White Christmas: Remaking Augmented Reality Censorship from Black Mirror with Identity-Preserving Instance Segmentation

Michael Lee (michaellee@cs.stanford.edu)
CS Department, Stanford University

and

Mario Baxter (mariobax@stanford.edu)
CS Department, Stanford University

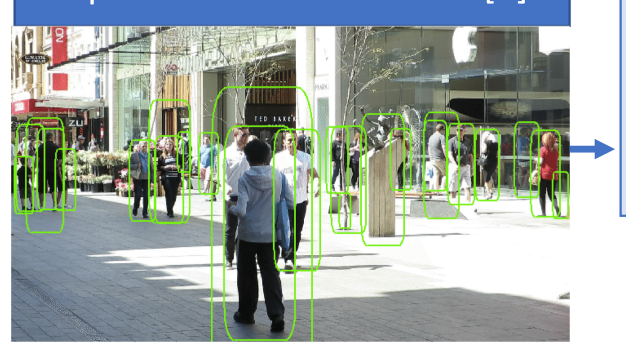
Introduction

Fig 1: Real-time censorship from *Black Mirror* episode *White Christmas* (Netflix, 2014)

- Dystopian censorship as portrayed in *White Christmas*
- Current video object tracking plus segmentation algorithms are very slow [13]
- Multiple object trackers are fast, operating at over 30 FPS [3]

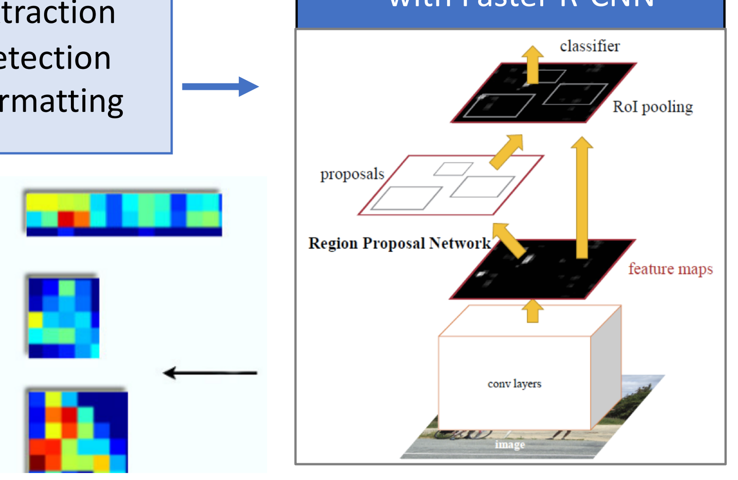
Method

Input: MOT 2016 Dataset [3]



Preprocessing
- Human data extraction
- Detection formatting

DeepSORT
Multiple object tracker with Faster R-CNN



Detections

Proxy: COCO Dataset

Parse out human examples, crop from BB

Pseudo-detections + masks

Branches

Support Vector Machine	RBF Kernel with gamma = 1E-7, individual pixel classification, ~1000 image samples, baseline.

1 pixel FOV | 5 pixel FOV | 15 pixel FOV

Deep Neural Network	10 hidden layers with 500 nodes each, BatchNorm, ReLU, final sigmoid activation. Adam optimizer with binary cross-entropy loss. Sigmoid output to full binary mask, evaluated with binary accuracy.

Fully Connected Conv. Network	DenseNet FCN w/ Bilinear Upsample	DenseNet with 53 convolutional layers and a bilinear upsampling layer, trained with Adam optimizer (LR 1E-3, weight decay=5E-5) with binary cross-entropy loss, BatchNorm, ReLU, final sigmoid activation. Evaluated with binary accuracy.

Results

Evaluation

Metrics all applied pixelwise between test set and prediction binary mask data

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$Precision = \frac{TP}{(TP+FP)}$$

$$Recall = \frac{TP}{(TP+FN)}$$

Data Examples

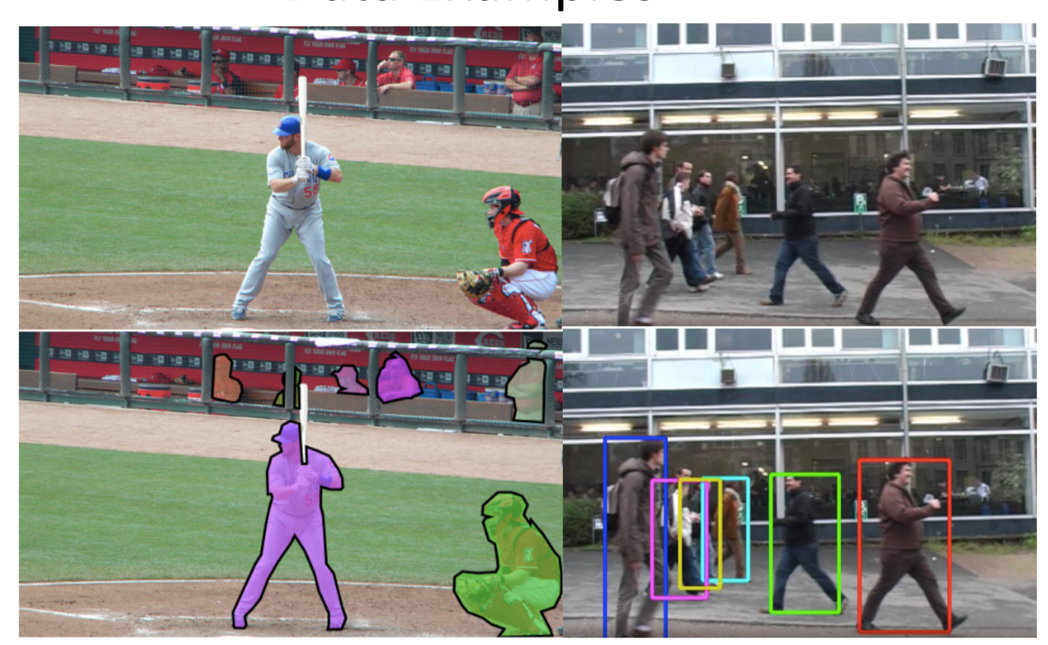


Fig 2: Left – Image from MS-COCO (upper) with annotated GT (below). Right – Image from MOT2016 (upper) with annotated GT (below)

Support Vector Machine

Fig 3: 3 image overfitted example on 5 pixel FOV SVM with 89% accuracy. Pixel-by-pixel evaluation.

10-Layer Neural Network

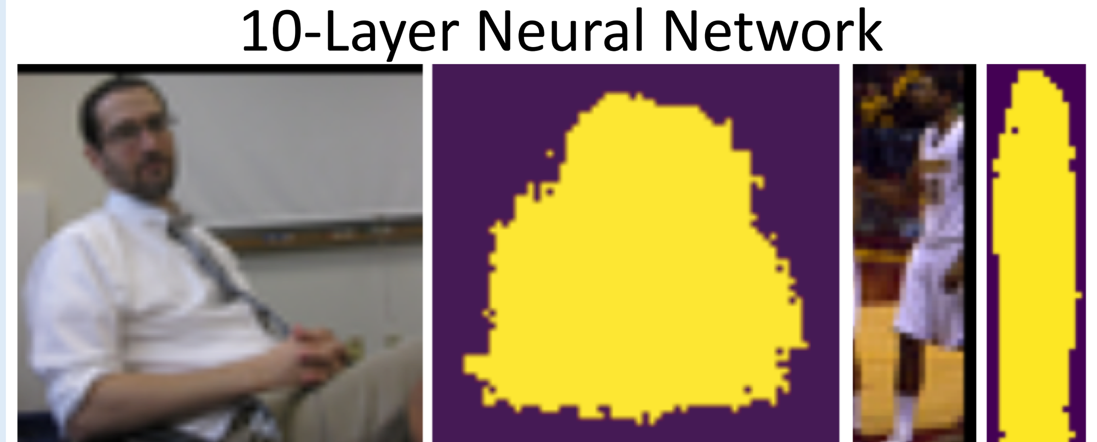


Fig 4: Left two images - average example of input/output pair for wide image. Right two – average input/output for narrow. Demonstrates 'blobbiness' of prediction.

Test Set Performance

Model	Accuracy	Precision	Recall	Speed
SVM 1 pixel FOV	54.17%	-	-	-
SVM 5 pixel FOV	56.05% (3 image overfit, 89.3%)	-	-	-
SVM 15 pixel FOV	51.20%	-	-	-
10-Layer Neural Network	61.38%	0.6012	0.6315	106 FPS
FCC Network w/ DenseNet	81.10%	0.7008	0.8341	35 FPS

FCC Network with DenseNet

Fig 5: Left 4 images – FCC mask of man from sequence 5 frames apart, Right 4 images – FCC mask of woman from sequence 5 frames apart

Full Pipeline

Fig 6: Full pipeline example. Top is the object tracking scene with bounding boxes. Below shows numbered mask examples from the frames from the FCC output.

Results (cont.)

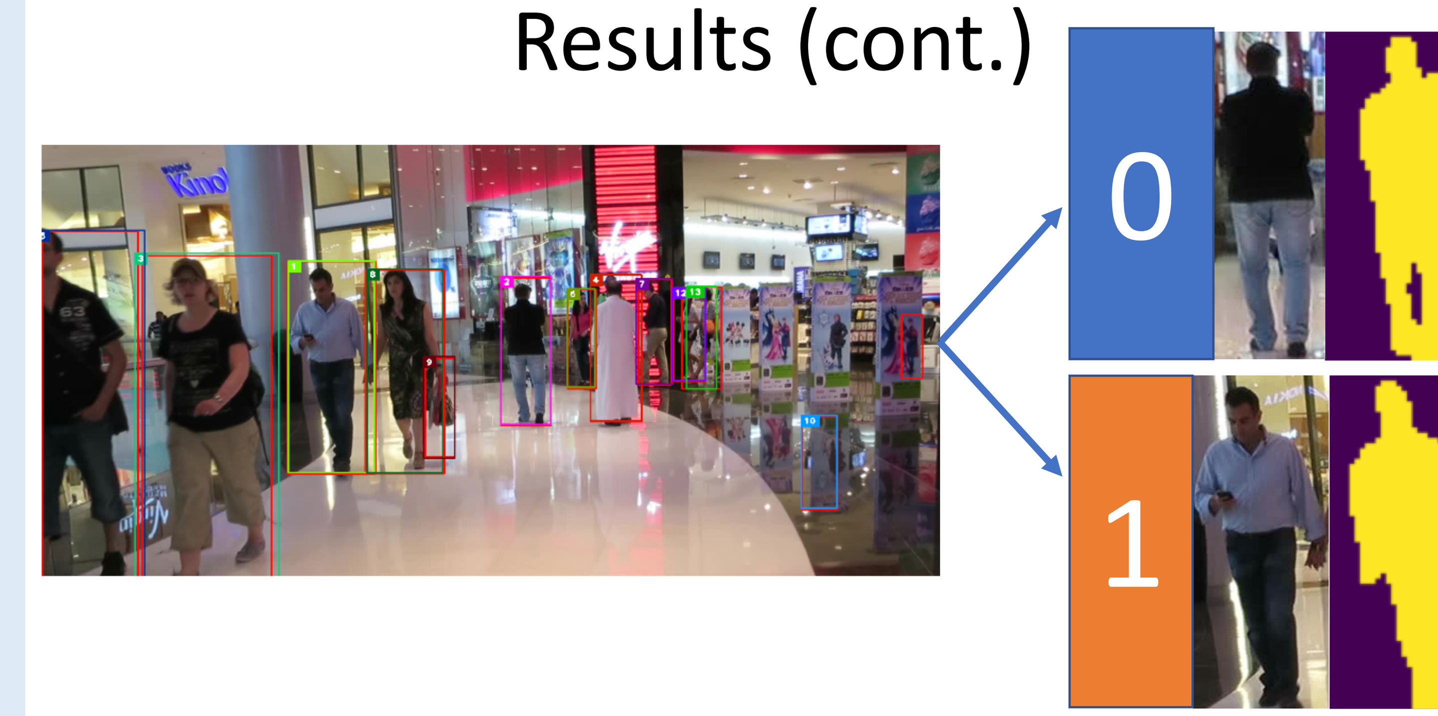


Fig 7: Another full pipeline example. Left is shopping mall single frame, while right is two selected instances from the frame.

Discussion

- SVM completely ineffective, likely not a good domain for application.
- Neural network had weak performance, with no person-like characteristics. However, clearly did learn general location.
- FCC Network had strong performance, with 81.1% binary accuracy.
- Failure cases of FCC still mostly robust to application space. This can be seen in high recall value of 0.834.
- High speed can be observed in all models.

Fig 8: Failure case of FCC network. Notice full coverage of subject despite overmasking.

Conclusion

- Overall, a strong pipeline for video object tracking plus segmentation, especially if precision unneeded
- Huge speed improvements allow for real-time application
- Real-time camera input as well as post-mask blurring needed to recreate *White Christmas.*