Transcript-Based Manipulation and Browsing

Tools for Placing Cuts and Transitions in Interview Video. Floraine Berthouzoz, Wilmot Li and Maneesh Agrawala, SIGGRAPH 2012.

Raw Footage: Hours of Material

Editing Process

Step 1: Index Footage

:40 Some of our customers that have been working in print for many years ...

1:15 So what we're trying to do ... (Noise)

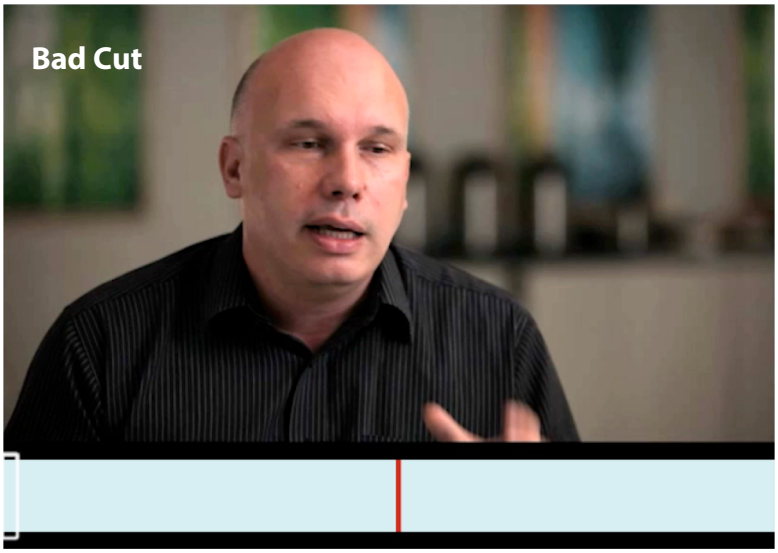
1:28 So what we're trying to do is to bring to our customers the tools that will allow them to express sophisticated layout ...

3:45 Now you'll be able to have much more creative layouts ... you can have as many regions as you want ... and the text will be able to flow across all of those regi

1:02 If the orientation of your device cha from landscape to portrait ... the layout and text will reflow across all those different regions ...

7:23 we have a fantastic engineering team ...

Step 2: Find Good Cut Locations



Not gesturing

Not in the middle of saying a word

Step 3: Insert Transition

Jump Cut | Fade-to-Black | Zoom-In

Visible Transition

Jump/Fade seen as time passing (short/long)

Zoom seen as continuous action from multiple cameras

Step 3: Insert Transition

Optical Flow Interpolation

Hidden Transition

Seen as single continuous shot

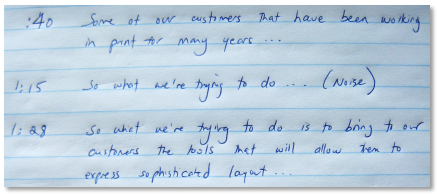
Difficult to smoothly interpolate large changes

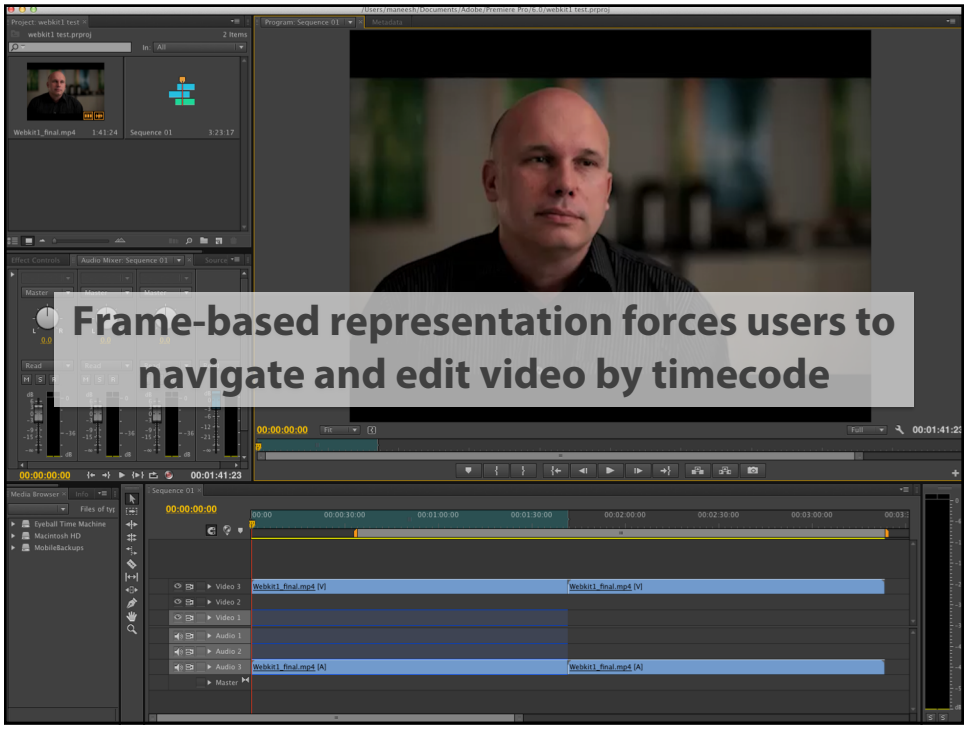
Current Problems

1. **Index footage**
   Must listen many times (very slow)

2. **Find good cut locations**
   Requires additional scrubbing

3. **Hidden transition**
   Interpolation introduces artifacts





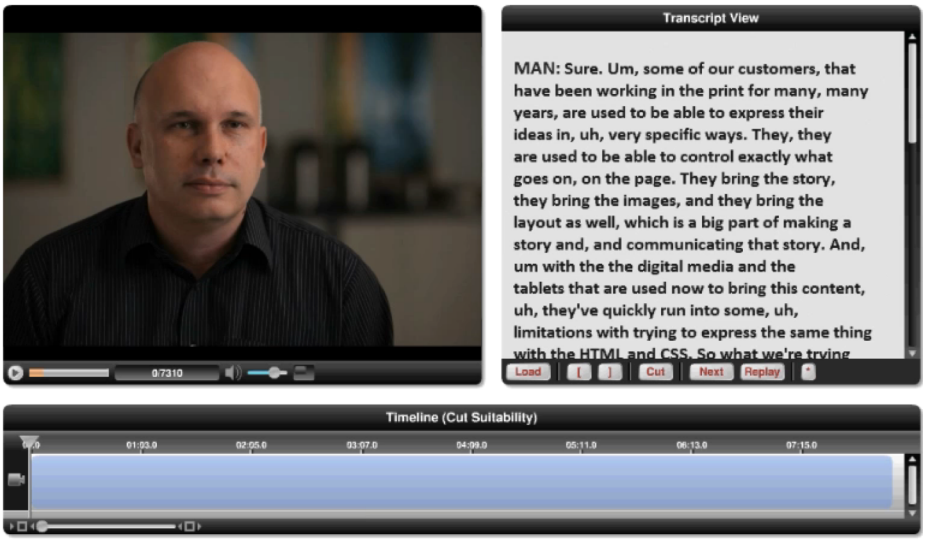
Conceptual Model

Man: Sure. Um, some of our customers, that have been working in the print for many, many years, are used to being able to express their ideas in, uh, very specific ways. They, they are used to being able to control exactly what goes on, on the page. They bring the story, they bring the images, and they bring the layout as well, which is a big part of making a story and, and communicating that story.

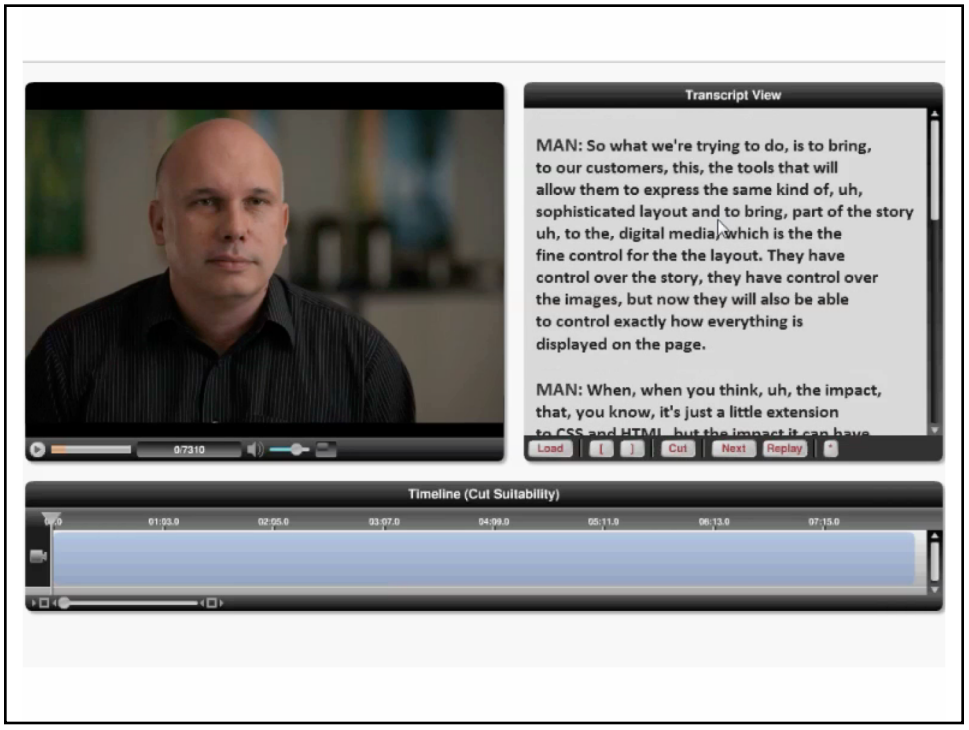
And, um, with the, the digital media and the tablets, that are used now to bring this content, uh, they've quickly run into some, uh, limitations with trying to express the same thing with the HTML and CSS.

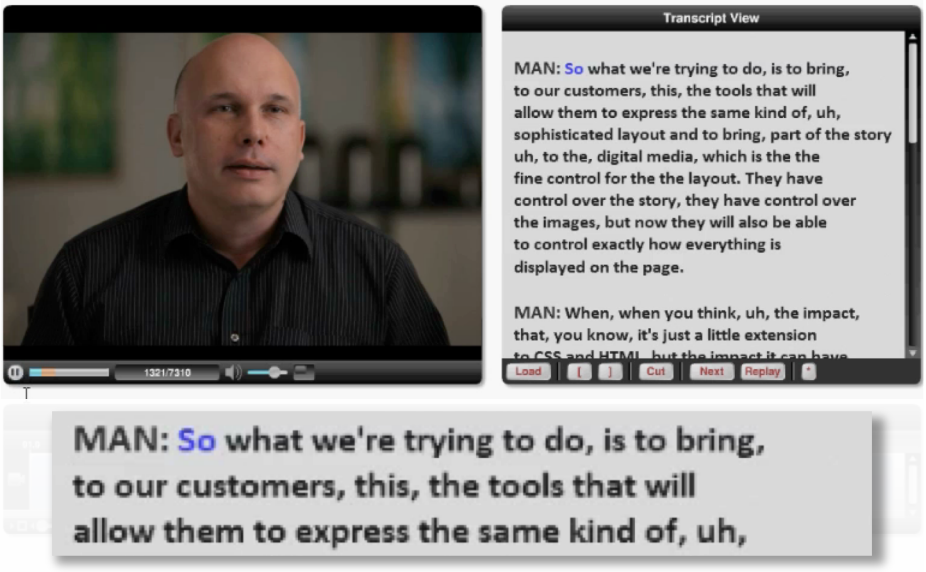
People think of interviews in terms of content

Goal: Let users navigate and edit video using this higher-level transcript-based representation

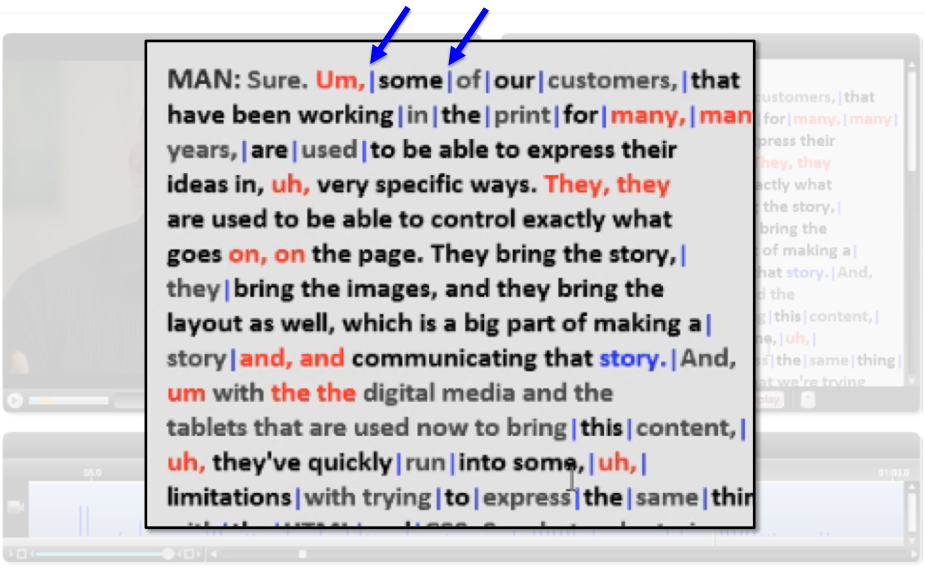



Player
Transcript View
Transcript View
MAN: Sure. Um, some of our customers, that have been working in the print for many, many years, are used to be able to express their ideas in, uh, very specific ways. They, they are used to be able to control exactly what goes on, they bring the images, and they bring the layout as well, which is a big part of making a story and, and communicating that story. And, um with the the digital media and the tablets that are used now to bring this content, uh, they've quickly run into some, uh, limitations with trying to express the same thing with the HTML and CSS. So what we're trying
Load
Cut
Next
Replay
Timeline (Cut Suitability)
Timeline

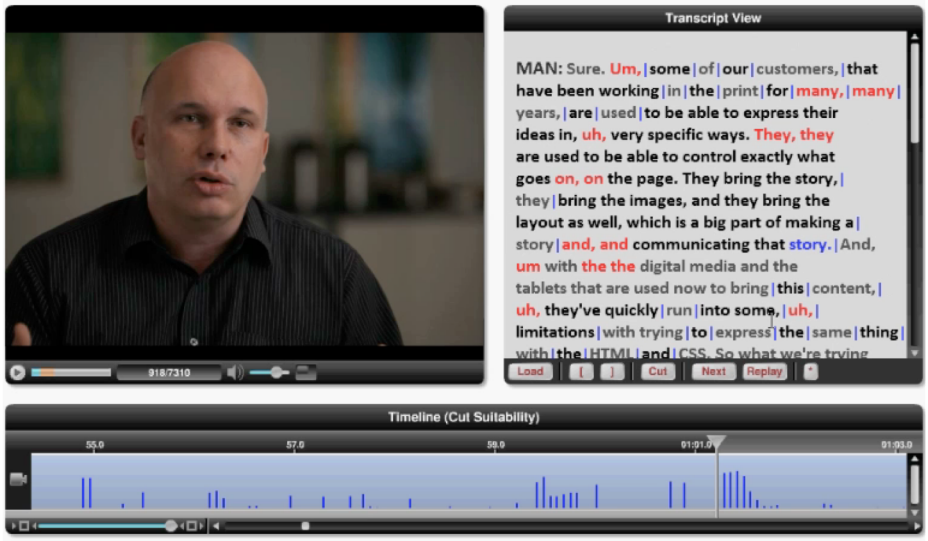

Transcript View
MAN: So what we're trying to do, is to bring, to our customers, this, the tools that will allow them to express the same kind of, uh, sophisticated layout and to bring, part of the story uh, to the, digital media, which is the the fine control for the the layout. They have control over the story, they have control over the images, but now they will also be able to control exactly how everything is displayed on the page.
MAN: When, when you think, uh, the impact, that, you know, it's just a little extension
Load
Cut
Next
Replay
Timeline (Cut Suitability)

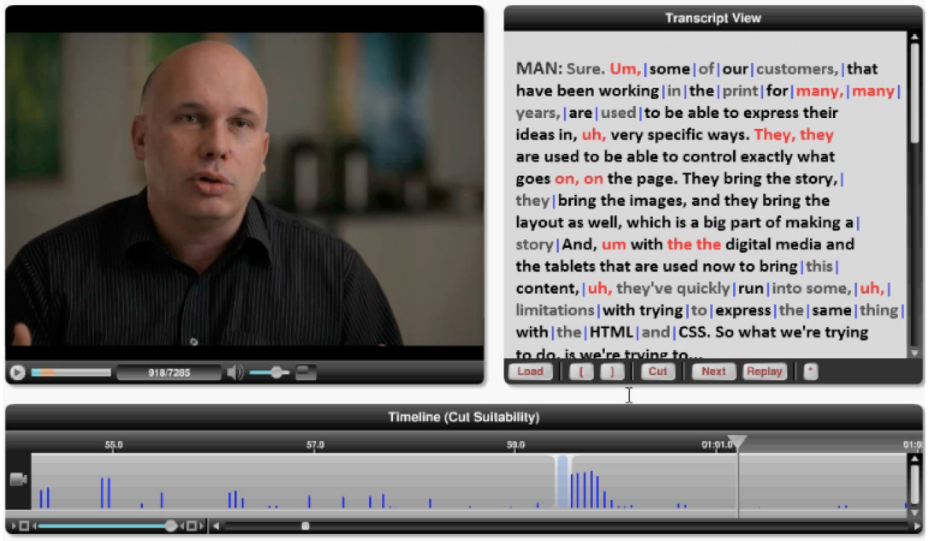


Transcript View Zoom In

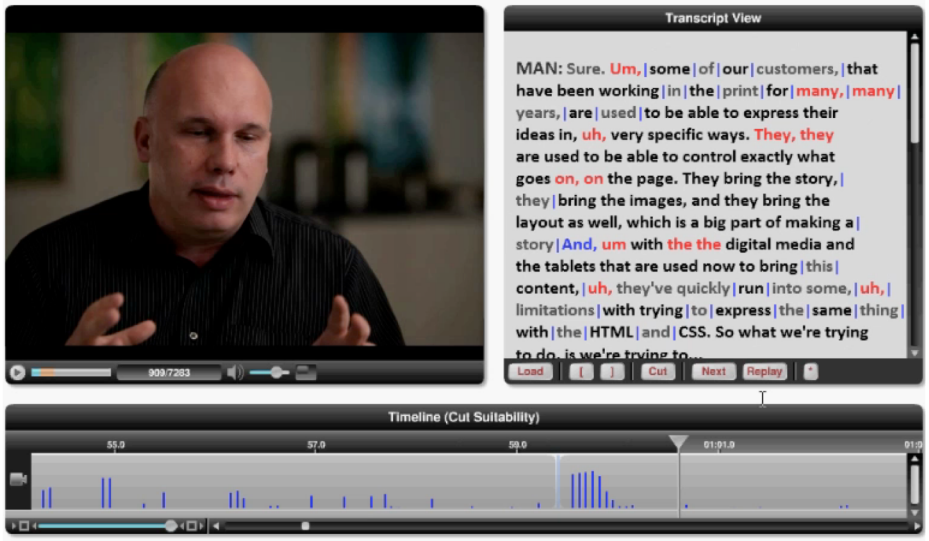


Transcript View Zoom In

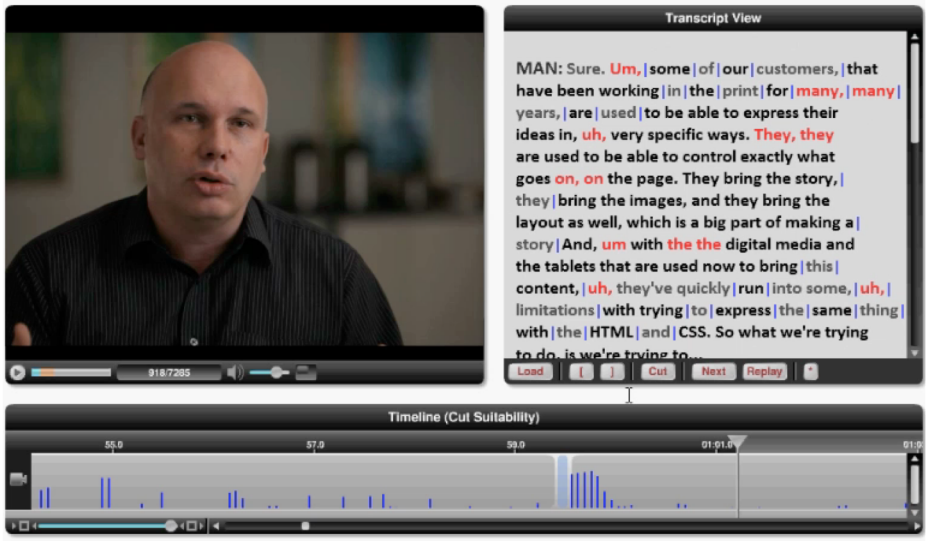




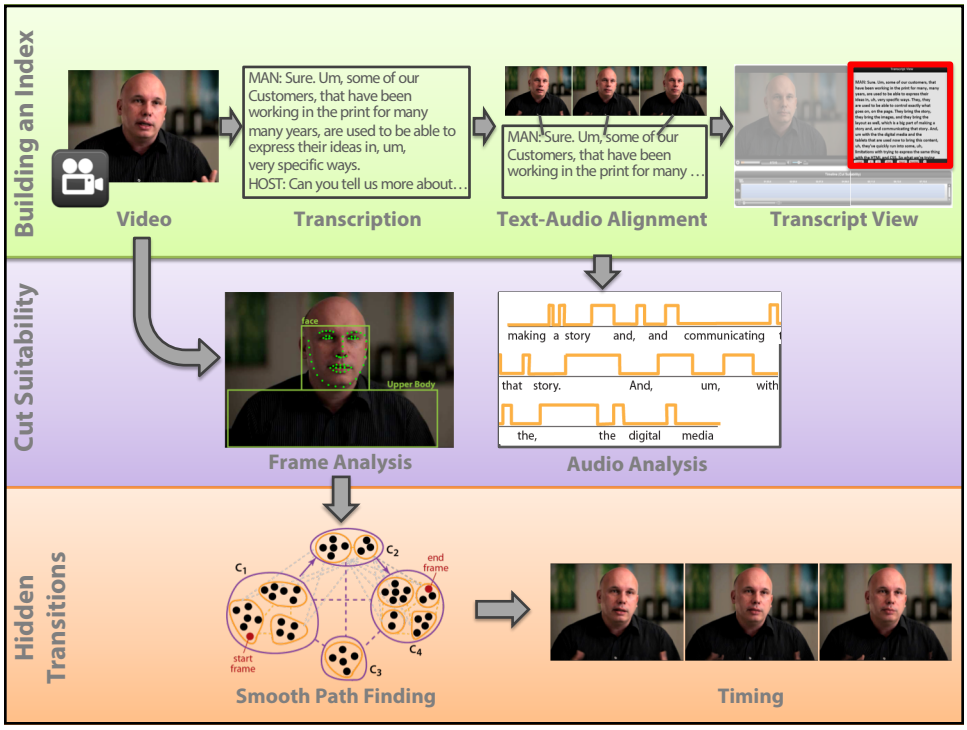
Hidden Transition

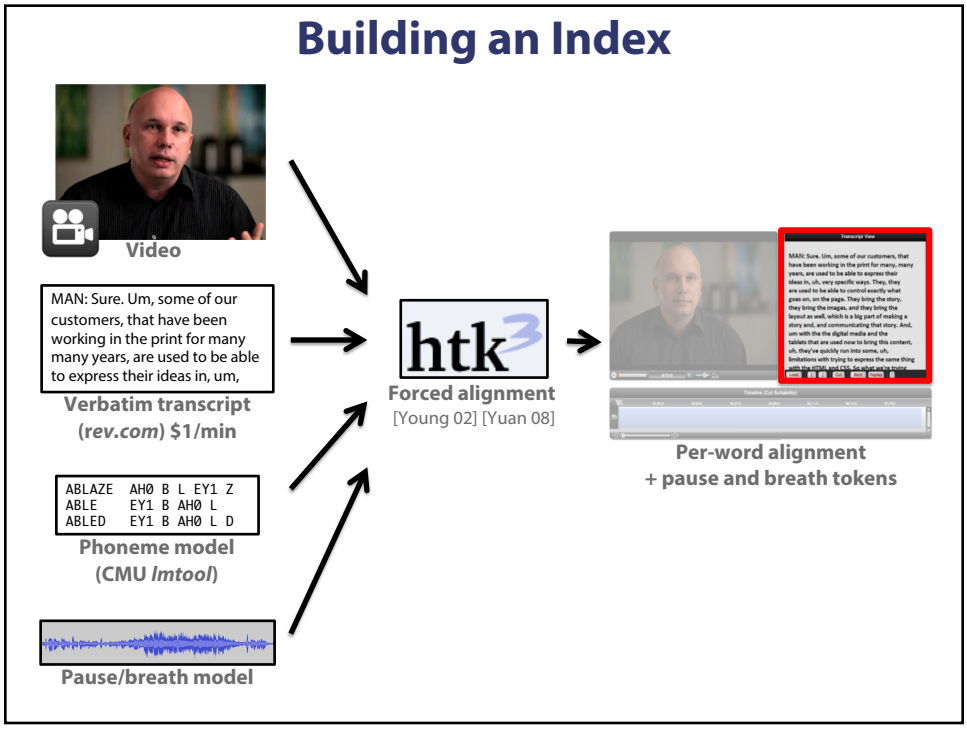


Jump Cut

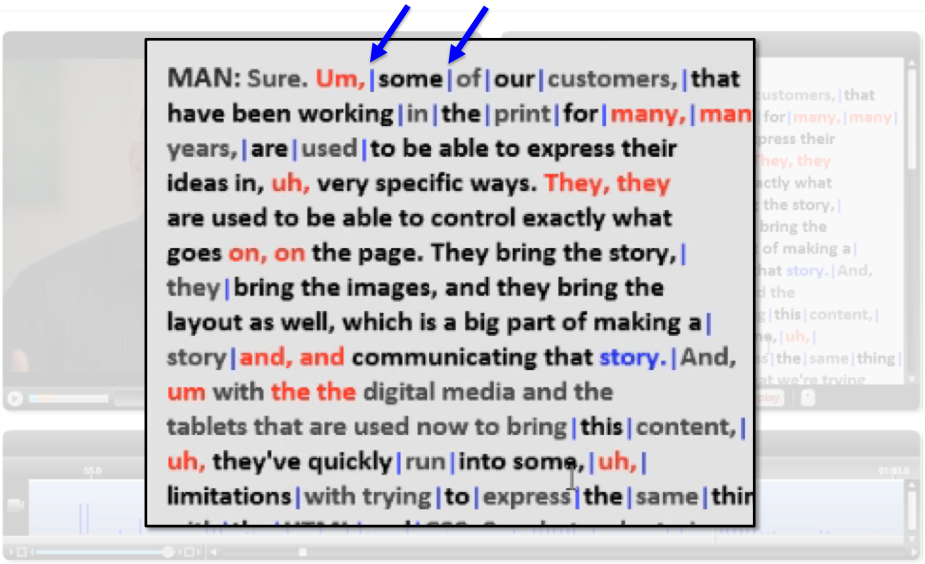


Hidden Transition


Building an Index
Video
MAN: Sure. Um, some of our Customers, that have been working in the print for many many years, are used to be able to express their ideas in, um, very specific ways. HOST: Can you tell us more about…
Transcription
MAN: Sure. Um, some of our Customers, that have been working in the print for many …
Text-Audio Alignment
Transcript View
Cut Suitability
face
Upper Body
Frame Analysis
making a story and, and communicating
that story. And, um, with
the, the digital media
Audio Analysis
Hidden Transitions
C1
C2
C3
C4
end frame
start frame
Smooth Path Finding
Timing


Building an Index
Video
MAN: Sure. Um, some of our customers, that have been working in the print for many many years, are used to be able to express their ideas in, um,
Verbatim transcript
(rev.com) $1/min
ABLAZE AH0 B L EY1 Z
ABLE EY1 B AH0 L
ABLED EY1 B AH0 L D
Phoneme model
(CMU lmtool)
Pause/breath model
htk3
Forced alignment
[Young 02] [Yuan 08]
Per-word alignment
+ pause and breath tokens

Cut Suitability Score



Scoring

Not gesturing

Not in the middle of saying a word

$$Score(i) = S_v(i)S_a(i)$$

Visual Score

Audio Score

Visual Frame Distance

$$Distance(i,j) = Spatial\ Distance + Appearance\ Distance$$

Spatial Distance

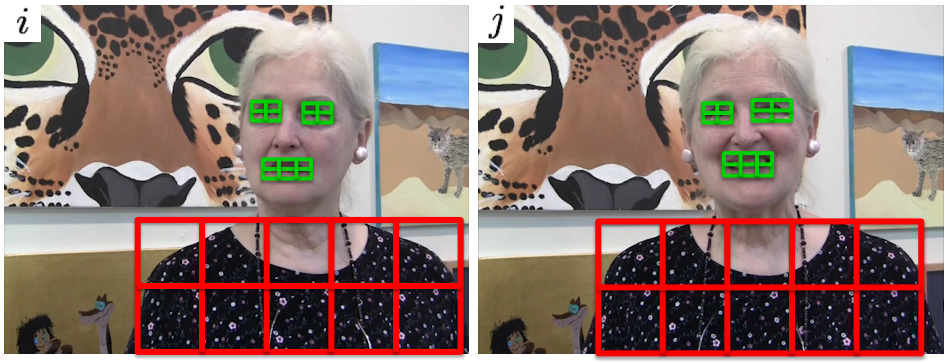
Facial Feature Tracking [Saragih 09]

Spatial Distance

Facial Feature Tracking [Saragih 09]

$$Spatial\ Distance(i,j) = ||p_i - p_j||$$

Appearance Distance

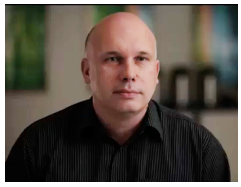


Compute histogram of oriented gradients (HOG) feature for each region

$$Appearance\ Distance(i,j) \simeq \chi^2_{i,j}(eyes) \times \chi^2_{i,j}(mouth) \times \chi^2_{i,j}(body)$$

Based on [Kemelmacher-Shlizerman 11]

Visual Score



Reference Frame r

$$S_v(i) = 1 - D_{app}(r, i)$$

visual

making a story and, and communicating that story. And,

Audio Score

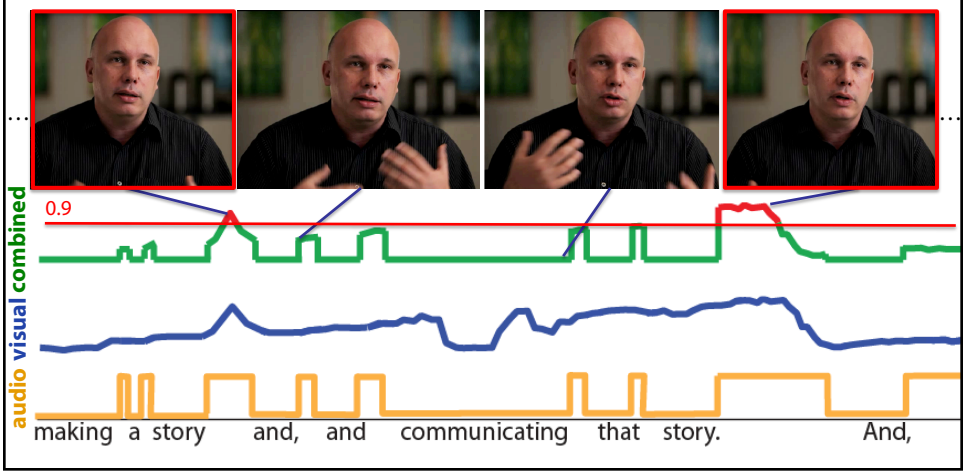
$$S_a(i) = \begin{cases} 1 & \text{frame } i \text{ between two words in transcript.} \\ 0 & \text{otherwise} \end{cases}$$

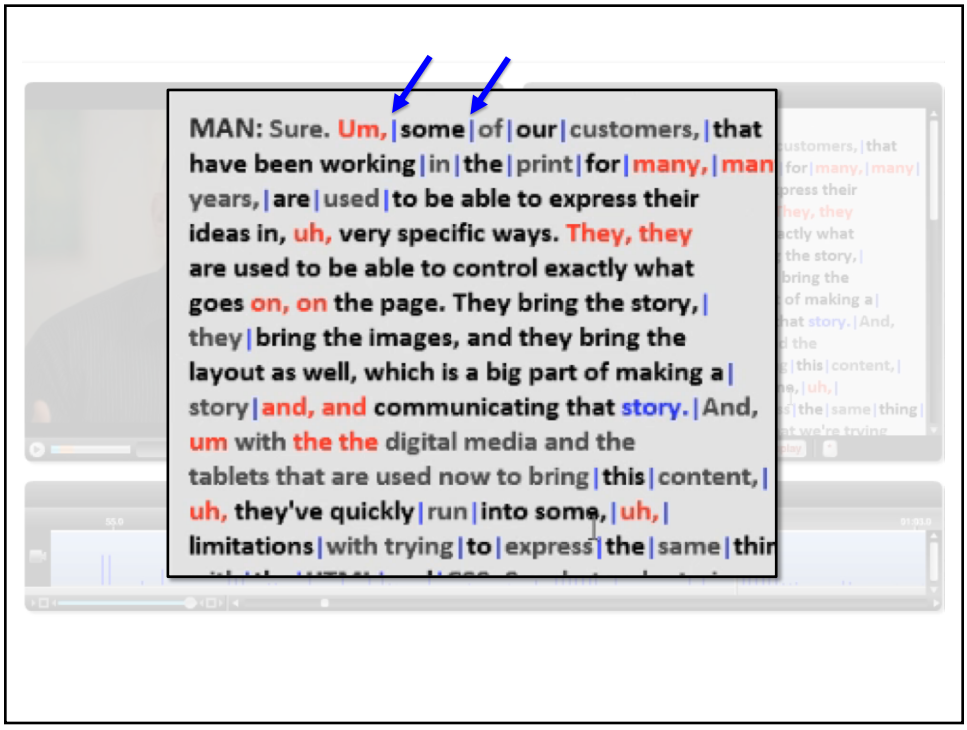
audio visual

making a story and, and communicating that story. And,

Cut Suitability Score

$$Score(i) = S_v(i) S_a(i) > 0.9$$



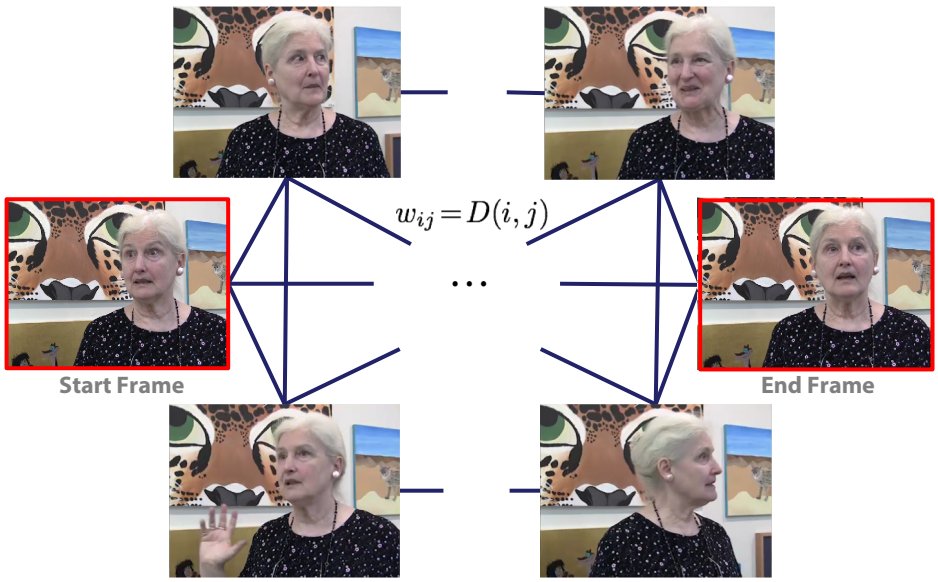


Hidden Transitions

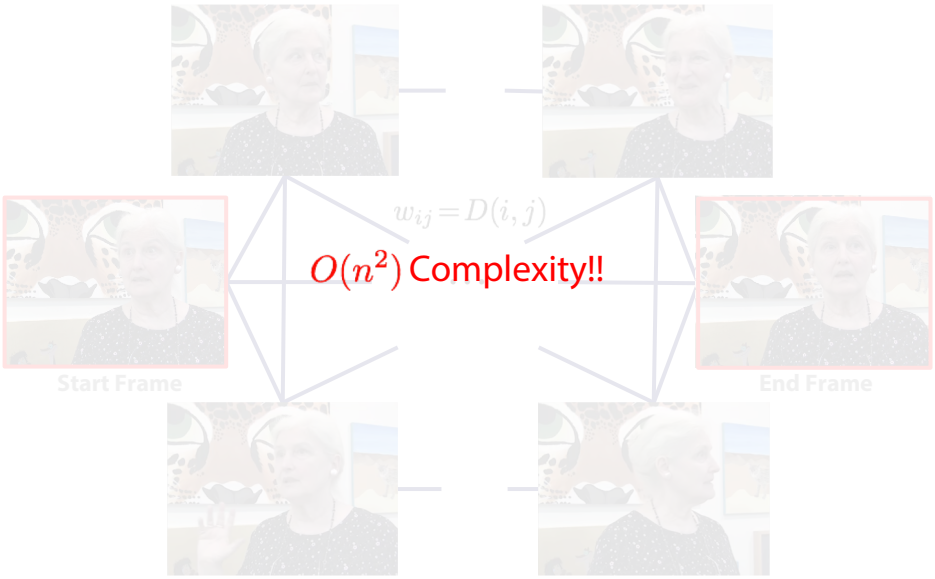
Jump Cut

Hidden Transition

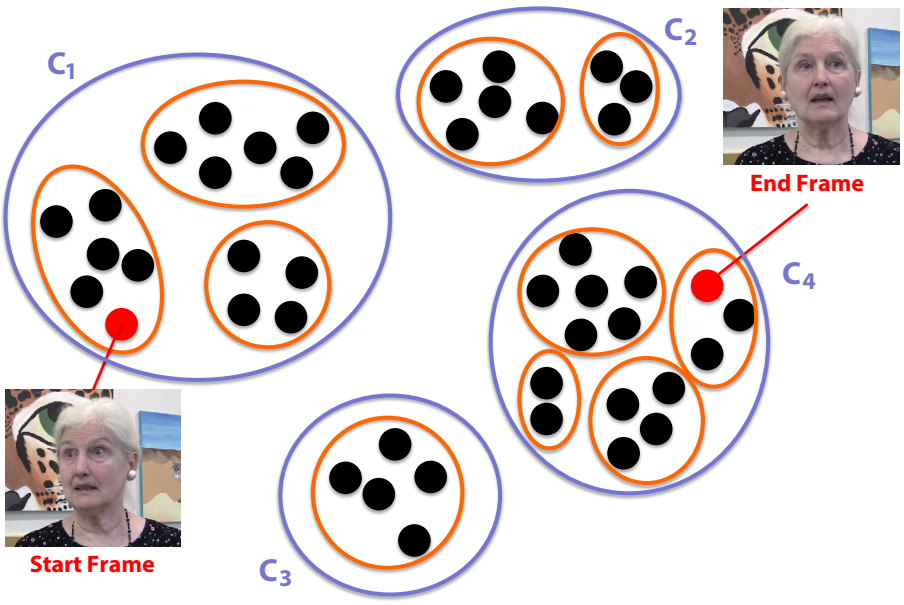
Brute-Force Approach



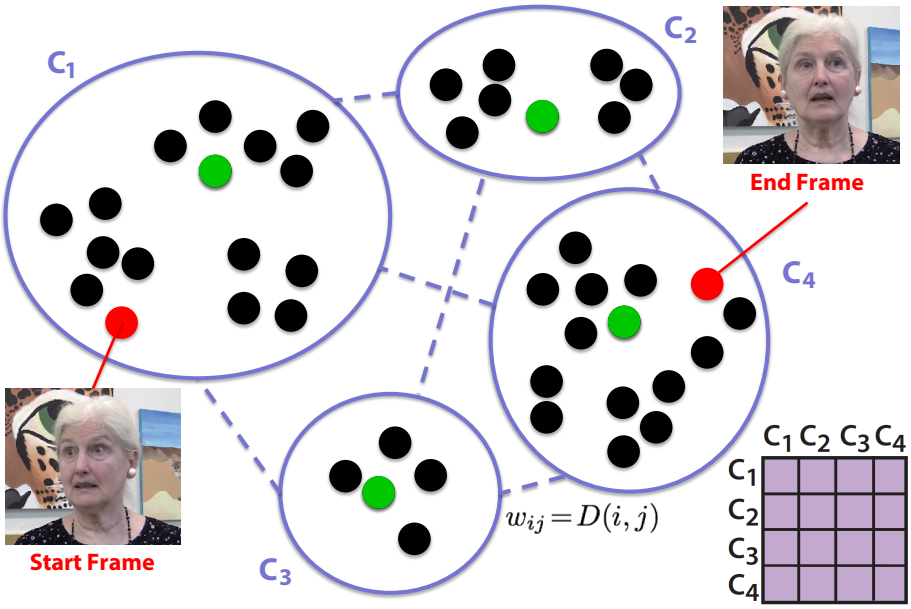
Brute-Force Approach



Hierarchical Clustering

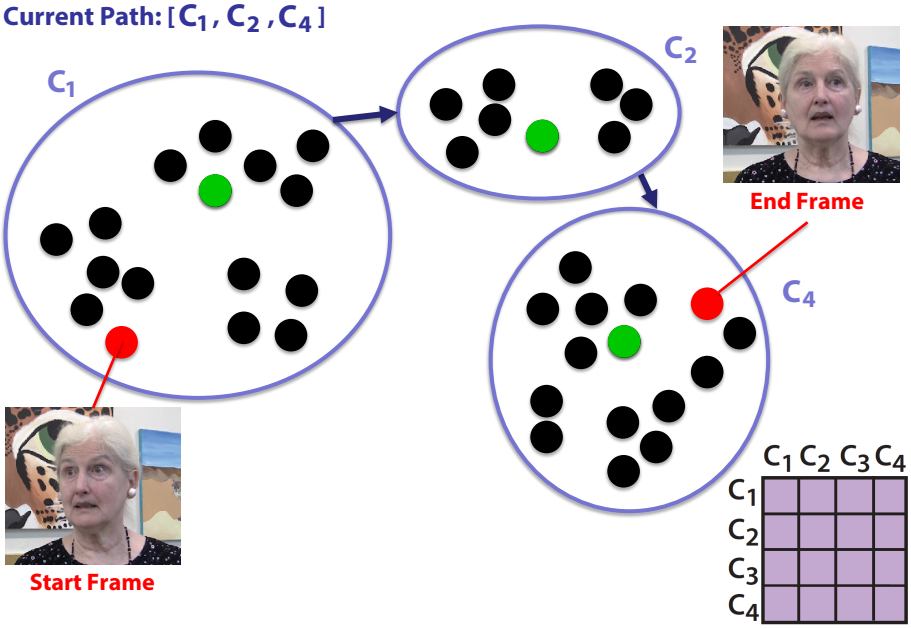


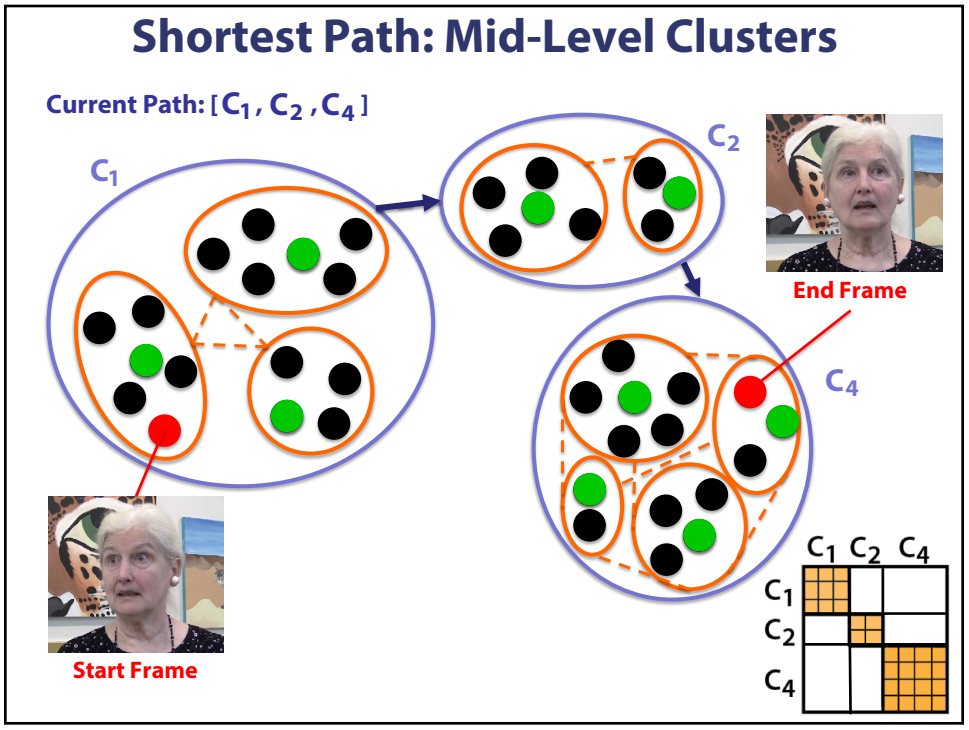
Shortest Path: Highest Level Clusters

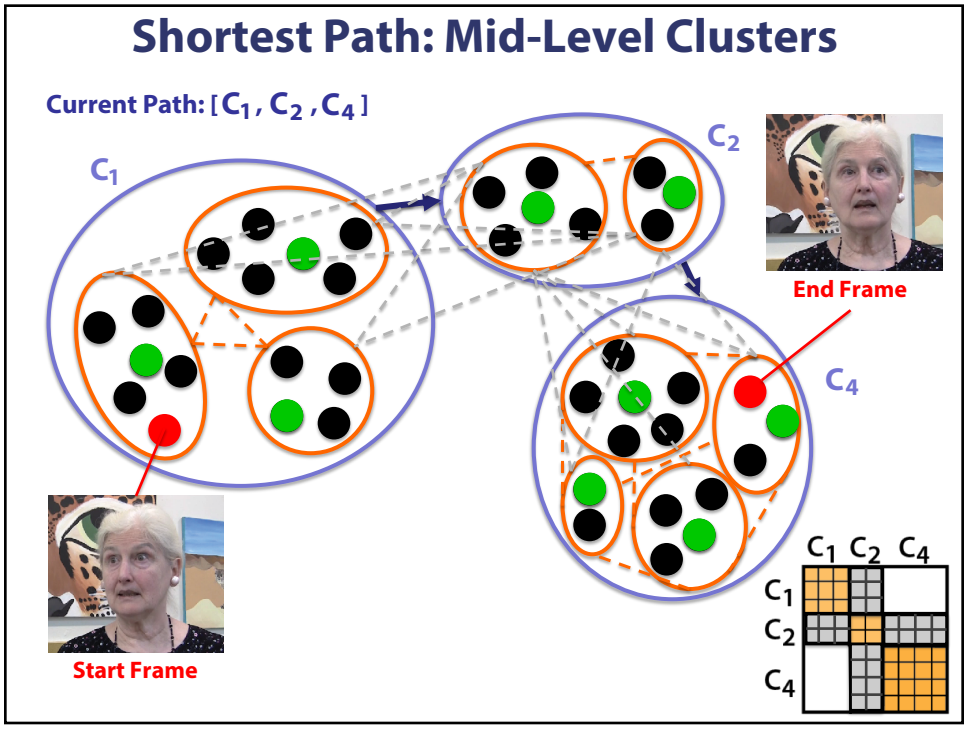


Shortest Path: Highest Level Clusters

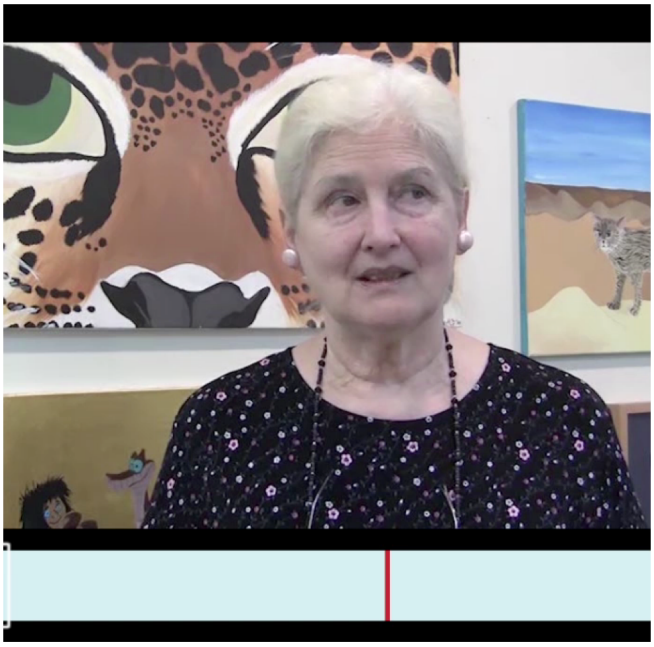
Current Path: [C_1, C_2, C_4]




Shortest Path: Mid-Level Clusters
Current Path: [C1, C2, C4]
C1
C2
C4
End Frame
Start Frame
C1 C2 C4
C1
C2
C4


Shortest Path: Mid-Level Clusters
Current Path: [C1, C2, C4]
C1
C2
C4
End Frame
Start Frame
C1 C2 C4
C1
C2
C4

Jump-Cut



Shortest Path Transition

Shortest Path

Shortest Path In-Between Frames

Optical Flow Interpolation

Shortest Path In-Between Frames

Interpolation Frames [Brox 04]

Data-Driven Timing

Raw Footage Frames

$N(i,j) = 2$

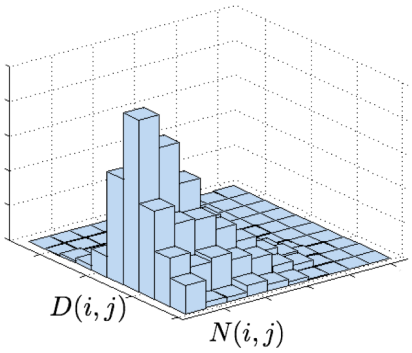
$D(i,j) = 0.02$

Number of intermediate frames

$$N(i,j) = |i - j|$$

Frame distance

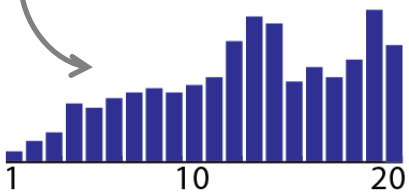
$$D(i,j)$$



Setting Number of Frames

Shortest Path In-Between Frames

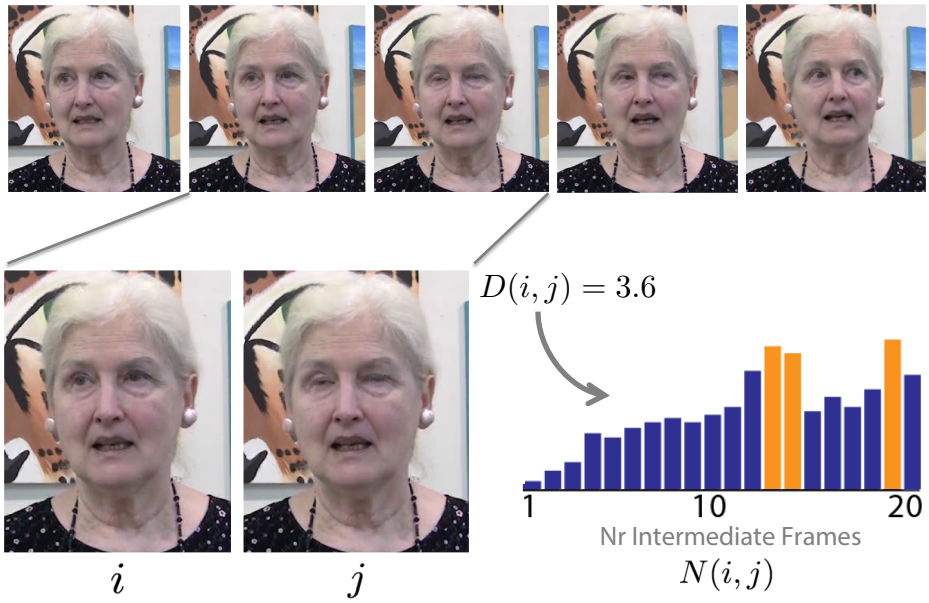
$D(i,j) = 3.6$



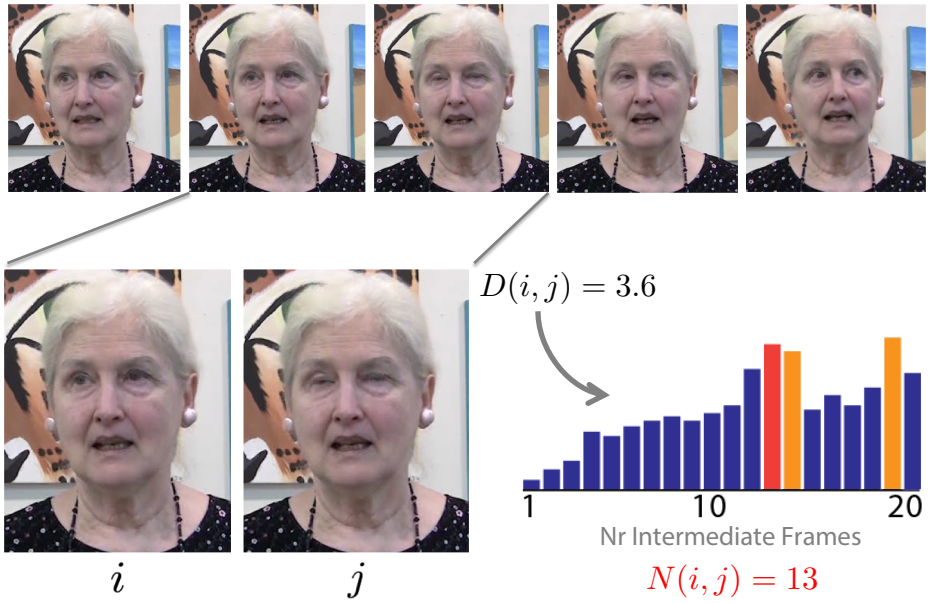
Nr Intermediate Frames

$N(i,j)$

Setting Number of Frames

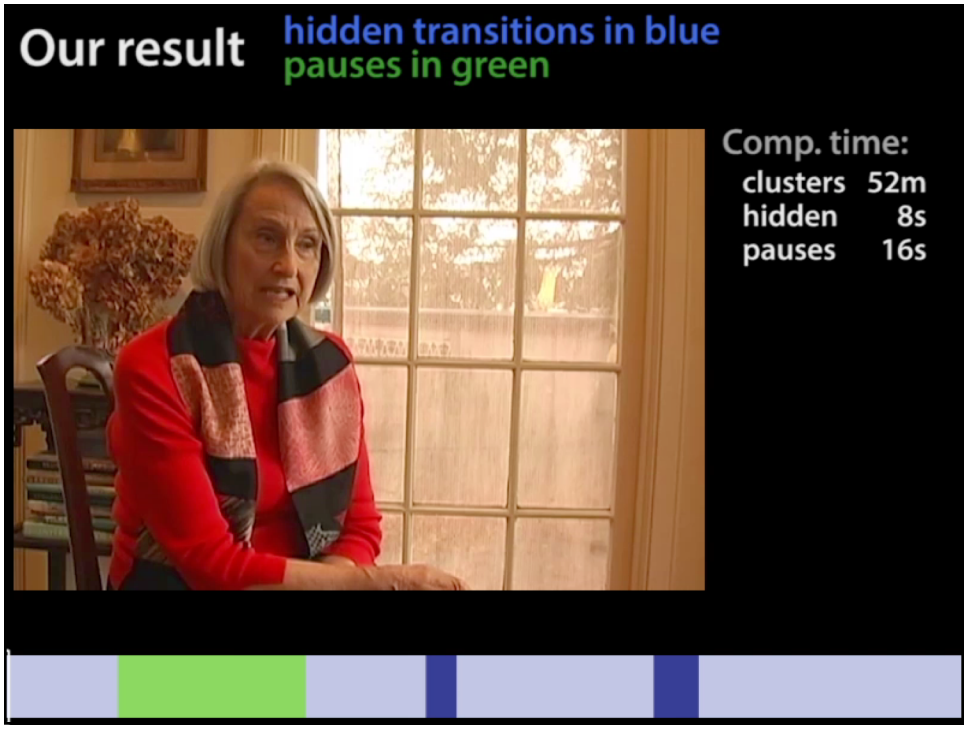


Setting Number of Frames

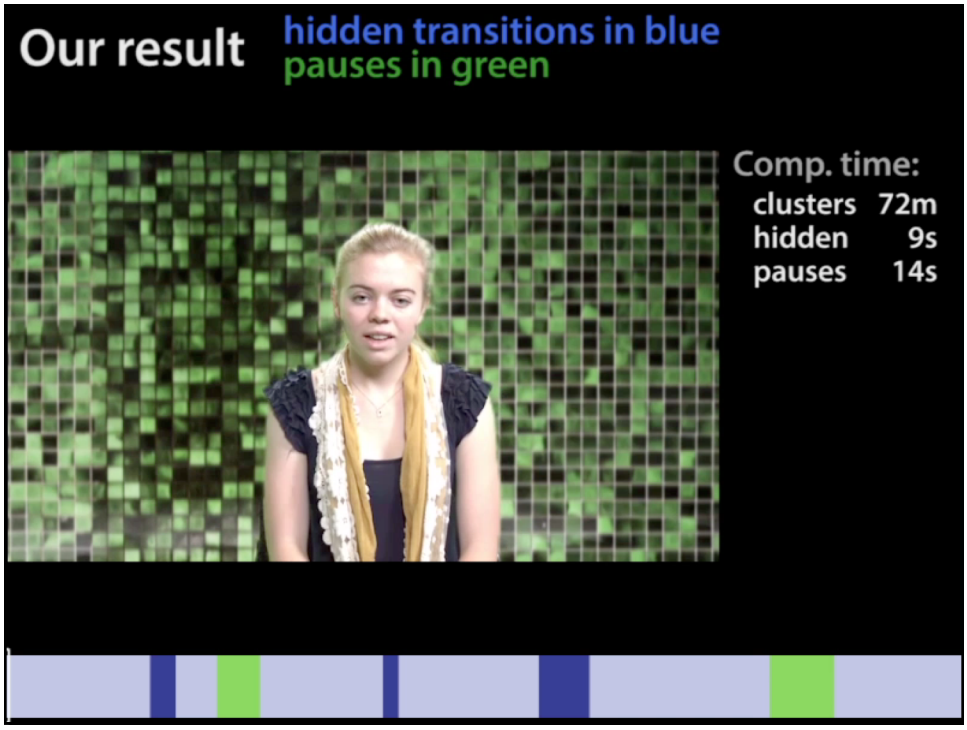


Hidden Transition

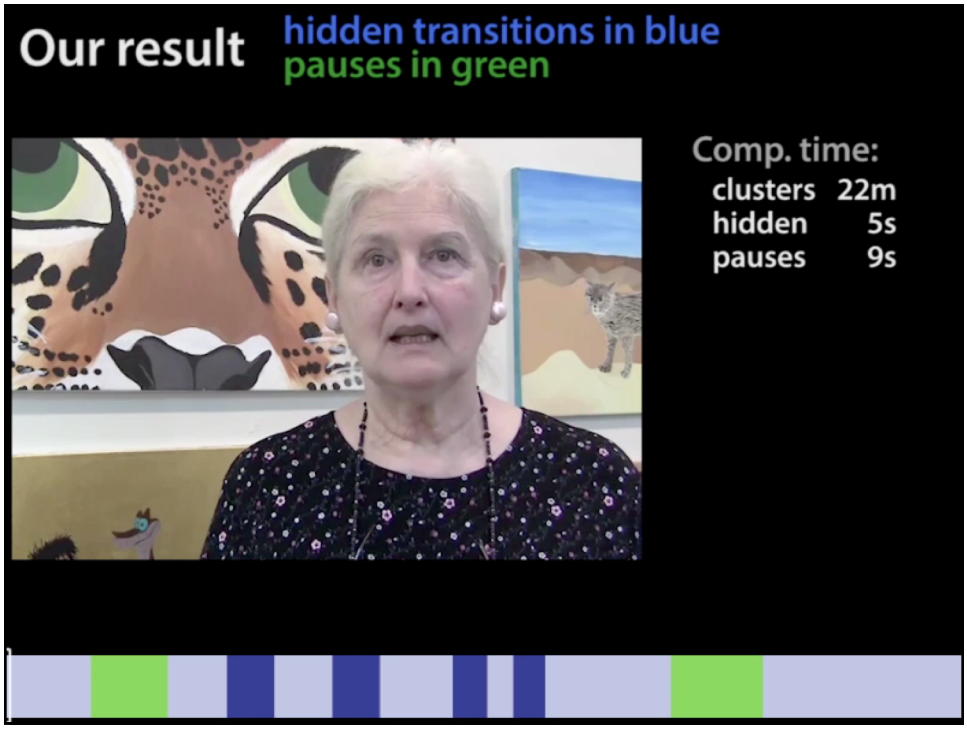
Jump cuts (in red)


Our result
hidden transitions in blue
pauses in green
Comp. time:
clusters 52m
hidden 8s
pauses 16s


Jump cuts (in red)


Our result
hidden transitions in blue
pauses in green
Comp. time:
clusters 72m
hidden 9s
pauses 14s


Jump cuts (in red)



Limitations

Poor hidden transitions when endpoints very different
Assumptions: Still background, fixed lighting and camera position

User Feedback

Professional Users

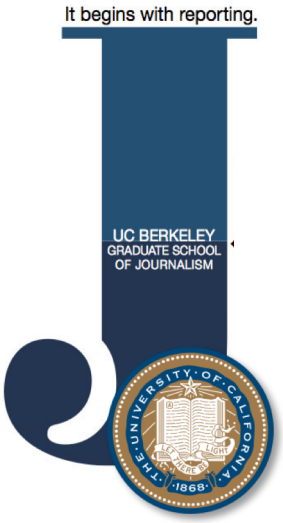
4 Video editors
5 Journalists

Interface

Would fit in existing workflow
Transcript view extremely useful
Easy to find good cut locations
Could quickly test different cut possibilities

Hidden Transitions

Useful alternative to visible transitions
Good for Web



Takeaways

Choose the *right* representation

Transcript much better than frames for navigation and editing of interview video

Principle of congruence

For effective interfaces, structure of external representation should match structure of mental representation

[Tversky 02] [Norman 86, 88]

What is appropriate representation for other types of video?