FOR TECHNOLOGIES

presents

MEDIA OVER IP

SYMPOSIUM 2018

In partnership with:

aws

BLACKBIRD

ARISTA

Quantum

aws elemental

Source digital

VERITONE

"We expect the future of live media production to be performed in a fully virtualized, software based, highly agile data center running on commercially available off-the-shelf servers, Ethernet switches, and enterprise storage instead of today's bespoke broadcast hardware and coaxial cable interconnects. IP is the ultimate on-ramp into this vision."

Thomas Edwards, *VP Engineering & Development*, Fox Networks

F2 TECHNOLOGIES

F2 Technologies

- Canadian Partner of :
  - Arista Networks
  - AWS and AWS Elemental
  - Evolphin Asset Management
  - Forscene
  - Quantum Storage
  - Source Digital
- IP Based Media Systems Integrator
  - Project Management
  - 5,000 Square Foot Build Facility
  - Augmented Labour Service

TECHNOLOGIES

Tom Ohanian Backgrounder

- Co-Inventor Avid Media Composer, Film Composer, Multicamera Systems. Employee #8.
- Academy Award and two-time Emmy recipient for scientific and technical invention. Multiple patent holder.
- CSO at Signiant. Developed user interface and workflow modeling.
- Began broadcast engineer career at age 19 and author of multiple industry textbooks.
- Paul Friedman still makes me carry his travel bags at the airport. Though Barry and Jeff are nice.

How Artificial Intelligence and Machine Learning Will Change Content Creation Methodologies

Tom Ohanian
President
TAO Associates

MEDIA OVER IP SYMPOSIUM 2018

The Two Core Questions

- Can Artificial Intelligence and Machine Learning, as intelligent learning systems, produce content that heretofore required humans to produce?
- Is it conceivable that the creative decisions that humans make be codified such that intelligent systems can accomplish those creative tasks?

Applying AI & ML to Content Creation

- Real-time speech recognition providing real-time subtitling and CC in over 80 languages and with 95-99% accuracy.
- Automatically creating personalized viewer highlights on a large scale.
- Automatic creation of different versions of promos by changing voiceovers by retyping words.
- Automated editing of footage from single or multiple cameras to create a coherent narrative of an event.
- Automatic creation of frame accurate, lip-synced images from a content library, creating content that is completely fabricated from various source elements.

Content Deluge, Lower CPM, Crowdsourcing

- YouTube Daily: 1B mobile video views. 5B videos watched. 432,000 hours of content uploaded—18,000 days or 49 years of content.
- Facebook: Daily: 100M hours viewed by 500 million people.
- Challenges of metadata tagging and categorization of content.
- Historical television model, based on CPM, redefined to audiences of thousands and hundreds.
- Make more content available with fewer resources. $100M motion picture = 20% for Post & Delivery. Cost of producing and delivering content for IP channels must be lower.

AI, ML, Neural Networks

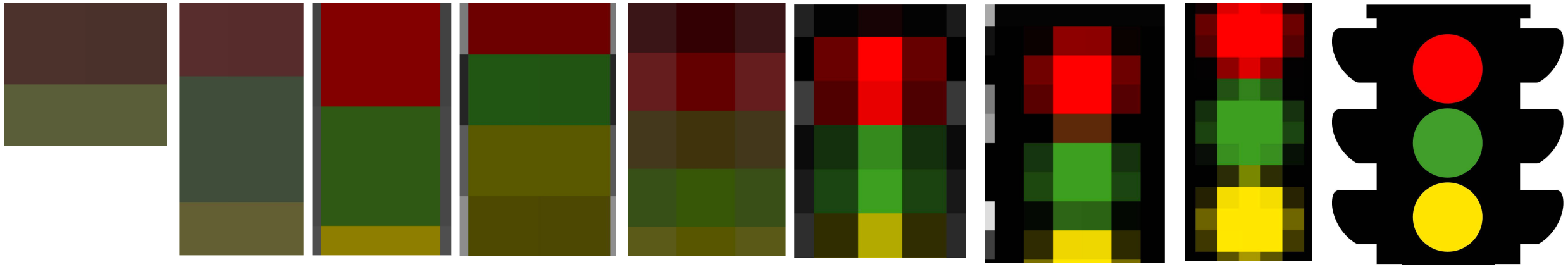
- AI can assist in image classification, facial recognition, etc. Rules by which systems logically undertake tasks: stock market trading.
- ML apps analyze large datasets and, based on the learning process, make determinations and predictions.
- Speech to text functionality + tonal analysis can provide the text of what is being said, indexed to the specific moment in time, and interpret items such as:
  - The sex of the speaker
  - Approximate age
  - The nature of the communication

AI, ML, Neural Networks

- Neural Networks combine AI and ML to process data.

- Assume a rectangle comprised of circles of colored pixels which are red, amber, and green and a surrounding area of pixels that are more uniform. Based on a library of similar shapes and pixel patterns, the conclusion may be that the rectangle represents:

AI, ML, Neural Networks



Three Phases of AI & ML to Content Creation

Phase 1: Decreasing Human Workload

Phase 2: Content Insight and Workflow Steering

Phase 3: Automatic Content Production

Building Blocks of the Three Phases

- *Speech to Text*
- *Language Recognition*
- *Cognitive Metadata Extraction*
- *Speech and Tonal Analysis*
- *Image Recognition*
- *Near Human Voice Quality Dubbing*
- *Real-time Data and Statistical Integration and Analysis*
- *General Automation Routines*

Phase 1

Decreasing Human Workload

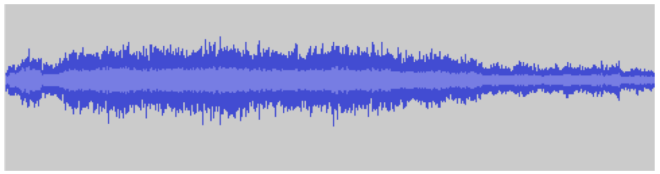
Phase 1

- **Phase 1 solutions decrease and eventually remove human workload.**
- Phonetics-to-text (PTT) conversion software began shifting the reliance from human operators to automated systems.
- Automatic speech recognition (ASR) systems can real-time process speech with increasing accuracy. RT speech recognition, subtitling and closed-caption creation in over 80 languages and with 95-99% accuracy is a reality. **That 5.1% error rate is on par to the error rate of human transcriptionists.**
- 240 motion picture theatrical versions requiring a human to watch a minimum of 200 unique compositions.

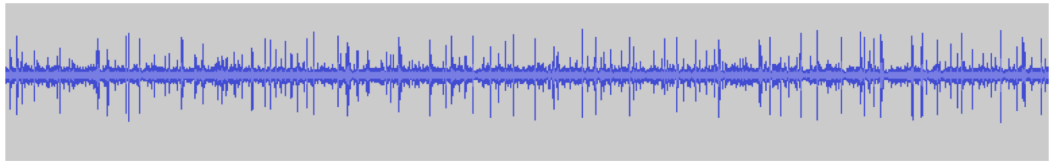
Phase 2

Content Insight and Workflow Steering

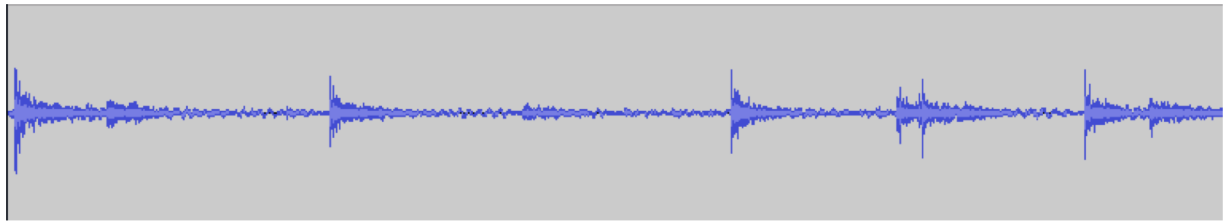
Phase 2: Tennis



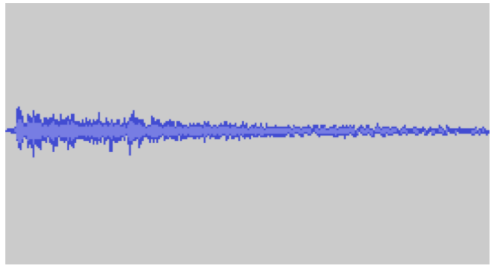
Viewer Catcalls



Viewer Cheers



Tennis Volley



Tennis Ace

Phase 2: Tennis

- Wimbledon: Automated methods for extracting and delivering added-value clips and content to viewers via solutions by IBM.

- *Real-time Data and Statistical Integration and Analysis:* Create rules by which potential clips identified based on court data and statistics. Breakpoints won, serves, scoring data, historical performances, etc., became part of a large dataset.

- *Image Recognition & Speech and Tonal Analysis:* The AI application analyzed crowd cheering and other crowd noises. Based on an historical video library of players, image recognition identified players and cataloged reactions based on a player's previous matches.

- *Highlight Clips*: Combining Phase 2 technologies and associating them with social media network data to use crowd-sourced information and correlate that real-time data (e.g. chat, tweet, etc.) as additional judging criteria for clip creation.

- End result: Automatic creation of highlight clips. By classifying players via facial pixel makeup, viewers could point a cell phone at a player to receive information unique to that player.

Phase 3

Automatic Content Creation

Phase 3

- Produce content according to data in the form of rulesets, large datasets of relevant examples, and creative conventions interpreted in the form of idiomatic expressions.

- *Speech Synthesis:* Artificial production of human speech by concatenating pieces of recorded speech. Text to Speech (TTS) uses a database of recorded speech from an individual to create new combinations of speech. Two main issues: Database must be very large and emphasis of the spoken phrase may be difficult to shape.

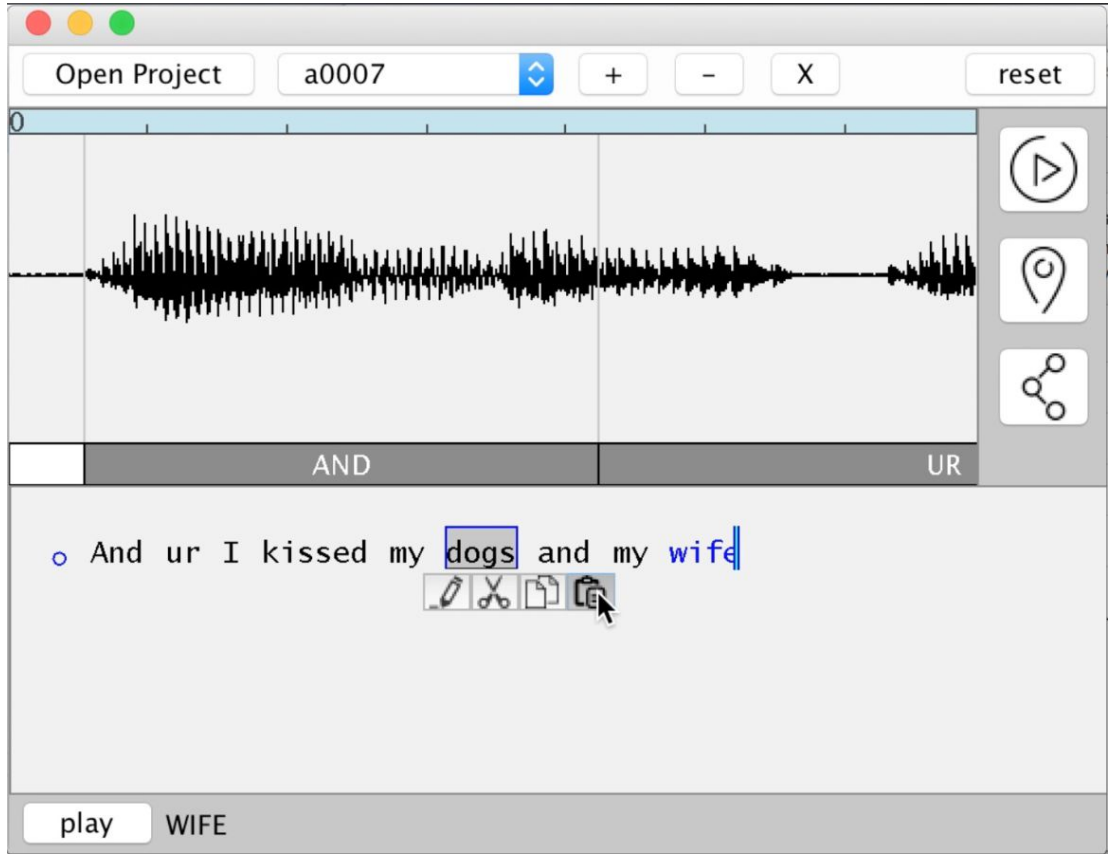
- Research by Heiga, Tokuda, and Black: Natural sounding synthetic speech advanced from knowledge-based (via a large database) process to a data-based method. This is Parametric TTS (P-TTS) where model parameters are adjusted to shape both content and characteristics of the speech. The output of the model is processed by algorithms in vocoders (voice encoders) and audio signals are generated.

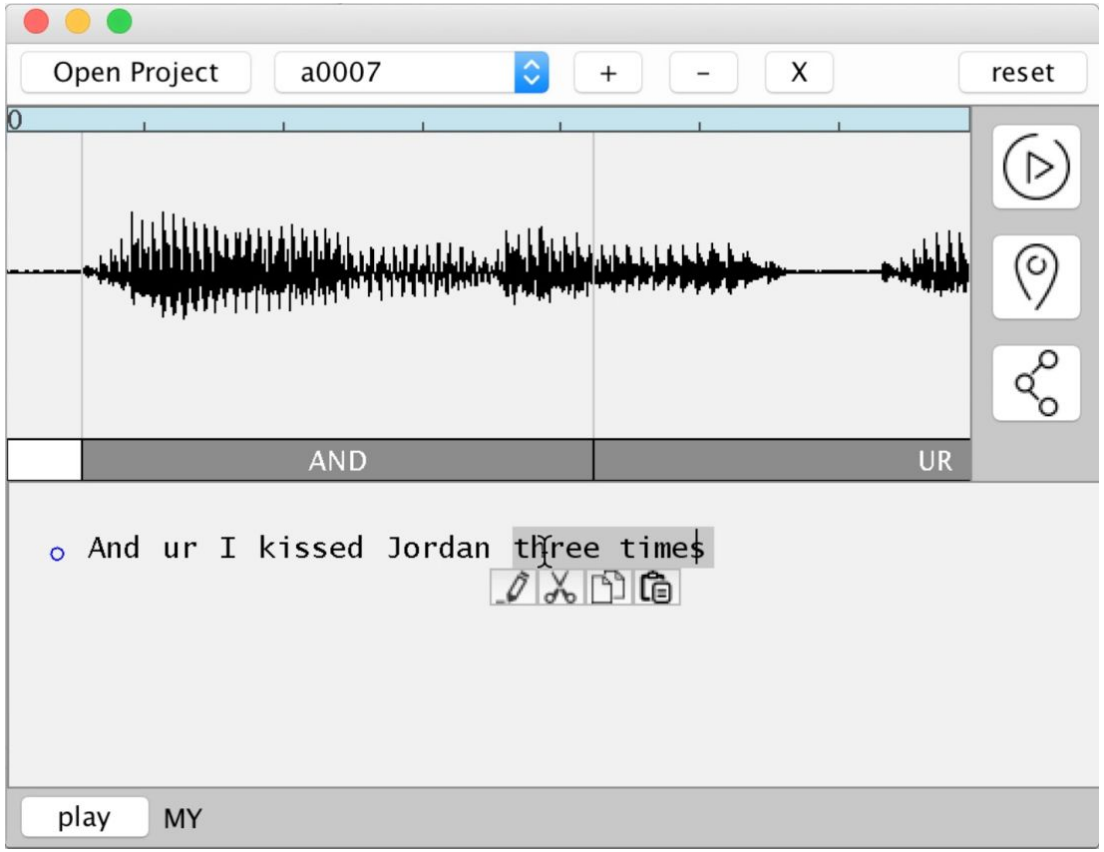
Phase 3

- *Changing Words in Voiceovers by Retyping Words*
- Generating synthetic speech provides countless possibilities.
- The number of promos for a network program can range from hundreds to thousands. Each episode, days of the week, durations (10, 20, 30, 60 seconds), voiceovers, highlighting different characters, network, affiliates, Pay-TV, OTT. Today, each of those versions is created manually.
- Instead of creating different versions by manually editing them and requiring new voiceovers be recorded, type the text of the new versions and have the audio changes automatically conformed.
- Resulting efficiency, flexibility, cost and time savings.

Phase 3: Project VoCo, Edit Speech in Text

Zeyu Jin, “Adobe VoCo for Adobe Creative Cloud”, Adobe Max





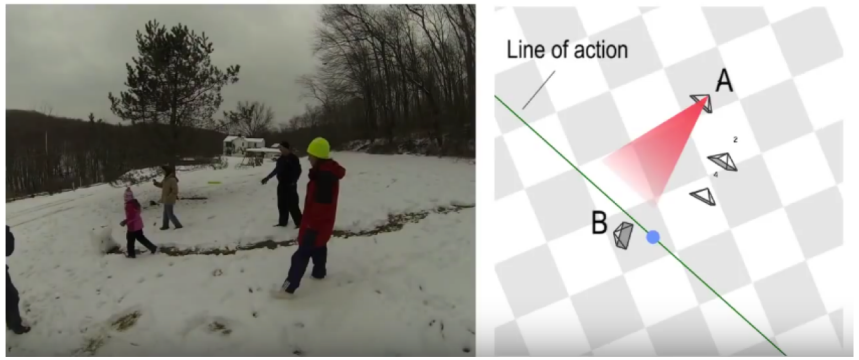
Phase 3: Automatic Content Creation

- *Multicamera originated footage that must be edited into a coherent final scene.*
- What is of primary interest? At what moment in time? When should a change be made?
- Number of cameras, angles, and focal lengths.
- Algorithms examining the principal point of interest, known rules of cinematography to create working models. that automate the construction of scenes based on this type of footage. Research by Arev, Hodgins, Park, Shamir and Sheik at Disney Research.

Phase 3: Auto Editing of Multicam Footage

- Arev, Hodgins, Park, Shamir, Sheik: Disney Research.

Four consumer / mobile cameras



Observing the 180-degree Line of Action Rule

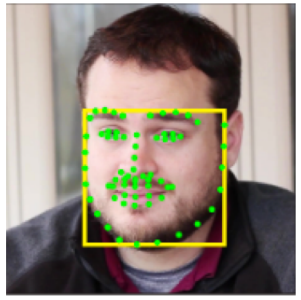
Avoiding jump cuts between nearby cameras

3D camera motion estimation to identify prime interest areas to decide when to cut to a different angle.

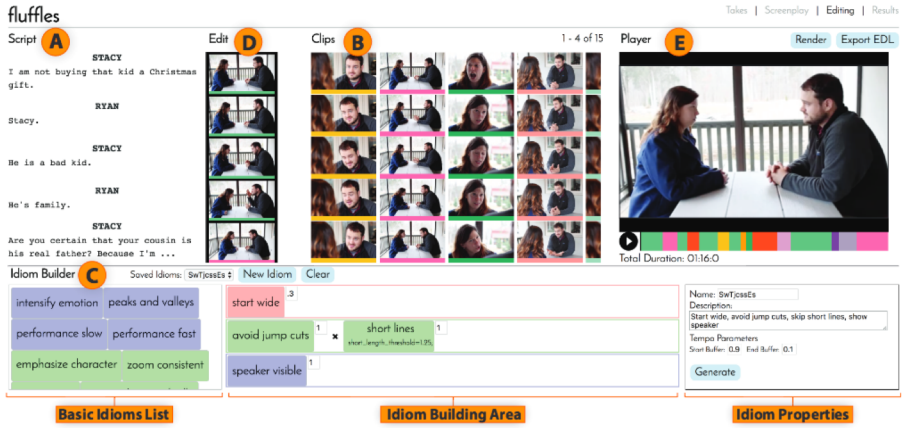
Phase 3: Auto Editing of Single Cam Footage

Agrawala, Davis, Leake, Truong: Stanford & Adobe Research.

Correlating the Input Script into Lines of Dialogue Spoken by Each Character



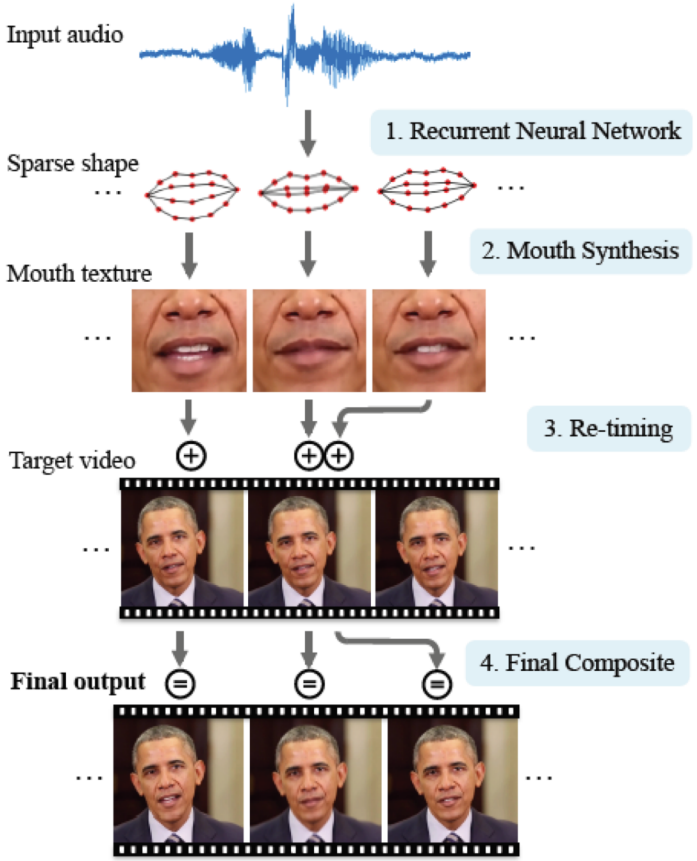
Facial Analysis & Tracking and Computing Speakers Visible by Changes in Mouth Area



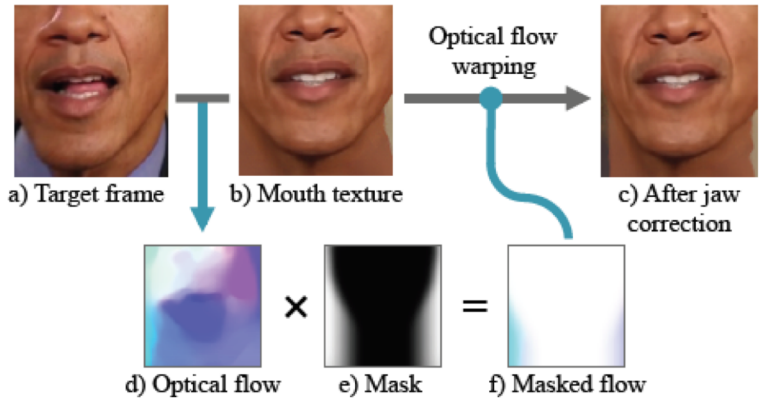
Choosing from the Idioms List and Placing into the Building Area results in Automatic Scene Construction.

Phase 3: Creating a Shot that Never Existed

Swajanakorn, Seitz, & Kemelmacher-shlizerman (University of Washington)



With a database of mouth shapes associated with time instances, mouth textures were synthesized and then composited with 3D matching to change what he appears to be saying. The result is that synthetic, photorealistic shots can be created.



Audio Converted to Time Varying Mouth Shapes and Fixing Jawline Discrepancies.

Conclusion

Artificial Intelligence, Machine Learning, and Neural Networks will make significant contributions to the content creation-to- consumption process.

- Make Friedman carry his own bags.