

Deep Video Analytics

Akshay Bhat, Cornell Tech, Cornell University.



A good overview of computer vision research by Tomasz Malisiewicz

<http://www.computervisionblog.com/2015/01/from-feature-descriptors-to-deep.html>

A quick summary

Sift, Graph Cuts



HOG, DPM



Deep Learning



??????

Internet



Caltech 101, Matlab OpenCV



VOC, Imagenet, Caffe, Theano



??????

Numerous high quality libraries & datasets

- OpenCV
- ROS
- Caffe (model zoo!)
- Theano
- Torch
- Tensor Flow
- CNTK
- MXNET
- Pytorch
- Caltech 101
- Imagenet
- COCO
- Too many to keep track!
 - Open Images
 - [Soundnet](#)
 - [Mapnet](#)
 - [CMU Video patch dataset](#)

A deluge of datasets!

- VideoNet
- Yahoo Flickr Creative Commons 100M
- ViCom
- Visual Genome
- YouTube-BoundingBoxes
- Youtube 8M
- imSitu by AllenAI
- Charades by Allen AI
- Udacity car dataset
- KITTI
- Caltech, INRIA, ETH Pedestrians
- Stanford Drone Dataset

State of the art pre trained models

- Imagenet classification
 - Inception
 - Resnet
 - VGG
- Detection models
 - R-CNN
 - YOLO
 - SSD
- Face detection / recognitions
 - Face-MTCNN
 - Facenet
- Semantic Segmentation models
 - Multipathnet
 - FCN
- Audio embedding models
 - Soundnet

Question

What is natural progression after libraries, large datasets and pre trained models?

Answer

A platform which seamlessly combines
Data + Models + User Interface.

What is hidden in plain sight?

A Relational Model of Data for Large Shared Data Banks

By Edgar F. Codd

Can we develop an equivalent of relational
model / databases for visual data?

Visual Data

=

{ Images, Videos, Annotations, Features }

Relational data : Postgres, MYSQL, SQLite

::

Text, HTML : Lucene/Solr, Elasticsearch

::

Videos & Images : _____

Relational data : Postgres, MYSQL, SQLite

::

Text, HTML : Lucene/Solr, Elasticsearch

::

Videos & Images : ***Deep Video Analytics***

Relational data : SQL

::

Text, HTML : inverted word index, Page Rank

::

Videos & Images : ***Approximate Nearest Neighbor***

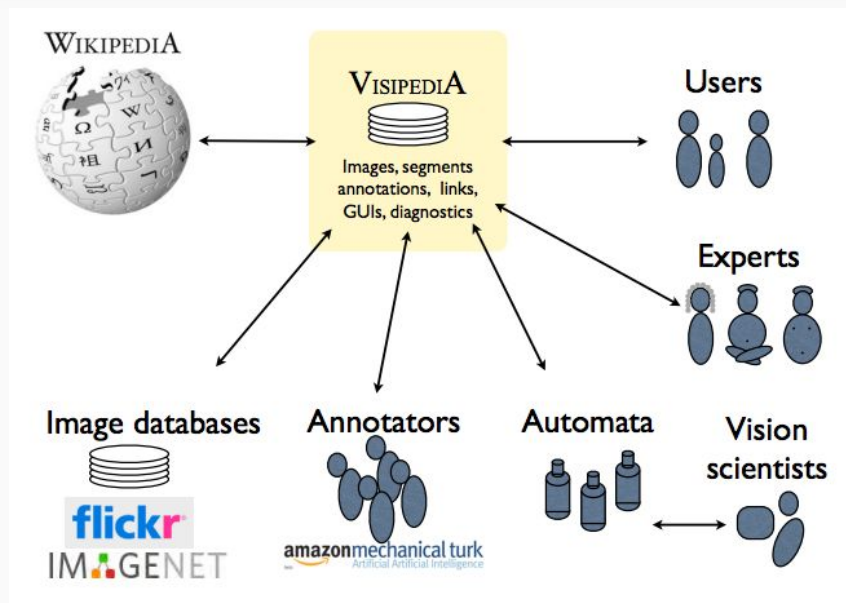
Previous attempts: CloudCV

- Large Scale Distributed Computer Vision as a Cloud Service
- Intended for researchers, and non-researchers
- Support for OpenCV, Graphlab, Cafe
- Image Classification, VQA, stitching, etc
- **Does not retains state!**

Previous attempts: NVidia DIGITS

- "DIGITS (the Deep Learning GPU Training System) is a webapp for training deep learning models. "
- Load/create datasets, train models, deploy models.
- Aimed at researchers
- Written in Python/Flask with Torch & Caffe supported
- **Retains uploaded images.**

Previous attempts: Visipedia



Taken from Vision of a Visipedia, Perona et. al.

Previous attempts: Visipedia

- Collaborative creation of visual data
- Pre-defined set of concepts E.g. Birds
- Different type of participants
 - Experts, Annotators, Citizen Scientists, Users, Computer scientists
- **Retains state!**

Previous attempts: VMX.ai

- Unsuccessful Kickstarter project Circa Jan 2014
 - by Tomasz Malisiewicz Pre Tensor Flow, Pre Deep Learning
- Allow developers to create real time detectors
- Support for training model, works via browser canvas
- <https://www.kickstarter.com/projects/visionai/vmx-project-computer-vision-for-everyone/description>

Model of a Visual Intelligence Platform

Provides images & videos,
along with metadata,
annotations

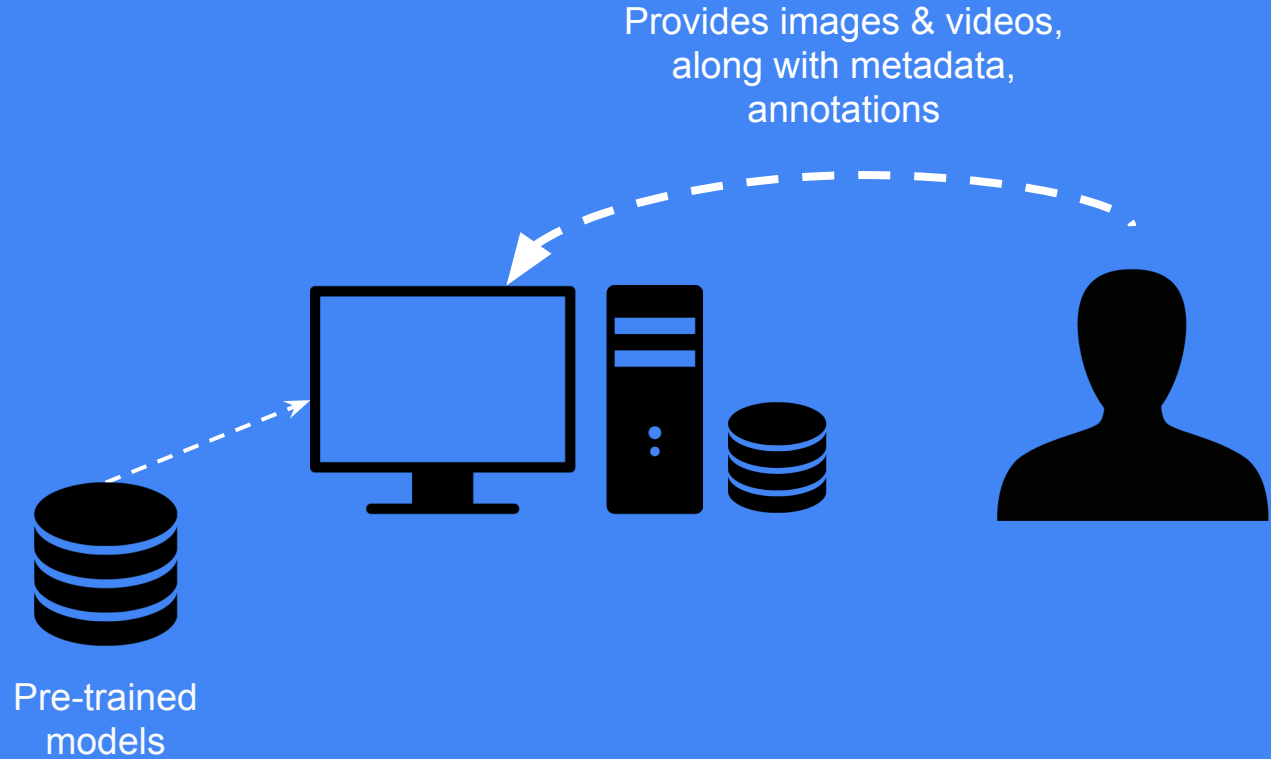


Model of a Visual Intelligence Platform

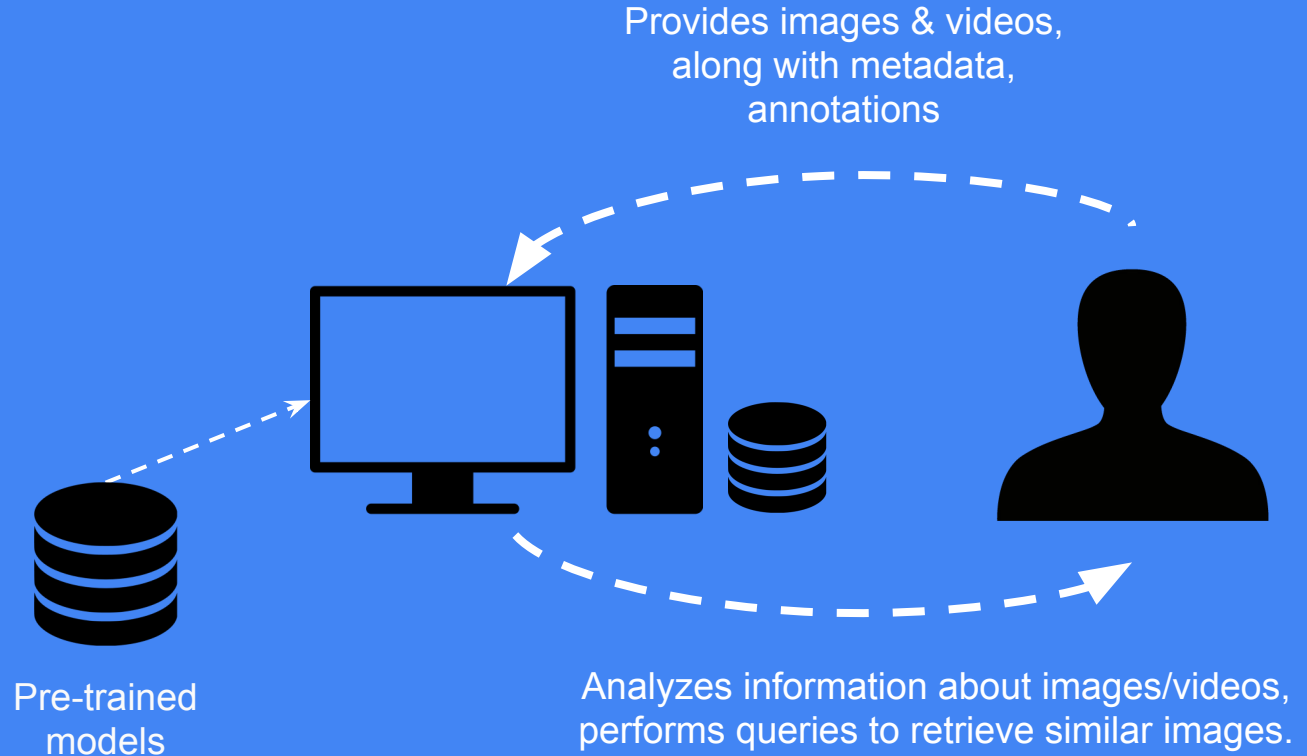
Provides images & videos,
along with metadata,
annotations



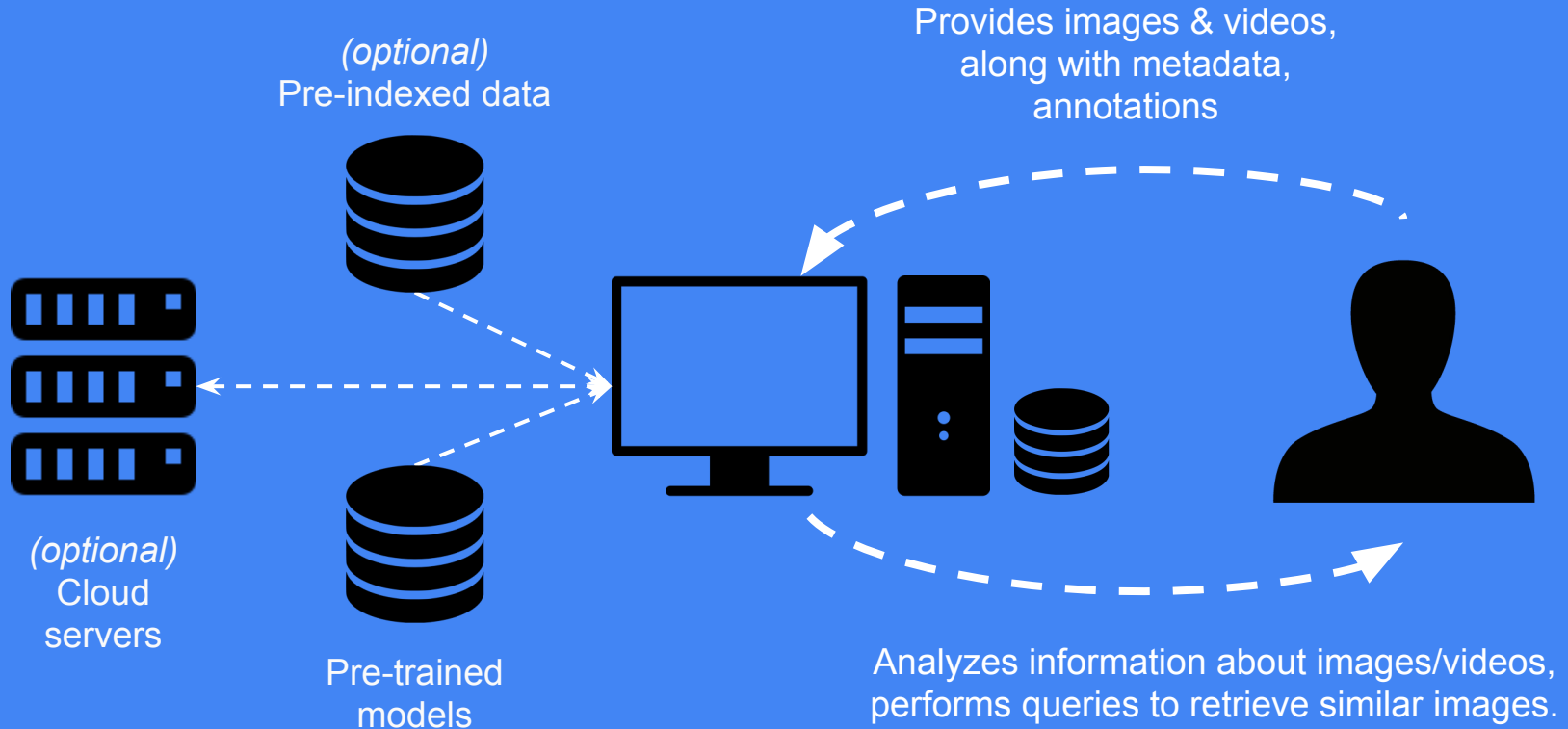
Model of a Visual Intelligence Platform



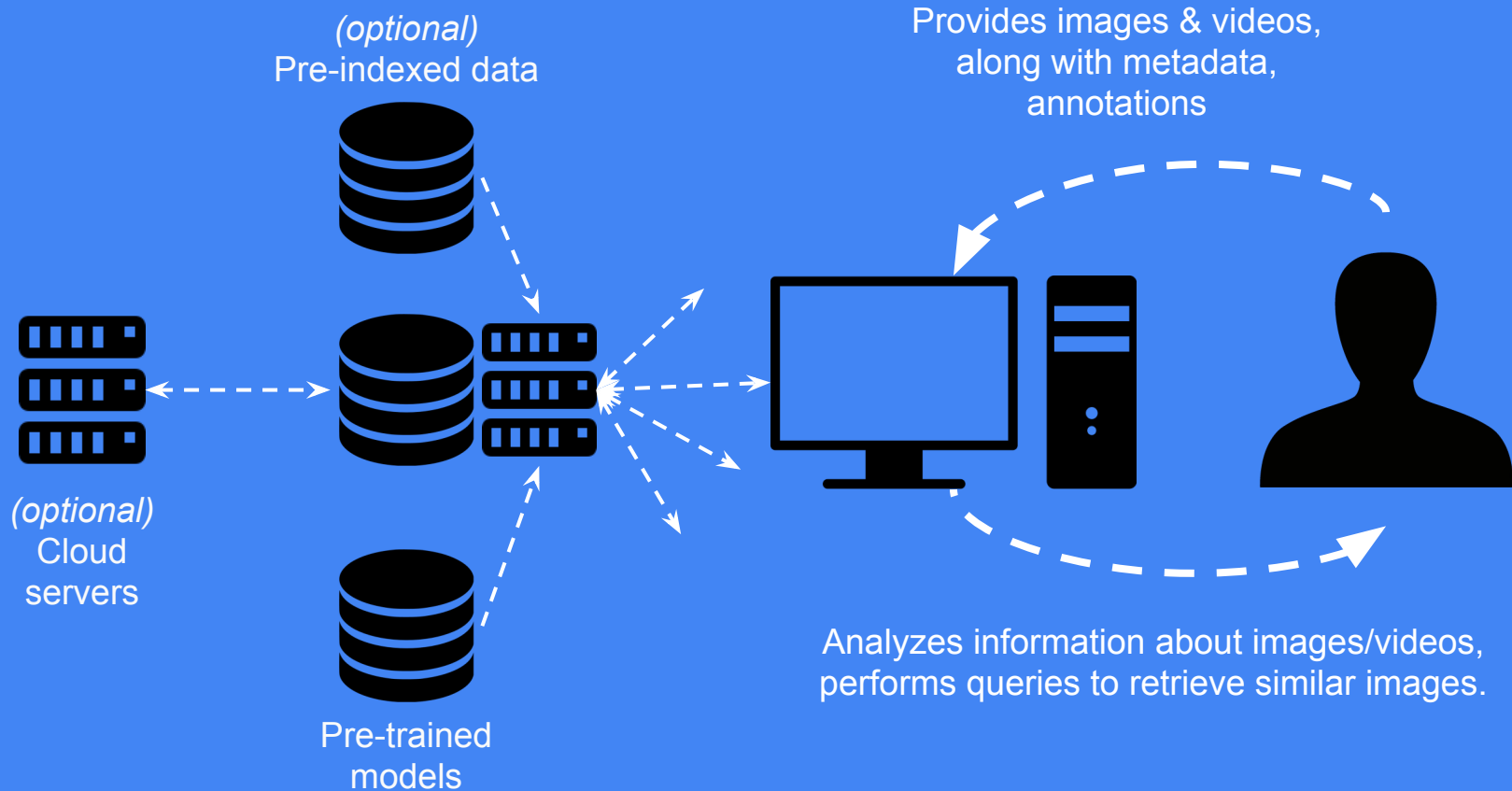
Model of a Visual Intelligence Platform



Model of a Visual Intelligence Platform



Model of a Visual Intelligence Platform



Question

Why not just modify lucene to index images?

Answer

Visual Search is significantly different compared to full text search. It requires a new user interface and ability to handle detections, segmentations, videos, etc.

Deep Video Analytics

Visual Search as a “Primary User Interface”

- Intended for **non-researchers**
- Make it easy for users to provide data (uploads, youtube-dl, etc.)
- Batteries-included approach with an indexing and detection pipeline
 - Tensor Flow Inception v3
 - Single Shot Detector trained on VOC & YOLO 9000
 - Face detection / alignment / recognition
 - More algorithms such Text detection, Audio features planned.
- Pre-indexed datasets from different domains can be quickly loaded
- Can be easily customized by developers & researchers.

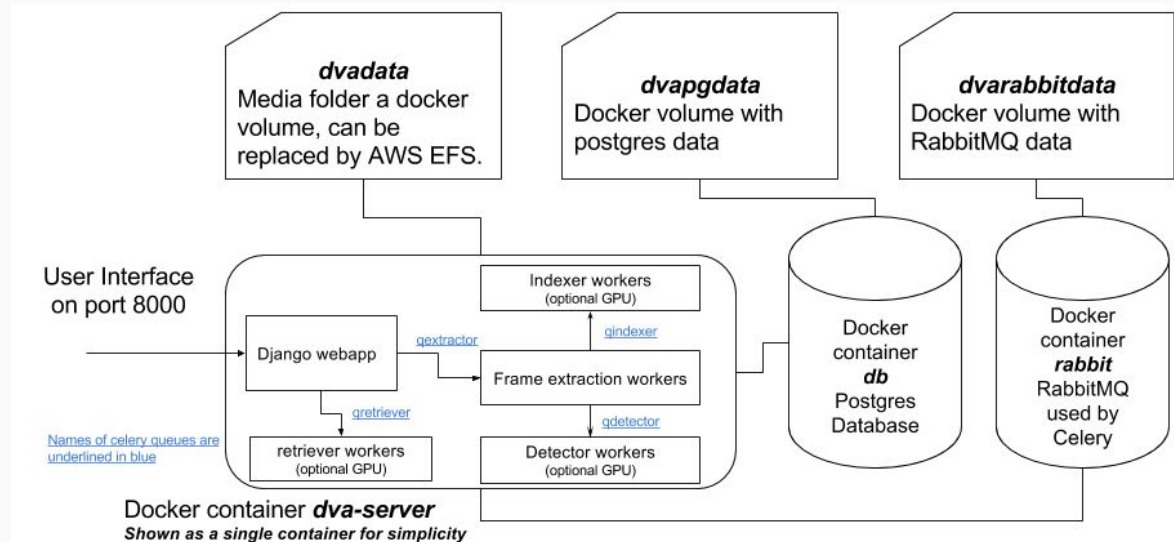
Technical requirements

- Must be useful without having to write code or config
- Must work on machines with and without GPUs
- Must allow uploads and reindexing operations
- Easy to adapt by technical users
- Easy to dynamically scale out using cloud computing

Emulating datacenter on a machine

Docker, Docker-compose, Nvidia-docker

Docker enables same codebase across all configurations {a laptop, multi-GPU machine, datacenter} .



Code organization:

dvaapp & dvalib

dvaapp: a django app/project

- Handles UI and data processing
- Data model
 - Video, Frame, Detection
 - Query, QueryResult
 - Event, etc.
- A set of celery tasks
 - Extract frames / process video
 - Perform indexing
 - Perform detection
- Uses dvalib to carry out tasks

dvalib: library for handling algorithms

- A database & celery agnostic library
- Interface with Tensor Flow & Pytorch for
 - extraction
 - detection
 - indexing
- Usable without having a running django instance, but designed to interface with it. E.g. assumptions regarding layout of directories containing videos, frames etc.

User Interface:

Visual Search as primary interface

Deep Video Analytics

Exact Search Completed

Account

By Akhay Bhat

Clean editor

Add Image

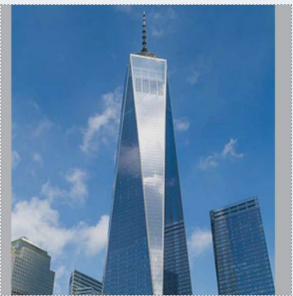
Exclude

Approximate Search

Exact Search

Clear stroke

Reset Zoom



Upload a video or multiple images in a single zip file (example of zip file with jpg images).

File:

Choose File

No file chosen

Upload

Submit youtube video or playlist using youtube-dl.

Submit (coming soon)


Data	Count	View
Videos / Datasets	6	view
Frames	4019	
Detections	11157	
Queries	7	view

Results:

[View results from past 7 queries](#)


In video at 70100

Found by tfinception




In video at 69900

Found by tfinception




In video at 40300

Found by tfinception



In video at 29500

Found by tfinception

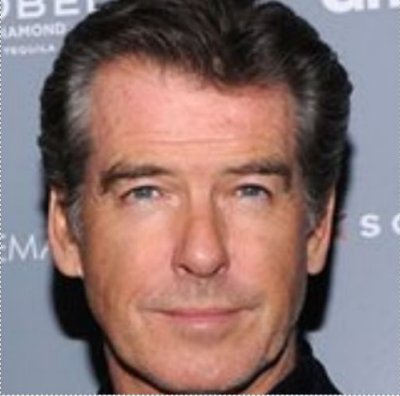


User Interface:

Search across frames + detections (faces, etc.)

Clear masks

Reset Zoom



provide a name

File:

Choose File

No file chosen

Upload

Submit youtube video url. We use youtube-dl.

provide a name

url of youtube video

submit

Videos / Datasets3view

Frames286

Detections514


Queries0view

Results:

View results from past 0 queries


Frame rank 1

In video at 4300 found by tfincception




Frame rank 2

In video at 4164 found by tfincception




Frame rank 3

In video at 4218 found by tfincception




Frame rank 4

In video at 1500 found by tfincception




Detection rank 1

In video at found by facenet




Detection rank 2

In video at found by facenet




Detection rank 3

In video at found by facenet




Detection rank 4

In video at found by facenet



User Interface:

Browse previous queries

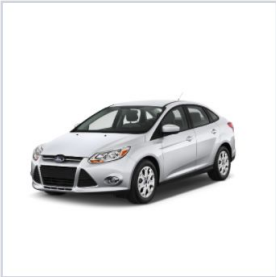
 **Deep Video Analytics**

AccountBy Akshay Bhat

List of past queries

click on query image to view results

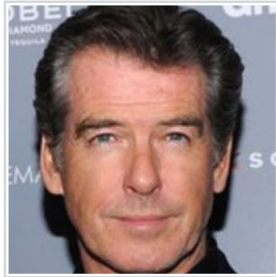
Created at March 9, 2017, 1:14 a.m.
results available



View results

Query again using image

Created at March 9, 2017, 1:22 a.m.
results available



View results

Query again using image

User Interface:

Browse videos, frames/images and detected objects

Deep Video Analytics

Account By Akshay Bhat

List of frames or images in tomorrow never dies

List of frames/images & detection

Show 10 5 entries Search

frame or detected object name	confidence	size in pixels	frame index	view
entire frame	100.0 %	921600	624	view
detected : SSD_person	96.8 %	25728	624	view
entire frame	100.0 %	921600	696	view
detected : SSD_person	99.7 %	25320	696	view
detected : SSD_person	96.2 %	10816	696	view
entire frame	100.0 %	921600	700	view
detected : SSD_humonitor	86.5 %	17876	700	view
entire frame	100.0 %	921600	726	view
detected : SSD_humonitor	83.4 %	16880	726	view
detected : SSD_person	53.3 %	135786	726	view

Showing 41 to 50 of 407 entries

Previous 1 4 5 6 ... 41 Next

Video

<https://www.youtube.com/watch?v=gTzCaw980c>



0:27 / 2:20

Metadata from ffmpeg

ffmpeg version 4.2.1 Copyright (c) 2000-2019 Fabrice Bellard, et al. built on Linux 4.15.0-47-generic, 64bit

Deep Video Analytics

Account By Akshay Bhat

List of Detections in WorldIsNotEnough at 3120

object	confidence	x	y	w	h
SSD_boat	99.4075655937	174	97	314	142

Q Query using this frame



Detected objects:

SSD_boat



User Interface:

Get status of running and finished tasks & resubmit tasks.

≡

Deep Video Analytics

AccountBy Akshay Bhat

System status and logs

Data	Count	View
Videos / Datasets	5	view
Frames	286	
Detections	514	
Queries	2	view

Events

Show 10 entries

Search:

Video	Operation	Started	Completed	timestamp	time taken in seconds	retry task
exampledataset	extract_frames_by_id	True	True	March 8, 2017, 11:31 a.m.	3.2	Retry
exampledataset	index_by_id	True	True	March 8, 2017, 11:33 a.m.	3.0	Retry
exampledataset	perform_detection_by_id	True	True	March 8, 2017, 11:47 a.m.	188.1	Retry
query_1	query_by_id	True	True	March 9, 2017, 1:14 a.m.	5.8	Retry
query_1	query_face_by_id	True	True	March 9, 2017, 1:14 a.m.	197.1	Retry
query_2	query_face_by_id	True	True	March 9, 2017, 1:22 a.m.	1.1	Retry
query_2	query_by_id	True	True	March 9, 2017, 1:22 a.m.	2.0	Retry
tomorrow never dies	extract_frames_by_id	True	True	March 8, 2017, 11:31 a.m.	81.2	Retry
tomorrow never dies	index_by_id	True	True	March 8, 2017, 11:33 a.m.	72.5	Retry
tomorrow never dies	perform_detection_by_id	True	True	March 8, 2017, 11:50 a.m.	663.7	Retry

Showing 1 to 10 of 13 entries

Previous12Next

Indexer log

Detector log

Open questions:

A work in progress

- How to rank results using auxiliary information?
- How to balance fast/static vs slow/dynamic indexes?
- How to incorporate external (pre & un) indexed data?
- How to incorporate text data extracted from images?
- Can the system continuously learn new categories?
- Can we create a real time plug-in?
- Can we create an android / iOS frontend app?

Thanks!

Contact me:

akshayubhat@gmail.com

www.akshaybhat.com

