VES103: Enhancing Media With Machine Learning in 2019

Streaming Media East 2019
Who Am I?

- Jun Heider
  - CTOO at RealEyes Media

- We at RealEyes:
  - Build video players seen by millions (Super Bowl x 2, Olympics x 3)
  - Design and deploy worldwide eCDN and live streaming infrastructure (Oracle)
  - Love building solutions to manage live and VOD assets (EA Games, Viacom)
  - Partner with industry leaders (Wowza, Nginx, and Adobe)
What's the Problem?
Media libraries are growing at an accelerated rate. Cataloging of these libraries is crucial to maintain efficient searchability and ROI on this content being ingested.

Staffing is human, expensive, and there’s only a limited number of hours per person.

- Ingest metadata generation and analysis might not be the most cost effective use of staff time
- Staff may not be sufficient enough to handle the volume of new content
- Multiples humans are prone to either error or not being in sync (eg - Taxonomy)
How Do We Solve it?
How Do We Solve it?

- **AI - Artificial Intelligence**: The capability of a machine to imitate intelligent human behavior. (Turing Test)

- **ML - Machine Learning**: Using statistical techniques to give computer systems the ability to "learn" with data, without being explicitly programmed. ("If a typical person can do a mental task with less than one second of thought, we can automate it using AI ..." - Andrew Ng)

- **MLaaS**: Machine Learning as a Service, others build it and expose an API for you to consume...at the very least.
Neural Network?

http://vis-www.cs.umass.edu/bcnn/
Supervised Learning?

- $A \rightarrow B = \text{Something}$

- Labelled training set
Transfer Learning?

- Tuned model use less training to tune to new topics
- Tennis → Soccer → Hockey
Training/Testing?

- Training Set - Let’s train the machine with a large amount of input
- Test Set - Let’s make sure the machine learning model is working
Bias?

- Some examples:
  - Sample Selection Bias - Skin tone, gender, age, etc (million-scale and global dataset)
  - Interaction Bias - Wait, high heels aren’t shoes?
  - Survivorship Bias - Why does my Netflix only show movies like Armageddon?
What can we do?
Object Detection

- Clarifai - There are visual elements that I can recognize in the video
Face Detection

- AWS - I see humans and maybe even X/Y (eyes, nose, roll, yaw, pitch)

```json
{
  "Timestamp": 91858,
  "Celebrity": {
    "Urls": [],
    "Name": "Jacob McFlikier",
    "Id": "Ivw5vn3Z",
    "Confidence": 93,
    "Face": {
      "BoundingBox": {
        "Width": 0.12109375,"Height": 0.2152777761220932,"Left": 0.30000001192092896,"Top": 0.5100555582046509
      },
      "Landmarks": [
        {
          "Type": "eyeLeft","X": 0.34126874804496765,"Y": 0.5943679809570312
        },
        {
          "Type": "eyeRight","X": 0.38824111223220825,"Y": 0.591336727142334
        },
        {
          "Type": "nose","X": 0.3776569664478302,"Y": 0.6430012583732605
        },
        {
          "Type": "mouthLeft","X": 0.3496983051300049,"Y": 0.6868669986724854
        },
        {
          "Type": "mouthRight","X": 0.3836694359779358,"Y": 0.6818998456001282
        }
      ],
      "Pose": {
        "Roll": -3.6193172931671143,"Yaw": 22.45437240600586,"Pitch": -0.8823250532150269
      },
      "Quality": {
        "Brightness": 79.97325897216797,"Sharpness": 78.74752044677734
      },
      "Confidence": 99.99906921386719
    }
  }
}
```
Valossa - We can put a name to the face

Who is the best pilot I ever saw?.mp4

Make this report's URL public: 

Search within this video.

On-screen concepts: Dennis Quaid, beauty, clothing, costume, portrait, light, speech, nose, hair, man, glasses, face, head, black, blue, adult, eye, person, close-up, darkness

Speech keywords & sounds: "picture", speech

human.face Both genders 

Dennis Quaid 
male
ID: 1; Confidence: 1
AMS Video Indexer - Extract insights from text on screen

Weather Network COVERAGE of Hurricane Michael - October 10-11 2018
Logo/Brand Detection

- Veritone - Logos and brand mentions

Logo Recognition

- mga entertainment 00:01 - 00:02
- ferrari 00:04 - 00:05
- starbucks 00:04 - 00:05
- nfl 00:04 - 00:05
- starbucks 00:05 - 00:06
- nfl 00:05 - 00:06
- mcdonald indonesia 00:05 - 00:06
- burger king 00:05 - 00:06
- starbucks 00:06 - 00:07
- mcdonalds 00:06 - 00:07
- nfl 00:06 - 00:07
- 漢堡王 00:06 - 00:07
- gambar coca cola 00:07 - 00:08
- nfl 00:07 - 00:08
Compliance

- Valossa - Adult content, violence, or other questionable content?

Matt Foley_ Van Down By The River - SNL.mp4

Make this report's URL public: □

Search within this video.

On-screen concepts: room, wall, furniture, person threatened, hand-to-hand combat

Speech keywords & sounds: ”nail”, ”buddy”, ”amigo”, ”gear”, speech
Scene/Shot Detection

- Google Video Intelligence - Let’s figure out the smaller chunks in the video
AMS Video Indexer - From spoken audio to transcript

What happens is the protein that scaring the heme.
Translation

- AMS Video Indexer - From language A to language B
Audio Analysis

- **Valossa - Audio, Non-Speech?**

```
111: {
  "ext.refs": {"gkg": "/m/028ght"}
},
  "label": "applause",
  "oecs": [
    {
      "shs": 22,"ss": 237.0,"she": 22,"c_max": 0.516,"id": "392","se": 240.0
    },
    {
      "shs": 27,"ss": 341.0,"she": 28,"c_max": 0.51,"id": "393","se": 343.0
    }
  ],
  "t": "audio.context",
  "cid": "N1BEBdIy6gK5"
},
110: {-
},
113: {
  "cateq": {"tags": ["fauna"]
  },
  "ext.refs": {"gkg": "/m/068hy"}
 },
  "cid": "m1VEYHdchm0",
  "oecs": [
    {
      "shs": 23,"ss": 282.0,"she": 23,"c_max": 0.775,"id": "394","se": 285.0
    }
  ],
  "label": "pet",
  "t": "audio.context"
},
```

- **AMS Video Indexer - Stats?**

```
"statistics": {
  "correspondenceCount": 22,
  "speakerTalkToListenRatio": {"1": 0.039,"2": 0.451},
  "speakerLongestMonolog": {"1": 38,"2": 12},
  "speakerNumberOfFragments": {"1": 4,"2": 2},
  "speakerWordCount": {"1": 73,"2": 9,"3": 34}
}
```
Sentiment Analysis

Valossa

Who is the best pilot?

Make this report’s URL public:

Search within this video.

AMS Video Indexer

2 Sentiments

- 6.76% Positive
- 1.99% Negative

Watson Video Enrichment

Average face and speech sentiments

- Face sentiment:
  - 45% positive
  - 55% negative

- Face visibility:
  - 122 seconds
  - 37% of video length

- Speech sentiment:
  - 100% positive
  - 0% negative

Sentiment:

Sentiment score ranging from -5 (negative sentiment) to +5 (positive sentiment) that indicates whether the sentiment is positive, neutral, or negative.

Emotions:

Emotion scores ranging from 0 (no emotion) to 1.0 (maximum emotion). The scores are -0.51 (Sadness), -0.22 (Disgust), -0.14 (Fear), 0.12 (Fear), and 0.08 (Disgust).
Trending Topics

- Vilynx - What’s popular?
Other Capabilities

- Object Tracking
- Context and Categorization
- Landmark Detection
- Color Analysis
- Speaker Separation
- Speech Analytics
- Audio Fingerprints
- Sports Event Logging
- Highlight Clip Generation
- And more each day...
2019 MLaaS Landscape?
AWS

- https://aws.amazon.com/rekognition/video-features/
- IMO most robust facial detection out there in the more prominent MLaaS services
- In our testing speech to text seems very accurate (AWS TRANSCRIBE)
Azure

- One of the most robust easy to get started solutions from both a UI perspective and an API perspective
Clarifai

- [https://clarifai.com/](https://clarifai.com/)
- Number of different models
- Seems more geared toward retail, marketing, and creative types
Google

- Video Intelligence: [https://cloud.google.com/video-intelligence/](https://cloud.google.com/video-intelligence/)
- AutoML Video Intelligence (Beta)
- Lots of languages for translations (100+)
- Shot Detection
Video Labels
Detect and label entities, such as dogs, flowers, and people, throughout the entire video.

- nature: 91%
- animal: 83%
- tree: 70%
- plant: 70%
- vegetation: 70%
- forest: 68%
- geographical feature: 50%
- flora: 50%
- plant: 50%
GrayMeta

- [https://www.graymeta.com/](https://www.graymeta.com/)
- Provides a general description with the tags
- Multiple file or folder upload and cloud storage ingest
- Adobe Premiere Pro Plugin (Make sure Photoshop is installed before trying to add this extension)
Valossa

- [https://valossa.com](https://valossa.com)
- Bounding box for faces and logos
- Robust compliance detection
- In depth audio analysis
- Training Gallery for Faces
- Fast processing
Veritone

- https://www.veritone.com/
- Operating System for AI
- Select different engines for your indexing workflow
- Multiple file or folder upload and cloud storage ingest
- Processing status screen
Vilynx

- http://vilynx.com
- Trending topics in news and social media
- Clip generation
- Multiple file or folder upload and MRSS Ingestion
Vilynx

NHL Playoff Game. Edit

Video
04:58
3 days ago

Video information
Clips editor
New auto summary
Explore
Add subtitles
Download subtitles

Public Video
Manage external ID
Delete Video

TOP 3 TRENDING TAGS
1. San Jose Sharks
2. New York City
3. Target Corporation
Others (Not Necessarily MLaaS)

- Avigilon (Surveillance)
- Cogniac
- DeepAI
- Deep Video Analytics
- Figure Eight
- Hive
- iCetana (Surveillance)
- IBM Watson
- IntelliVision (Surveillance)
- Intel OpenVINO (Edge)
- MulticoreWare (Edge)
- Neurala
- ParallelDots
- Prime Image
- Raypack
- Speechmatics (Transcription)
- Stainless.ai
- Tedial
- Trint (Transcription)
- Zorroa
Making MLaaS Work for You
Ask Yourself

- What’s your content profile?
- What’s your functional use case?
- How accurate do you need it?
- Do you have developers?
- Did you play around and assess?
Ask Yourself

- What’s your content profile?
  - Movies/TV
  - News Footage
  - Sports
  - Nature Clips
  - User Generated Content
  - Home Videos

- Each service has been training on different data
UI Available - Play with It
API Available - Play with It

Video Indexer Developer Portal

Operations

Get Video Index

Query parameters

- location: trial
- accountid: Value
- videoId: Value
- language: Value

Headers

- Add parameter
- Add header
### Multiple Vendors - Play with Them

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSFT VI v2 Labels</td>
<td>Valossa</td>
<td>Google VI</td>
<td>AWS</td>
<td>Avg Confidence (if misdetected)</td>
<td>Note (if misdetected or not detected)</td>
</tr>
<tr>
<td>Person (Yes)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>Person (Yes)</td>
<td>AWS-98%</td>
<td>(AWS) Just wanted to note that a human comes into the scene about 10 seconds after the AI says it should</td>
</tr>
<tr>
<td>Indoor (Yes)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>No Tag</td>
<td>No Tag</td>
<td>(MSFT) It was very accurate except for one time, it labels part of the credits as outdoor.</td>
</tr>
<tr>
<td>Outdoor (26/26)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>No Tag</td>
<td>MSFT-0.8704</td>
<td></td>
</tr>
<tr>
<td>Man (Yes)</td>
<td>Male (Yes)</td>
<td>No Tag</td>
<td>No Tag</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text (Yes)</td>
<td>No Tag</td>
<td>Text (Yes)</td>
<td>Text (Yes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building (Yes)</td>
<td>No Tag</td>
<td>Building (Yes)</td>
<td>Building (Yes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tree (6/7)</td>
<td>No Tag</td>
<td>Tree (plant) (Yes)</td>
<td>No Tag</td>
<td>MSFT-0.9857</td>
<td>(MSFT) Counted vines flowing off a building as a tree</td>
</tr>
<tr>
<td>Wall (Yes)</td>
<td>Wall (5/32)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>Valossa-0.947</td>
<td>(Valossa) It caught 32 instances that were a wall, about only 5 of them actually had what I would consider a wall</td>
</tr>
<tr>
<td>Sky (Yes)</td>
<td>Sky (Yes)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>No Tag</td>
<td></td>
</tr>
<tr>
<td>Ground (Yes)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>No Tag</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glasses (Yes)</td>
<td>No Tag</td>
<td>Glasses (Yes)</td>
<td>No Tag</td>
<td>(MSFT) Didn’t identify glasses in other frames when other content was by the man’s face</td>
<td></td>
</tr>
<tr>
<td>Standing (Yes)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>No Tag</td>
<td>MSFT-0.8728</td>
<td>(MSFT) No type of sport was in the scene nor the entire video.</td>
</tr>
<tr>
<td>Sport (No)</td>
<td>Sports (No)</td>
<td>Individual Sports (sport(s)) (No)</td>
<td>No Tag</td>
<td>Valossa-0.818</td>
<td>(Valossa) Same exact scene as MSFT, no type of sport is in the video</td>
</tr>
<tr>
<td>Weapon (Yes)</td>
<td>Weapon (Yes)</td>
<td>Weapon (Yes)</td>
<td>Weapon (Yes)</td>
<td>Google-42%</td>
<td>(Google) Same exact scene as the two above. The scene is a girl chasing a guy</td>
</tr>
<tr>
<td>Young (Yes)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>No Tag</td>
<td>No Tag</td>
<td></td>
</tr>
<tr>
<td>Factory (Yes)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>Factory (Yes)</td>
<td>No Tag</td>
<td></td>
</tr>
<tr>
<td>Sidewalk (Yes)</td>
<td>No Tag</td>
<td>No Tag</td>
<td>No Tag</td>
<td>MSFT-0.8028</td>
<td>(MSFT) Misdetect factory as city</td>
</tr>
<tr>
<td>City (No)</td>
<td>No Tag</td>
<td>City (geographical feature) (No)</td>
<td>No Tag</td>
<td>Google-97%</td>
<td>(Google) Misdetect factory as city</td>
</tr>
</tbody>
</table>
Multiple Vendors - Play with Them
Multiple Vendors - Review Them
Multiple Vendors - Normalize

AWS
{
  "aws-labels": [
    {
      "Timestamp": 11166,
      "Label": {
        "Name": "Train",
        "Confidence": 74.20560455322266
      }
    }
  ],
}

GOOGLE
{
  "google-entity": {"entityId": "/m/0204fg","description": "skyline","languageCode": "en"},
  "categoryEntities": [{"entityId": "/m/0in32","description": "city","languageCode": "en"},
    "segments": [{"segment": {"startTimeOffset": {"seconds": 18,"nanos": 416666000000},
      "endTimeOffset": {"seconds": 24,"nanos": 958333000000},
      "confidence": 0.989143987913288}
    }
  ],
}

AZURE
{
  "azure-id": 17,
  "name": "weapon",
  "appearances": [{"startTime": "00:01:24.0830000","endTime": "00:01:25.4160000"}
}

UNIFIED
{
  "id": 17,
  "object-name": "weapon",
  "entity-id": "/m/0204fg",
  "confidence": 74.20560455322266,
  "occurrences": [
    {
      "startTime": "00:01:24.0830000",
      "endTime": "00:01:25.4160000"
    },
    {
      "startTime": "00:03:30.0920000",
      "endTime": "00:03:45.4160000"
    }
  ]
}
Bring MLaaS Insights Into Your App

AEM Assets + AEM Video Indexer DEMO
Let’s train the model

Import videos

AutoML Video Intelligence uses your videos to train a custom machine learning model. Learn more about preparing your data.

- Upload labels in your CSV, or upload un-labeled videos, and use our labeling tool.
- At least 100 video segments per label is recommended.
- Processed videos will be stored on Cloud Storage. Standard pricing applies.

Select a CSV file on Cloud Storage

The CSV file should contain paths to your train, test, and/or unassigned CSV files. Videos must be .MOV, .MPEG4, .MP4, or .AVI. Learn more.

Example CSV:

TRAIN, gs://domestic-animals-vcm/horses/videos/train.csv
TEST, gs://domestic-animals-vcm/horses/videos/test.csv

BROWSE
Closing Thoughts?
Thank you!

- MLaaS for video use cases is super!
- Play around and see which one(s) are best for your use case
- Train when you need more accuracy

- www: https://realeyes.com
- email: jun@realeyes.com
- twitter: @coderjun
- linkedin: http://www.linkedin.com/in/junheider

I would love feedback: http://bit.ly/re-sme2019