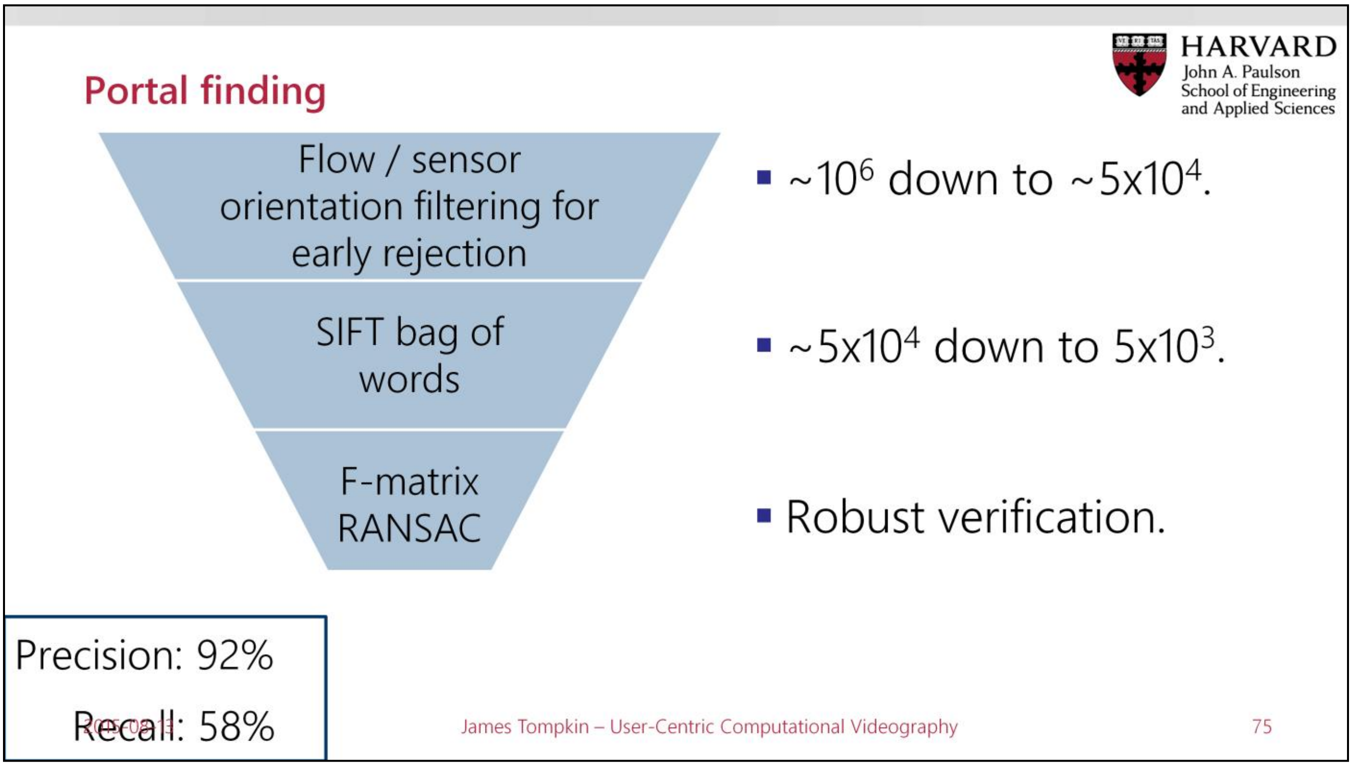
We take a coarse to fine approach to do this.

We first subsample frames and run fast optical flow over the videos to compute approximate camera motions. This lets us correctly sample the visual content as the camera translates, rotates, zooms, or stays still, and allows us to reject about 95% of the frames.



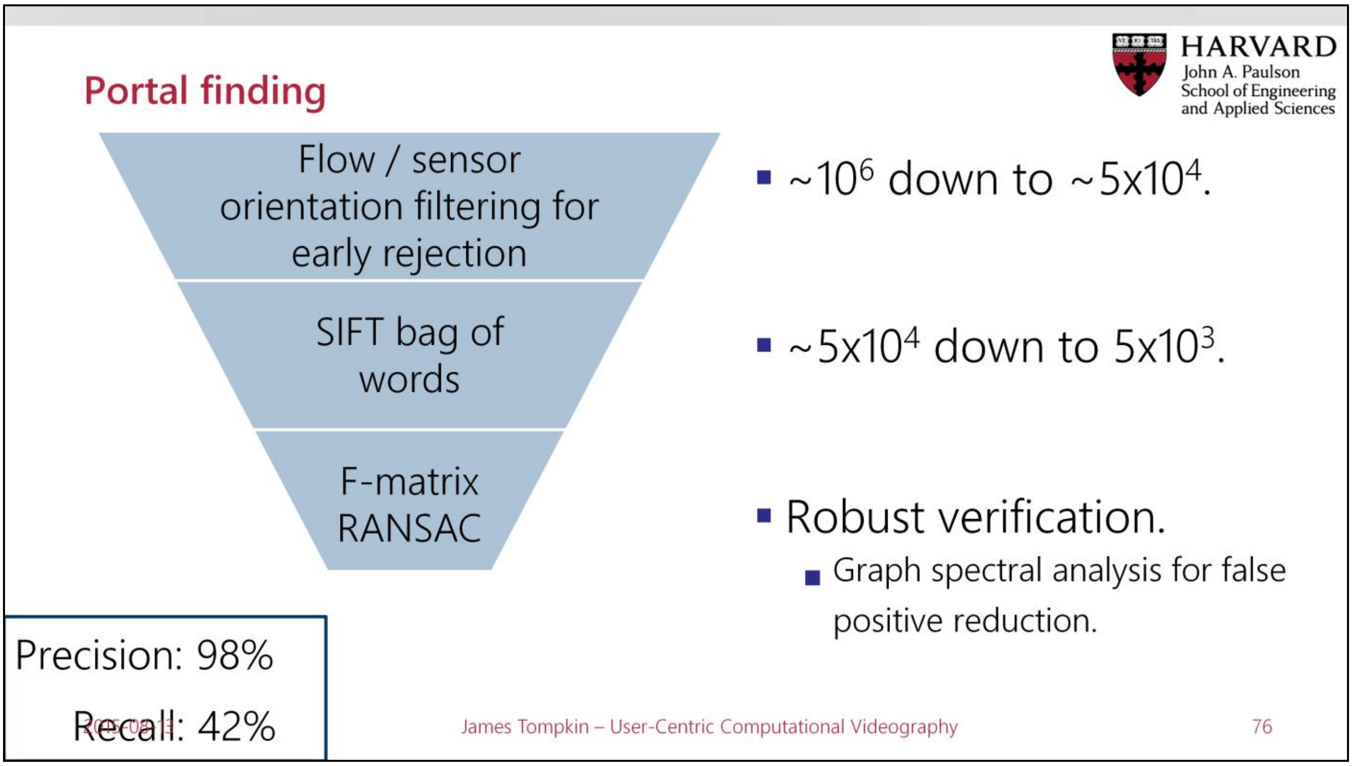
Next, we compute SIFT features and perform bag of words histogram matching. This finds candidate frames with the same visual features, though not perhaps the same content.

At this stage, from the precision and recall, we recover most of the true matches, but also many false matches.



To robustly verify these candidate matches, we see whether the feature correspondences fit a camera model by the fundamental matrix transformation. This greatly increases precision.

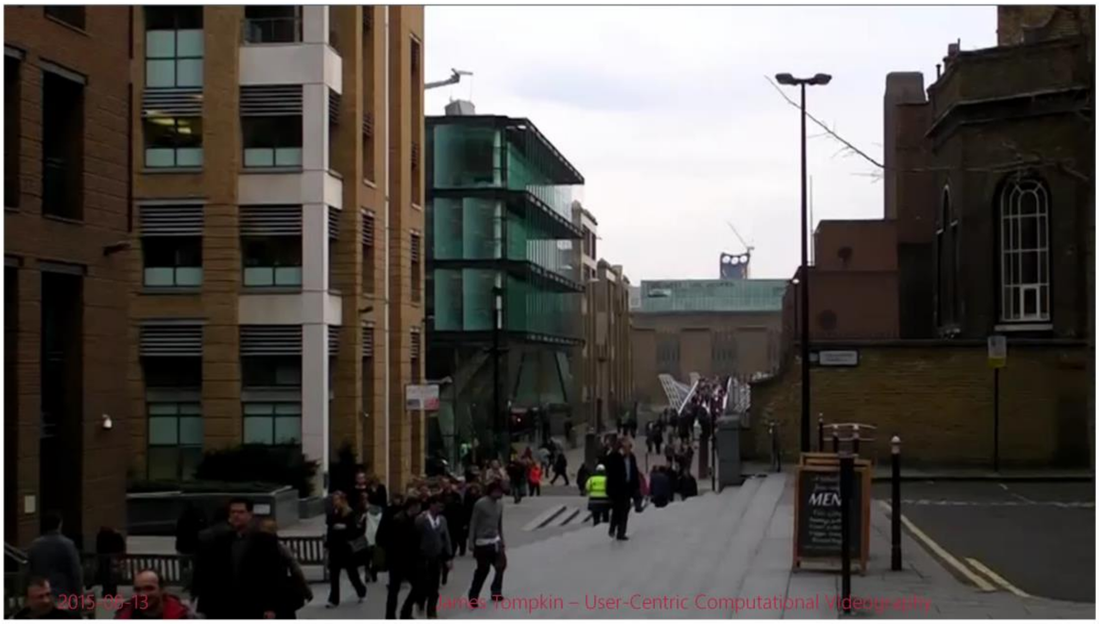
However, the accuracy of matches is crucial – we do not want to join unrelated videos, as this would be very disorientating for the users. A precision of 92% produces one error in ever 12 portals, and this is too many.



We could increase the correspondence thresholds in our robust verification, so that only 1 in 50 portals is in error, but then our recall drops considerably.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Interfaces – Interactive Exploration



[Tompkin J., Kim, K.I., Kautz, J., Theobalt, J. Videoscapes: Exploring Sparse Unstructured Video Collections, SIGGRAPH 2012.]

To not interrupt the viewing experience, a discrete camera icon appears in the corner when a portal approaches in time.

If the user wishes to switch videos, then they need only move the mouse and the navigation bar appears. Otherwise, the current video continues to play.

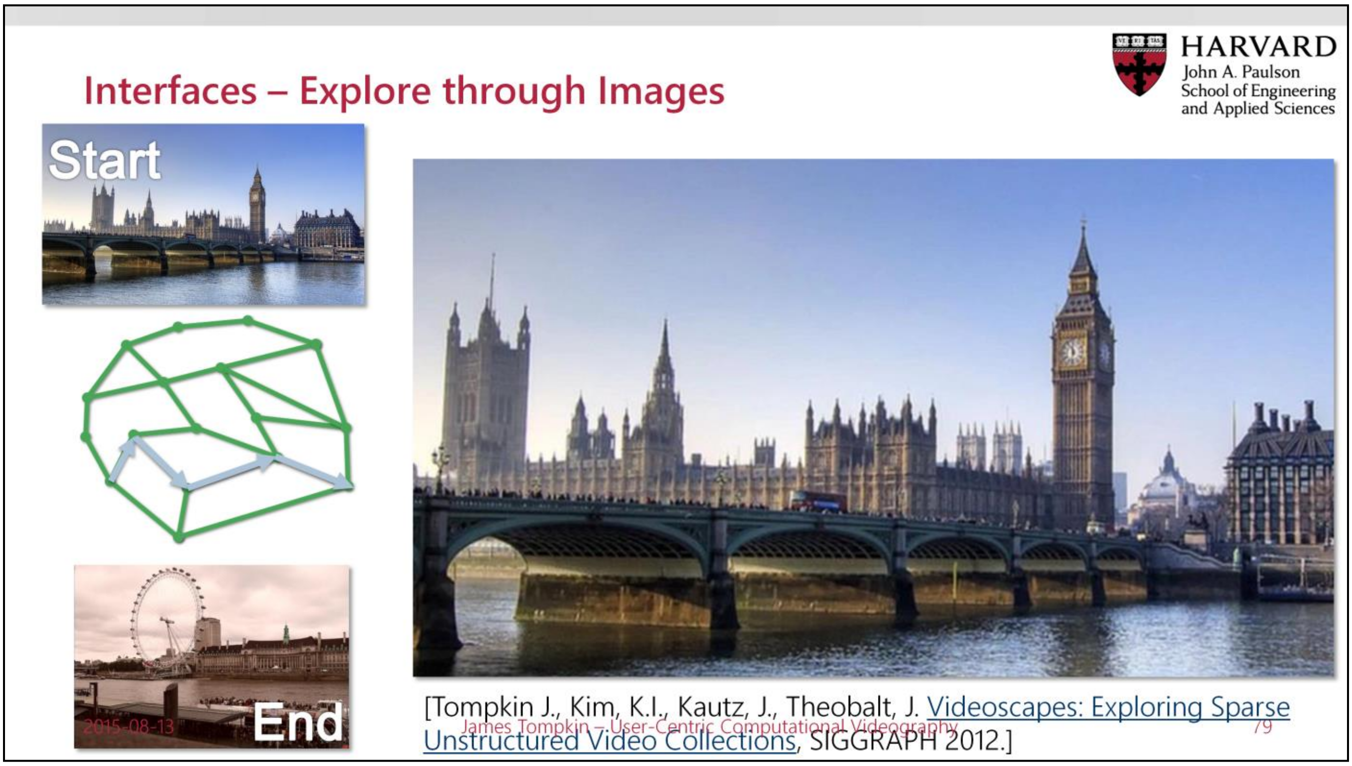
Interfaces – Overview Exploration

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

[Tompkin J., Kim, K.I., Kautz, J., Theobalt, J. Videoscapes: Exploring Sparse Unstructured Video Collections, SIGGRAPH 2012.]

With portal eyes, we can easily define paths around the graph. These are tours, or geographical summarizations, through the video collection. Here, a user allows the software to automatically pick a path through the graph, though one could be defined manually as well.

We now fast forward through the path to show that our system can cover wide geographical areas. Here, we inlay a mini-map which tracks the location of the current camera and provides geographical context. The dynamics and liveliness of a place like this are clear to see.



With this system we can also provide content searches with images – here, the user provides a pair or series of images, and the system generates an exploration of the video collection that takes in those views. This can also be accomplished with labels.

The Videoscape graph can be directed or undirected. Here, we show an undirected graph, so some of the content plays backwards. This could be used to generate creative effects for certain datasets. In this specific example, the system smoothly and automatically moves from one side of the river to the other to take in both views.

Social camera editing

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

[*Ido Arev, *Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir, Automatic Editing of Footage from Multiple Social Cameras, ACM SIGGRAPH 2014.]

If you have a lot of compute power, or a smaller set of videos, you can solve for the 3D pose of every frame of each video within the shared space, and not just at portals as we do.

This lets you compute things such as shared attention.

Then, using cinematographic rules, we can automatically edit the video.

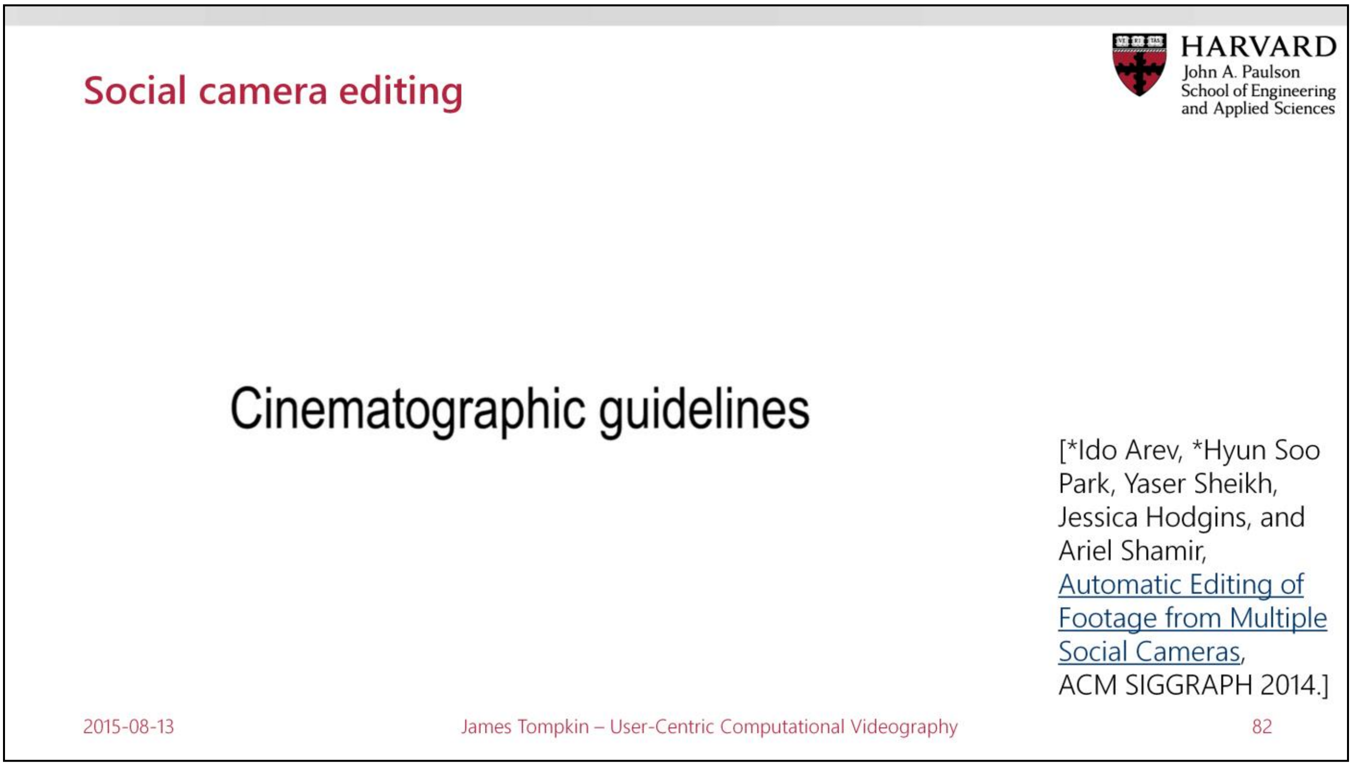
Social camera editing

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

[*Ido Arev, *Hyun Soo Park, Yaser Sheikh, Jessica Hodgins, and Ariel Shamir, Automatic Editing of Footage from Multiple Social Cameras, ACM SIGGRAPH 2014.]

Note how the ball is in view – it hasn't switched to the view of the person carrying the ball (whose camera probably wouldn't see the ball).

Very sophisticated abilities come from having camera pose and this idea of shared attention.



This kind of automatic editing application conforming to cinematographic rules based on shared attention is only possible when we have a combined alignment of all cameras into a world space.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Exploring videos in contexts

Image © [www.italyguides.it]

In Florence, in Italy, there’s a door in room, on the Aventino hill, where many people queue just to look through its keyhole.

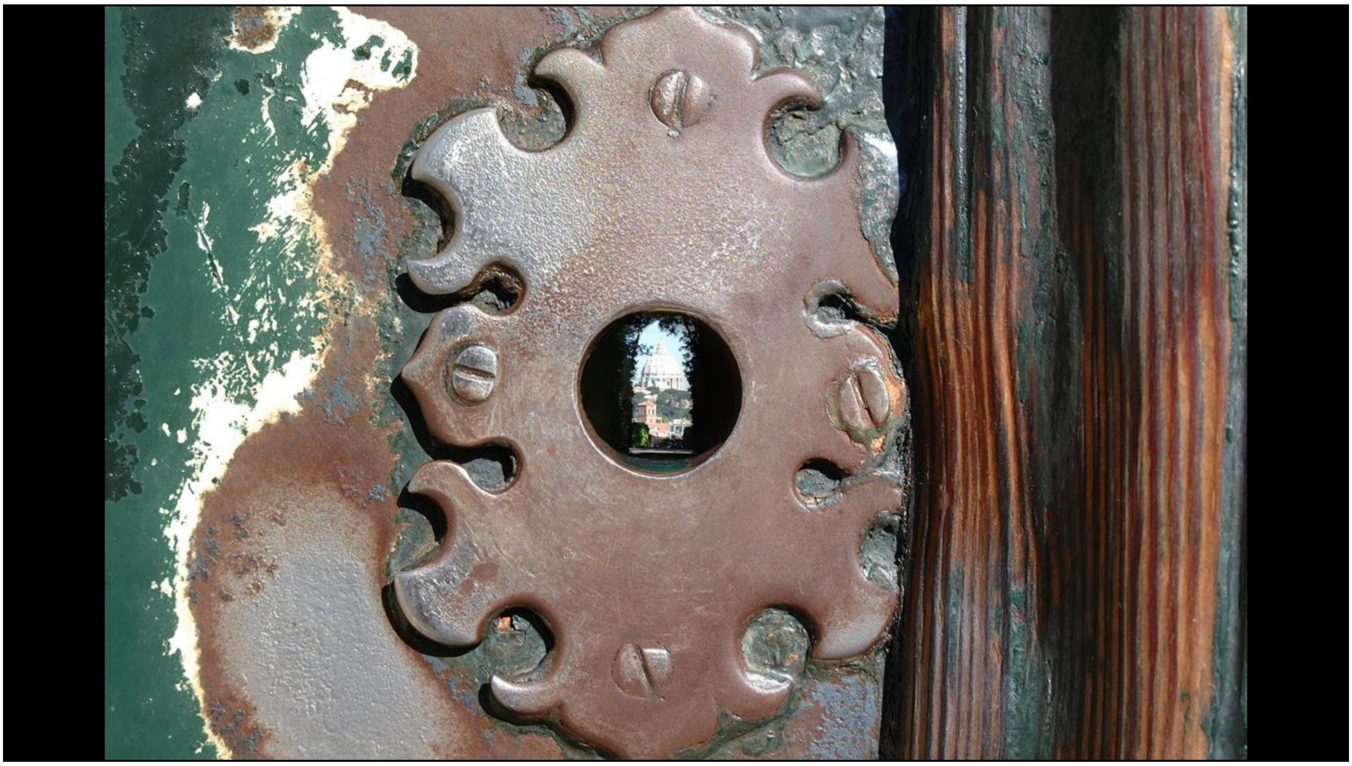


Image © [Panoramio user Carlo Magni]

What you could see once you make it to the top of the queue is this…

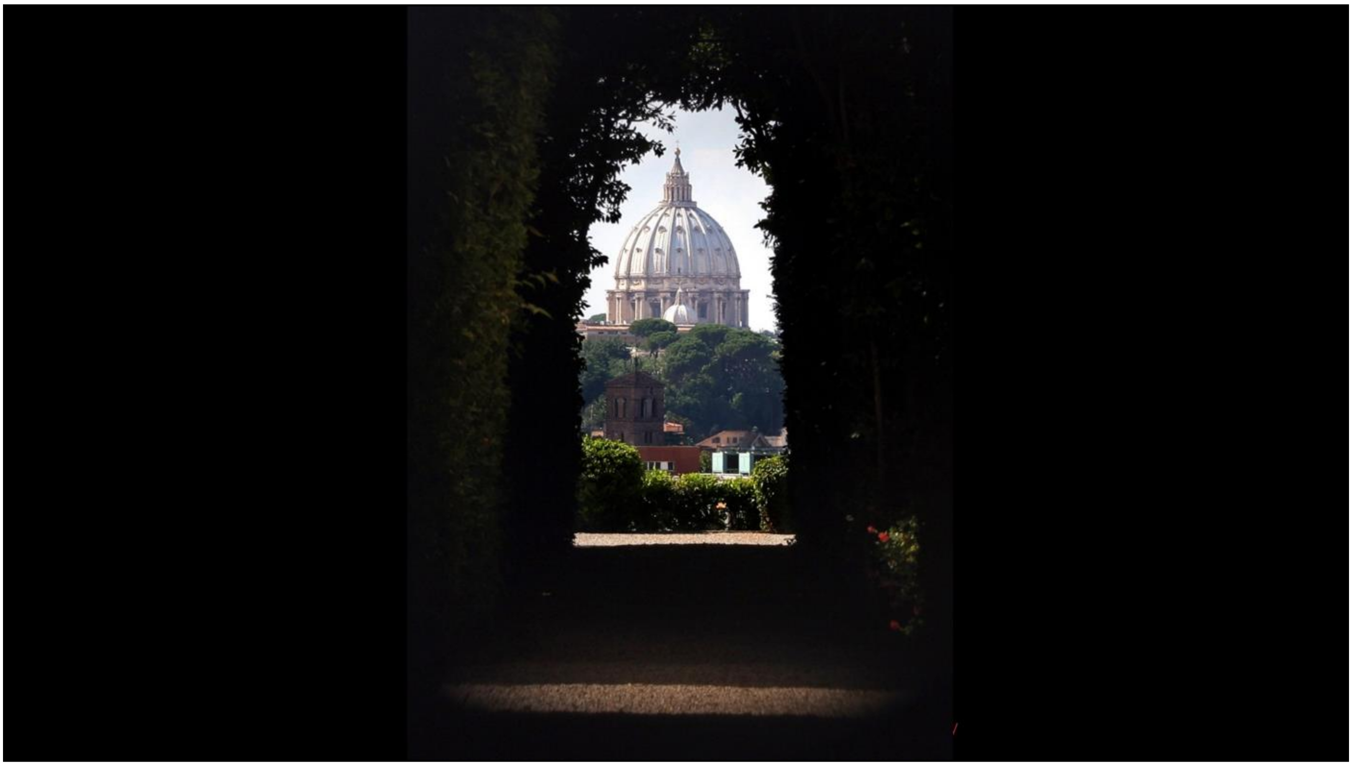


Image © [http://it.wikipedia.org/wiki/File:StPetersBasilica_Keyhole_2.jpg]

Or actually, this, the stunning Duomo of Saint Peter's Basilica.
View is constrained by the keyhole.

A lot of times, this is what capturing video is like – we see only through the field of view of the camera.

Image © [Panoramio user emilliriz]

However, few people then make it into the actual garden and have a look at the bigger picture, or as we will call it during the rest of this talk, the context!

There are many works which attempt this, and I’m just going to highlight two areas of work which attempt to contextualised videos spatially.

Thanks to Fabrizio Pece for these past 4 slides.

Geometric contexts

- Ulrich Neumann & co. @ USC

[U. Neumann, S. You, J. Hu, B. Jiang, and I. O. Sebe,
Visualizing Reality in an Augmented Virtual Environment,
presence: Teleoperators and Virtual Environments Journal, Vol 13, No. 2, pp. 222-233, April 2004.]

Geometric contexts

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

- Ulrich Neumann & co. @ USC

[U. Neumann, S. You, J. Hu, B. Jiang, and I. O. Sebe,
Visualizing Reality in an Augmented Virtual Environment,
presence: Teleoperators and Virtual Environments Journal, Vol 13, No. 2, pp. 222-233, April 2004.]

Video Collections in Panoramic Contexts

James Tompkin, Fabrizio Pece, Rajvi Shah, Shahram Izadi, Jan Kautz, Christian Theobalt
UIST 2013

Now we tried to investigate how to explore videos in contexts ourselves but, as building world geometry can be difficult, we decided to use the plentiful sources of panoramic imagery as our contexts

Video Collections in Panoramic Contexts

James Tompkin, Fabrizio Pece, Rajvi Shah, Shahram Izadi, Jan Kautz, Christian Theobalt
UIST 2013

Input:
- Collection of videos
- Panorama(s)

Output:
- Video temporally and spatially embedded within the panorama.

Contemporaneous capture: [Pirk, S., Cohen, M. F., Deussen, O., Uyttendaele, M., and Kopf, J. Video Enhanced Gigapixel Panoramas. SIGGRAPH Asia 2012 Technical Briefs.]

Our system, VidiContexts, takes as input a collection of videos taken from mobile phones
<ADVANCE>

and a panoramic image of a location. Our videos are captured together with timestamps, GPS, and sensor orientation data.

<ADVANCE>

As output, the system shows the videos temporally and spatially embedded within the panorama. This then allows us to analyse the data and produce heat-map like visualizations of attention, and to allow direct manipulation with scrubbable foci.

To achieve this, videos need to be registered to the panorama, firstly exploiting the orientation information, and then using image features.

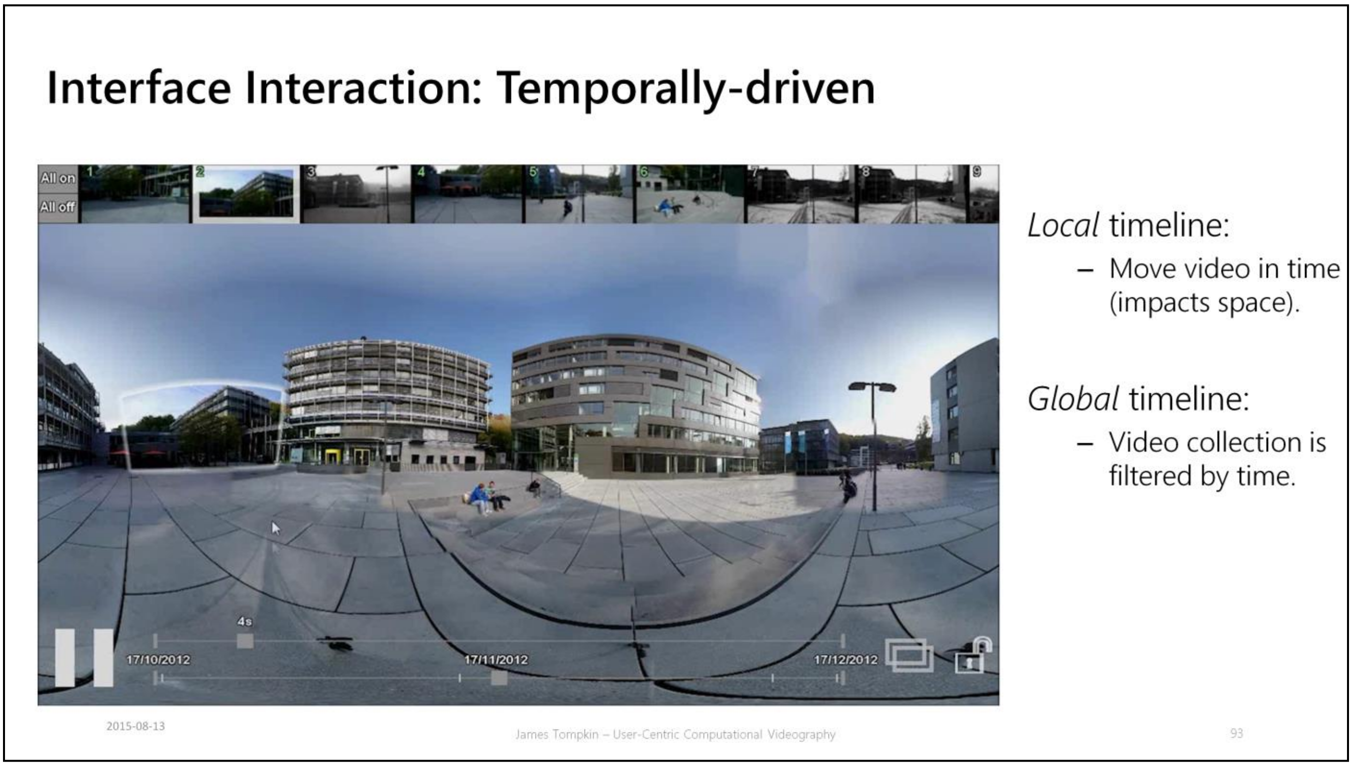
Here’s the VidiContexts interface.

<ADVANCE>

The interface can be divided in 4 regions:

1. Video Foci – here we can see some of the spatio-temporal comparison that is possible with the system, showing snow in the courtyard in winter on the left, in autumn with rain and pink sky in the middle, and in summer in full sun on the right.
   <ADVANCE>
2. Video Collection menu, with a draggable video selection strip
   <ADVANCE>
3. Playback controls, together with global and local timelines
   <ADVANCE>
4. Context and camera control, with overlapping video switcher, perspective change and lock to a video menu

The interface presents the context in either look-around perspective projection or as a full equirectangular map projection with an infinitely-rotating canvas.
The system allows a variety of interactions: including pan and zoom on the context and smoothly switch between perspectives.



Besides common interaction, VidiContexts offers two other type of interactions:

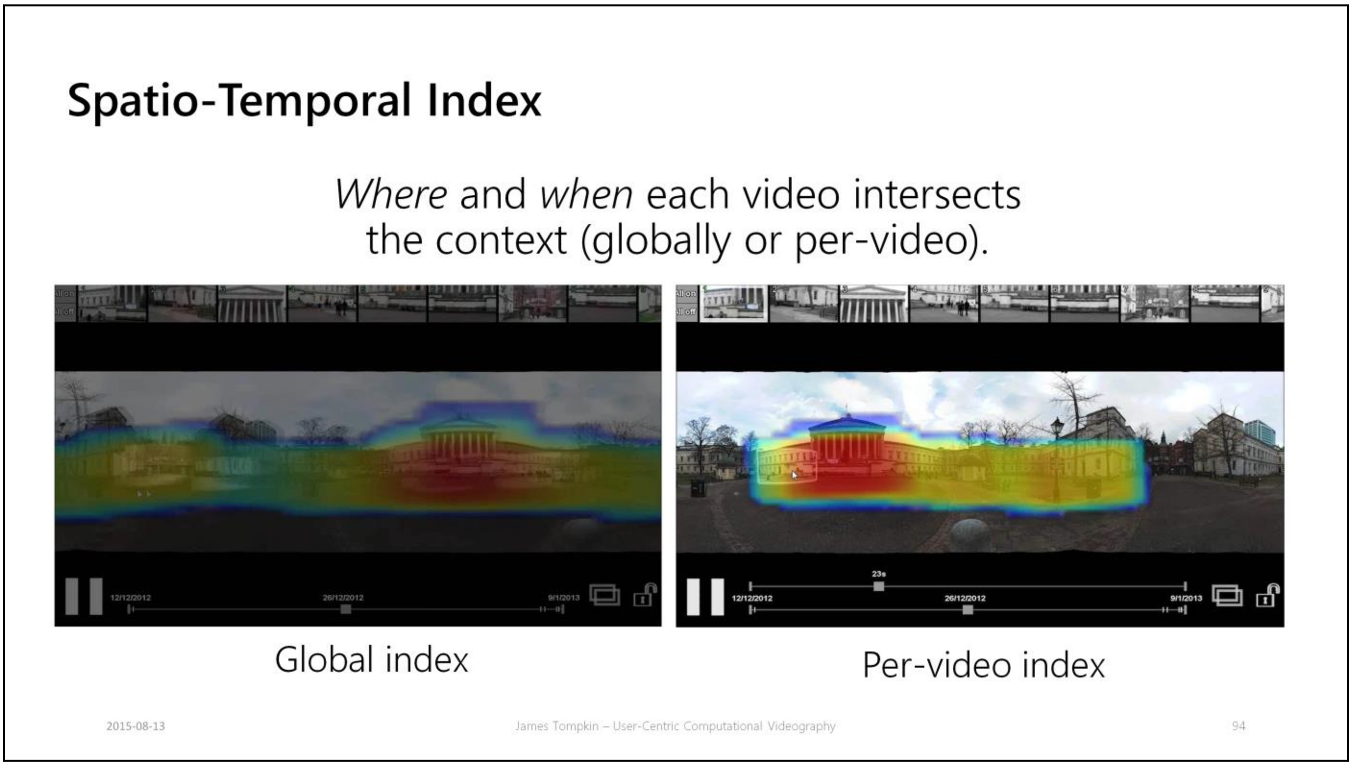
Firstly, temporally-driven interactions which enable the user to interact with a single video or the entire collection.

Users have two type of timeline with which modify the position in space and time of the videos: a local and global timeline.

Unlike a normal video player, adjusting the local timeline affects both the dynamic content within the video and the spatial position of the video in the context.

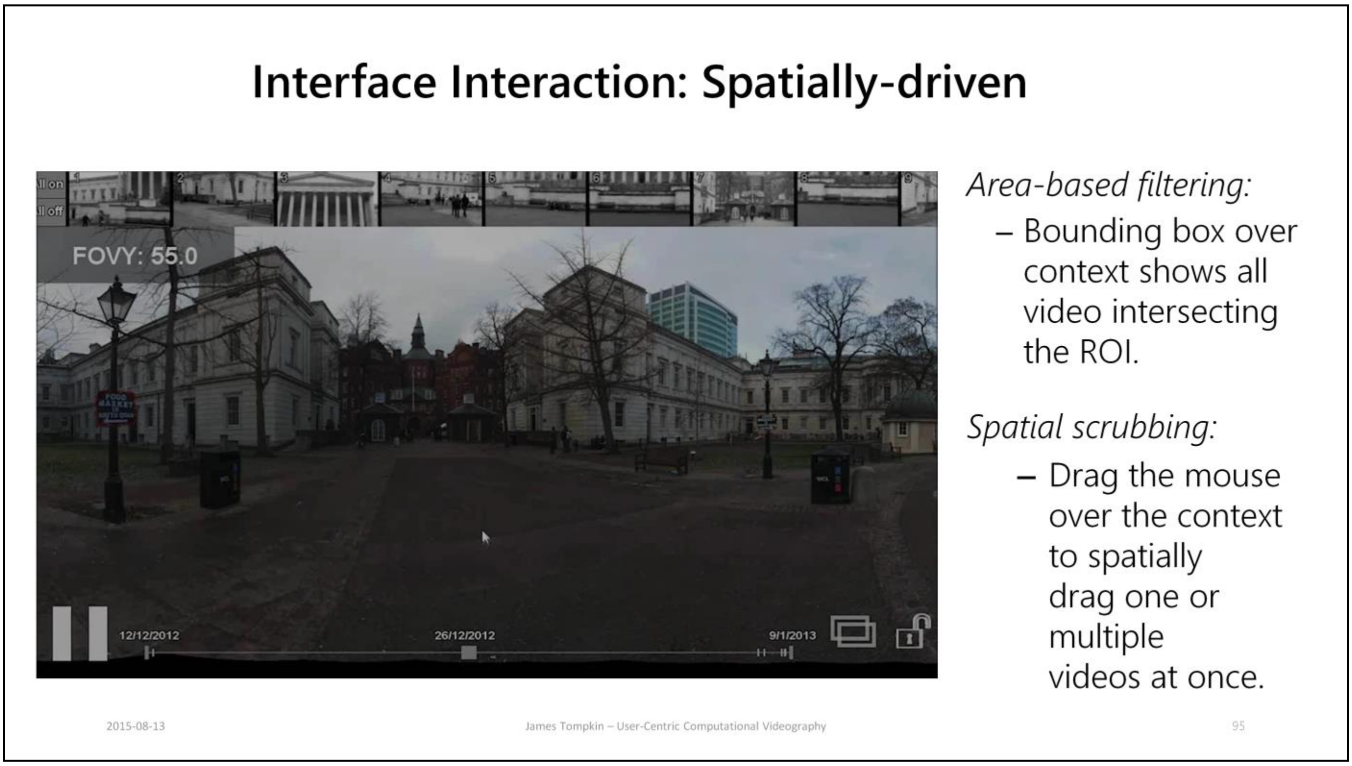
The global timeline, on the other hand, allows the user to synchronously visualise and playback video, and to filter the collection by time and date.

Unlike for outside->in video collections, here we see two videos which don't overlap visually at all, but because we're registering videos to a shared context then it's possible to compare them directly.



With all frames registered to the panorama, we can also build a spatio-temporal index which represents attention over the panoramic space.

We visualize this as an heat-map, which tells the user when and where the entire collection or each individual video intersects the context.



With this index, we can create other type of interactions are the spatially-driven:

Users can draw a bounding box on the context to reveal all videos that intersect the ROI.

Also, the user can drag a single or multiple videos at once to move them in time.

[USER STUDY NEXT!]

Video Collections in Panoramic Contexts

James Tompkin, Fabrizio Pece, Rajvi Shah, Shahram Izadi, Jan Kautz, Christian Theobalt

http://gvv.mpi-inf.mpg.de/projects/Vidicontexts/

This work has a nice feedback to the real world.

Given the system, you can then go to the real place with your tablet and pull up (spatially located) all the previously-recorded things that happened there.

Like a handheld space-time telescope! Similar in idea to ‘timescope’ idea of Naimark and others, but with modern technology.

Ultimate context

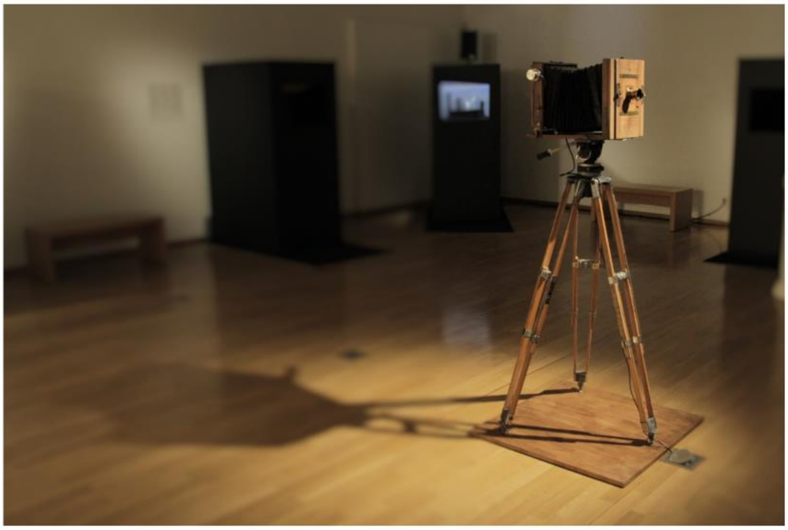
HARVARD John A. Paulson School of Engineering and Applied Sciences

- Michael Naimark – 'One World Model' idea
  - Every video camera registered and accessible.

- Planet Labs – SIGGRAPH Next
  - Satellites capturing every 5 meters of the earth once a day.

Panoramic contexts – special example

HARVARD John A. Paulson School of Engineering and Applied Sciences



Alfred Hitchcock's

Rear Window (1954)

Jeff Desom

and

James Tompkin

Take Rear Window film from Alfred Hitchcock.

Rear Window

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

[Rear Window, Paramount Pictures]

2015-08-13 James Tompkin – User-Centric Computational Videography 100

Jimmy Stewart has broken his leg; can only see out of the window of his apartment.
Discovers a murder has taken place!

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Rear Window

[Rear Window, Paramount Pictures]

Sees only through his camera lens; never sees the whole picture. The film never shows the whole panorama.

Rear Window Augmented

With Jeff Desom as artist, and similar technologies as just shown, we can reconstruct the world with these little video vignettes in place.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Rear Window Augmented

2015-08-13 James Tompkin – User-Centric Computational Videography 103

To complete the experience, we can build it into an old camera. Here, the camera ‘sees’ onto a video projection of the whole scene behind on the wall.

Rear Window Augmented

Move the camera by hand…

Rear Window Augmented

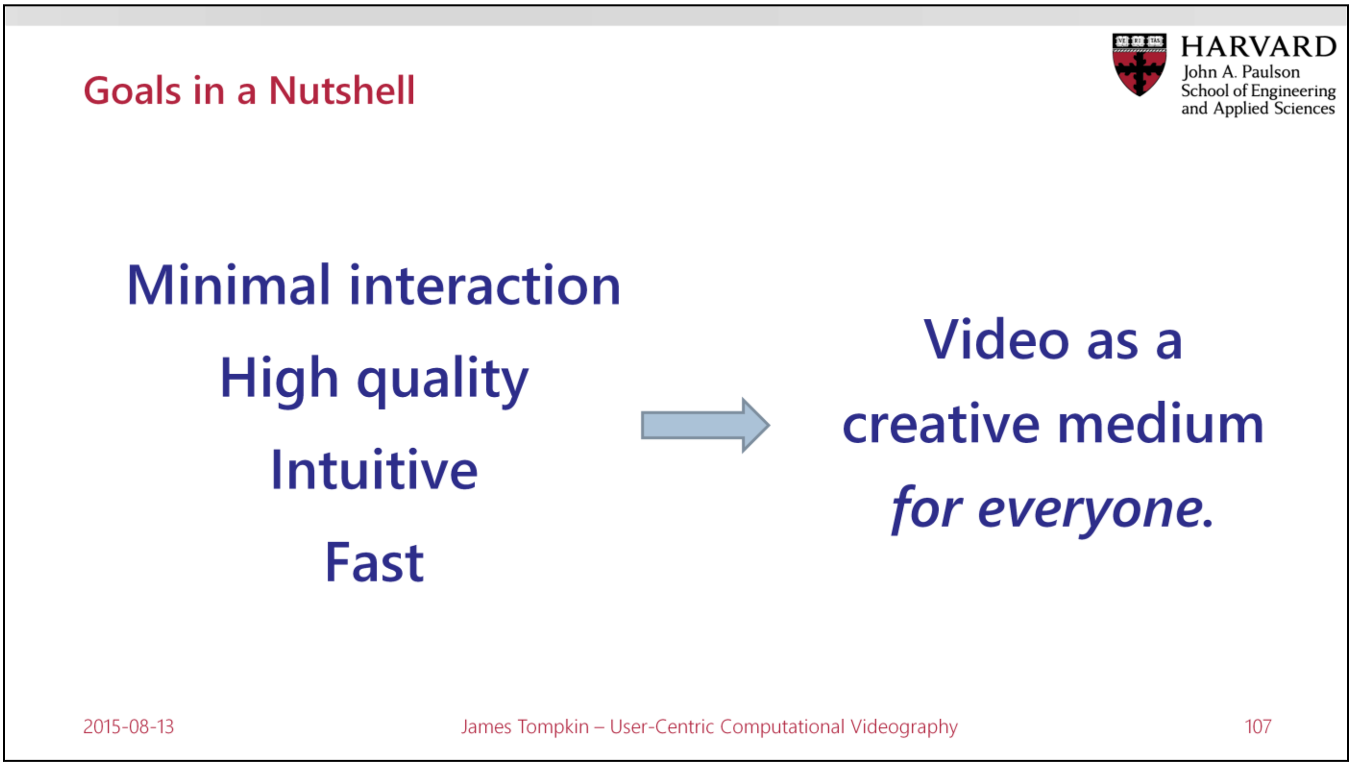
HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

2015-08-13 James Tompkin – User-Centric Computational Videography 105

Use the scroll wheel to zoom in. To complete the voyeur experience, you can push the button in like a shutter, and it’ll take a picture and upload it to twitter so that you can share it with your friends.

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Conclusion



So if we go back to the goals of what we called ‘user-centric’ video processing, we had these four criteria.

And as a community (and with others), we have made progress towards the goal of <ADVANCE> making video a creative medium for everyone, not just for large teams of people with expert skill and man-years of effort.

We want everyone to improve the capture of their moments, to express themselves creatively with video, and to explore videos and video collections in compelling ways.

In correcting capture ‘errors’ / ‘features’

Editing / authoring

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Minimal interaction - High quality – Intuitive - Fast

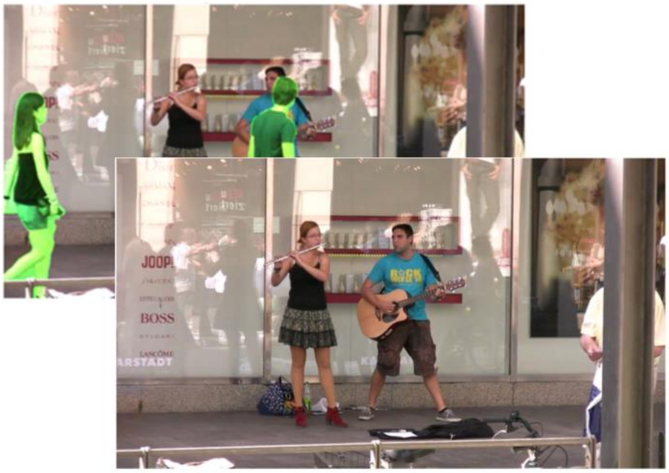
In making advanced edits with lightweight assumptions

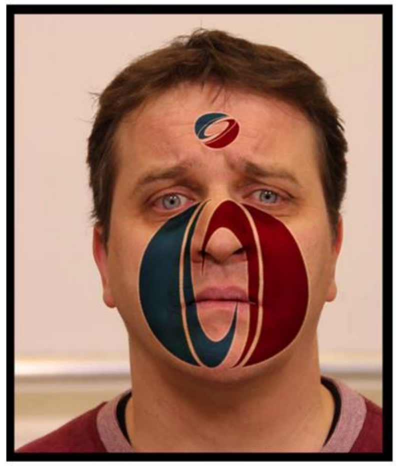
In discovering real world scene properties through analysis to allow editing

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Editing / authoring

Minimal interaction - High quality – Intuitive - Fast





2015-08-13 James Tompkin – User-Centric Computational Videography 110

In making advanced edits with lightweight assumptions

In discovering real world scene properties through analysis to allow editing

Viewing / exploring

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Minimal interaction - High quality – Intuitive - Fast

In my section on exploring video, we just saw new ways to find and manipulate the events within videos.

Very little or no interaction we can achieve unstructured video-based rendering to explore views onto a scene, or automatically edit sports footage to follow the action.

In creating new presentation forms

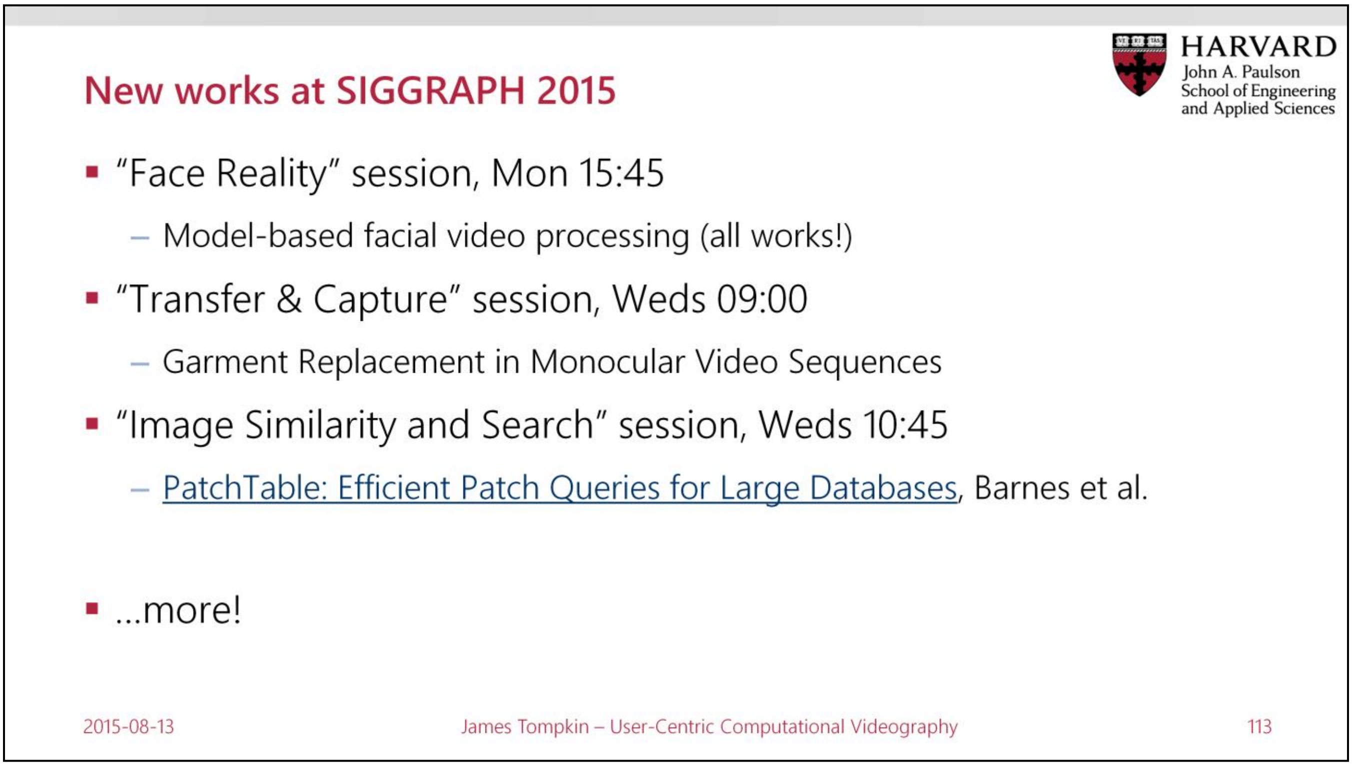
In exploring videos and video collections

New works at SIGGRAPH 2015

HARVARD John A. Paulson School of Engineering and Applied Sciences

- "Video Processing" session, Tues 14:00
  - Sampling Based Scene-space Video Processing, Klose et al.
  - Gaze-driven Video Re-editing, Jain et al.
  - AudeoSynth: Music-driven Video Montage, Liao et al.
  - High-quality Streamable Free-viewpoint video (model-based), Collet et al.
- "Let's Do The Time Warp" session, Tues 10:45
  - Recovering Painted Strokes from Time-Lapse Video,
  - Time-lapse Mining from Internet Photos,
  - Real-time Hyperlapse Creation via Optimal Frame Selection

Just this conference, many new works on video:



PatchTable – showed artistic video stylization

I think the quality of these works shows that as a community we are making great strides towards our outlined goals, but also that we are not running out of steam and that there are compelling challenges ahead – there are new ideas to improve the quality, editing capability and ease of use, and applications of videos.

How close are we?

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

2015-08-13 James Tompkin – User-Centric Computational Videography 114

Foreground / background separation with pixel-perfect, matted video segmentation from lazy, inaccurate input

Convincing structure-aware inpainting

Temporally consistent

Interactive rates.

With lightweight models, in limited cases, we can do all of these, but the techniques are still often brittle.

Critically, we are missing robustness and generality. In many cases we are moving towards the use of explicit model-based methods to more correctly model the scene and provides steps towards this robustness and generality.

Further still, even this future demonstration is quite a limited application of video cut and paste, but hopefully over this course we’ve shown that there is much more that could be done. For instance, we could try to move the camera in this edit, we could try to changed the appearance of the objects or insert new objects seamlessly, or we could try and move beyond the standard output presentation of this material as typical video, and perhaps use more free-form exploration, either for single and multiple videos, even potentially providing the geospatial context of the field situated in the world.

Now, there’s still a long way to go to achieve all of this, and it will require advanced in computer vision for scene analysis, computer graphics for appearance modelling, and in interaction for intuitive techniques, BUT, in showing you these techniques perhaps it seems feasible, and you will be inspired to work (or to continue to work) to help achieve it.

Thank you!

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

- Jiamin thanks Ravi Ramamoorthi, Maneesh Agrawala, Aseem Agarwala, A*STAR NSS PhD Fellowship, Light. Co.
- Christian R. and Christian T. thank Intel Visual Computing Institute.
- Christian T. also thanks Bonneel et al. for SIGGRAPH Asia 2014 slides.
- James thanks NSF CGV-1110955 – ‘Analyzing Images Through Time’

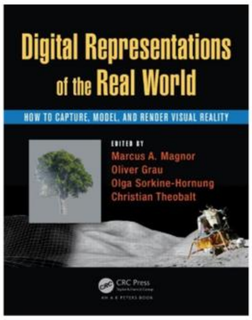
intel VISUAL COMPUTING INSTITUTE

erc

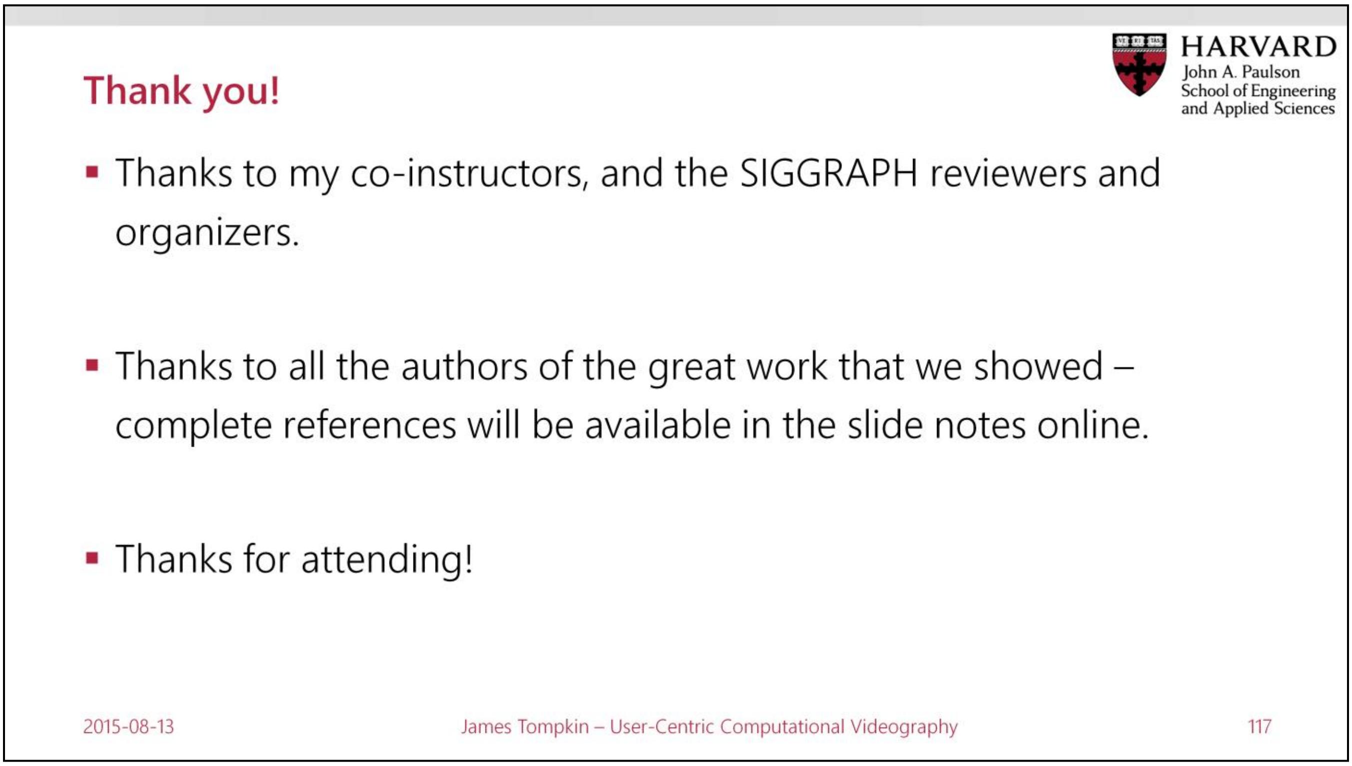
ERC Starting Grant
CapReal

Before we open up the floor, we’d just like to thank

Resource for model-based work



Marcus Magnor, Oliver Grau, Olga Sorkine-Hornung, and Christian Theobalt (Eds.)
Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality
Taylor and Francis (AK Peters), 2015



Thanks to Jiamin Bai and Christian Theobalt my co-instructors, and especially thanks to Christian Richardt for organizing the course and keeping us all in order.

We're excited to hear what you think about the state of video processing

Course website:
http://gvv.mpi-inf.mpg.de/teaching/uccv_course_2015/