Automated Feature Labelling: Head Detection

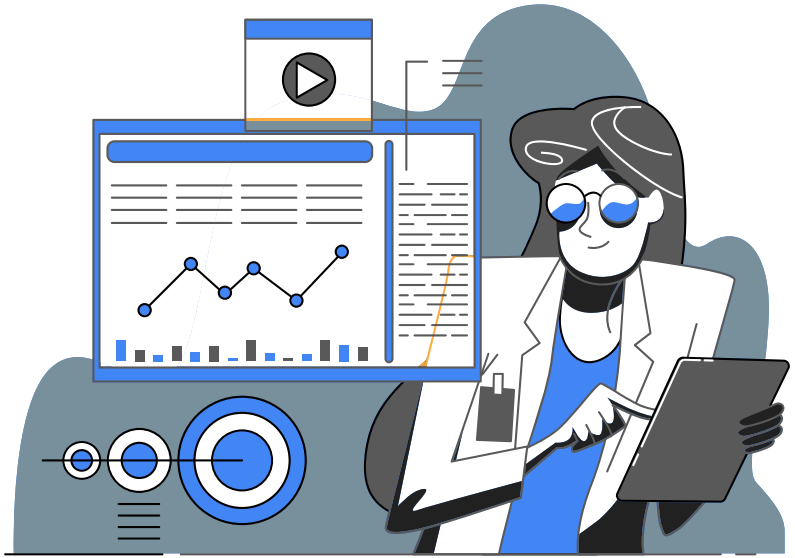
By: Anirudh

CS-153 Computer Vision

Head Detection!

<- Is that actually a head!?..

Table of Contents



Motivation

1. ~35% of pixels in movies and YouTube videos as well as ~25% of pixels in photographs belong to people (1)
2. Head Detection is hard -
   a. State of the art systems reach about 65% precision (as of 2015) (1)
   b. Differences in pose, minimal features, especially from the back view, partial blocking and more
3. Harder than face detection!
4. A chance to work with Deep Learning Models

Related Work

Context-aware cnns for person head detection(Tuan-Hung Vu, Anton Osokin and Ivan Laptev) Tuan-Hung Vu, Anton Osokin and Ivan Laptev, 2015. Context-aware cnns for person head detection

1. Really Hard to implement
2. Uses three different models in conjunction with each other
    a. Local Model - R-CNN based, trained on ImageNet
    b. Global model - Spatial 3D heatmap of objects locations
    c. Pairwise model - Trained to reason about relationships

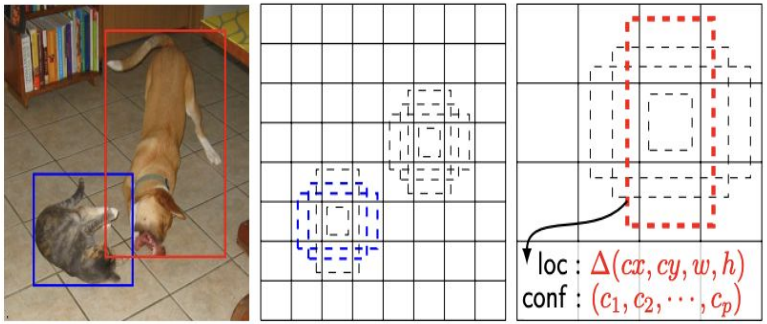
Single Shot Multi-Box detector (Wei Lie Et. Al)

1. Single Neural Network
2. Outperforms state-of-the-art faster R-CNN model
3. Very fast (59 fps on NVIDIA Titan X), 74% accuracy
4. Possibly useable in embedded systems due to its speed

Data

1. Gaze Data from Prof. Breeden
    a. 15 film clips, 1-4 minutes long, from a variety of different genres
    b. Frame by frame annotations of presence of face
    c. Other unused information, such as
        i. Gaze features
        ii. Visual styles
        iii. Temporal pacing

Single Shot Multibox Detector. How does it work



(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

1. Training only needs input images and ground truth boxes
2. In evaluation, prediction stage, the model produces predictions of different scales from feature maps of different scales, and explicitly separate predictions by aspect ratio.

Model Used

1. **Trained SSD Based model**
2. Uses concepts and code from Wei Liu et. al implementation of SSD object detectors for head detection in particular
3. Trained on HollywoodHeads dataset (pre trained yay!)
4. Model Dependencies
    a. TensorFlow backbone
    b. Keras (SSD implementation)
    c. CUDA only

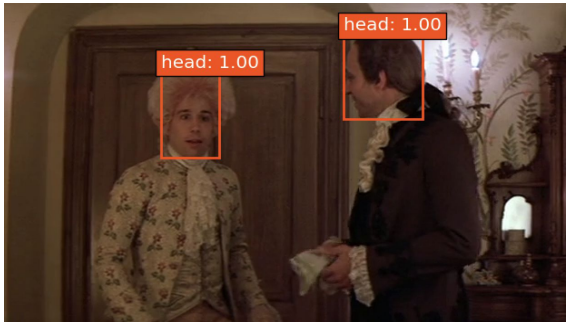
Training using just frames, bounding box truth labels.

Methods

1. Model Features
    a. Only using pre-trained weights
    b. Lot of flexibility with hyperparameters
        i. Aspect-ratios per layer
        ii. Confidence threshold
        iii. Etc
2. Converted input frame from base dimensions to 512x512 dimensions, input into model, and then converted the coordinates of Bounding Box back to base dimensions.
3. Run on XSEDE, 20 FPS using 4 GPUs.

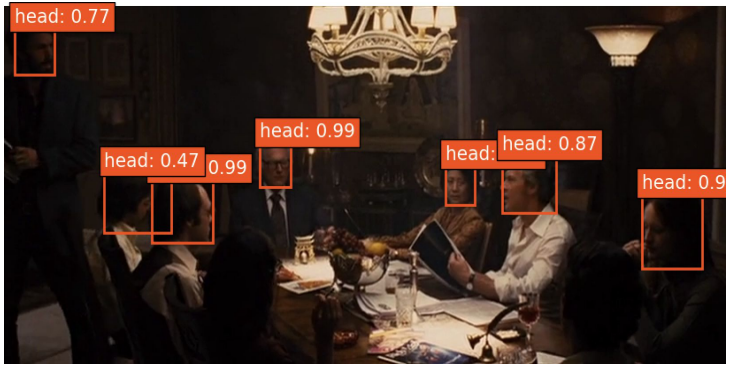
Results

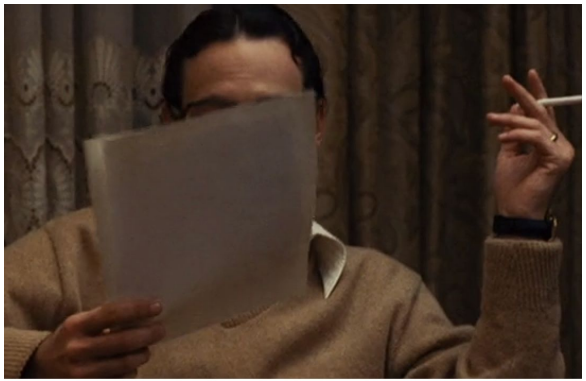
Amadeus



Results

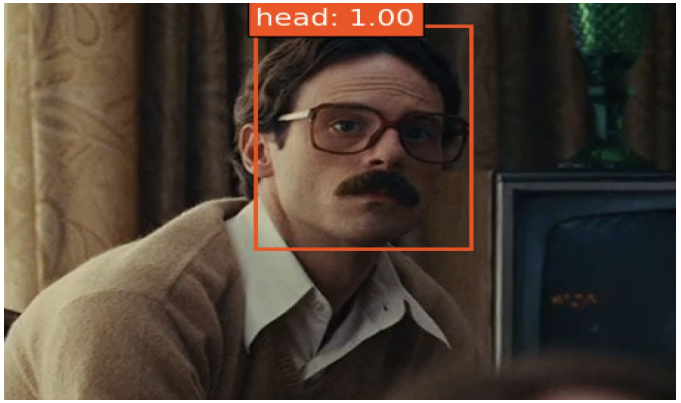
Argo (did really well!!)

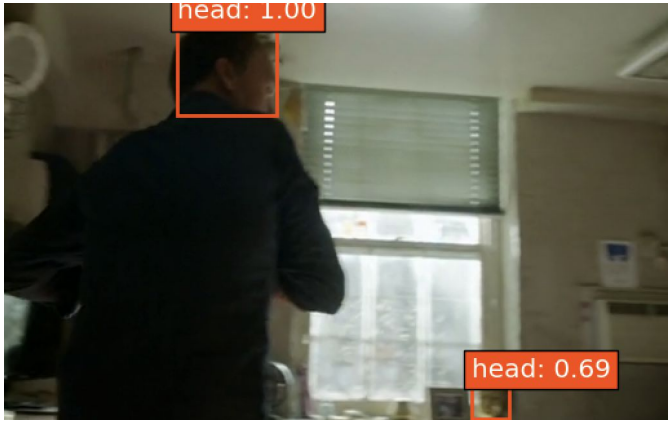




Results

Still Argo





Birdman

Results

1. Blocked faces are hard
2. Half covered faces as well
3. When heads have little to no facial features, accuracy drops tremendously

What's Next?

1. Evaluation metrics
    a. **Quantitative:** Compare generated bounding boxes vs Data face labels
    b. Qualitative:
        i. Visual inspection of random frames for each movie clip
        ii. Stitch together frames and create videos for visualization
2. Extending to cuts/shots detection in frames?...
3. Someway to carry information from frame to frame? And not treat each picture as independent

References

1. Tuan-Hung Vu, Anton Osokin and Ivan Laptev, 2015. Context-aware cnns for person head detection
2. Wei Liu Et. Al, Single Shot MultiBox Detector - https://doi.org/10.48550/arXiv.1512.02325
3. https://github.com/AVAuco/ssd_head_keras
4. Marin-Jimenez, M.~J, Kalogeiton, V. and Medina-Suarez, P. and Zisserman, 2019. Revisiting people, Looking at each other in videos.
5.

Thank you!

Questions?