

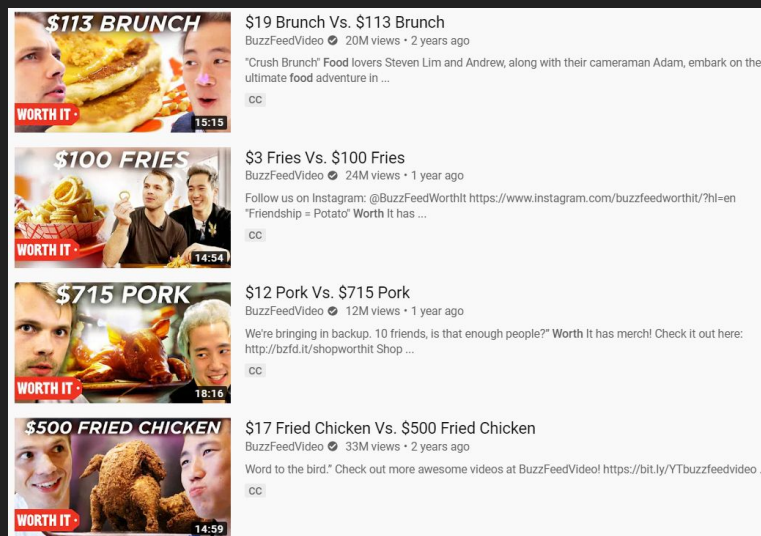
Food Video Localization

Jerry Yu

Problem

- Lots of food locations on Yelp, Google maps, but no videos
- Lots of food videos on Youtube but unindexed
- Hard to find videos of restaurants
- Enable a video based Yelp

- We want to know:
 - Where are the food scenes in the video?
 - Where are the locations visited?
 - What are the foods eaten?

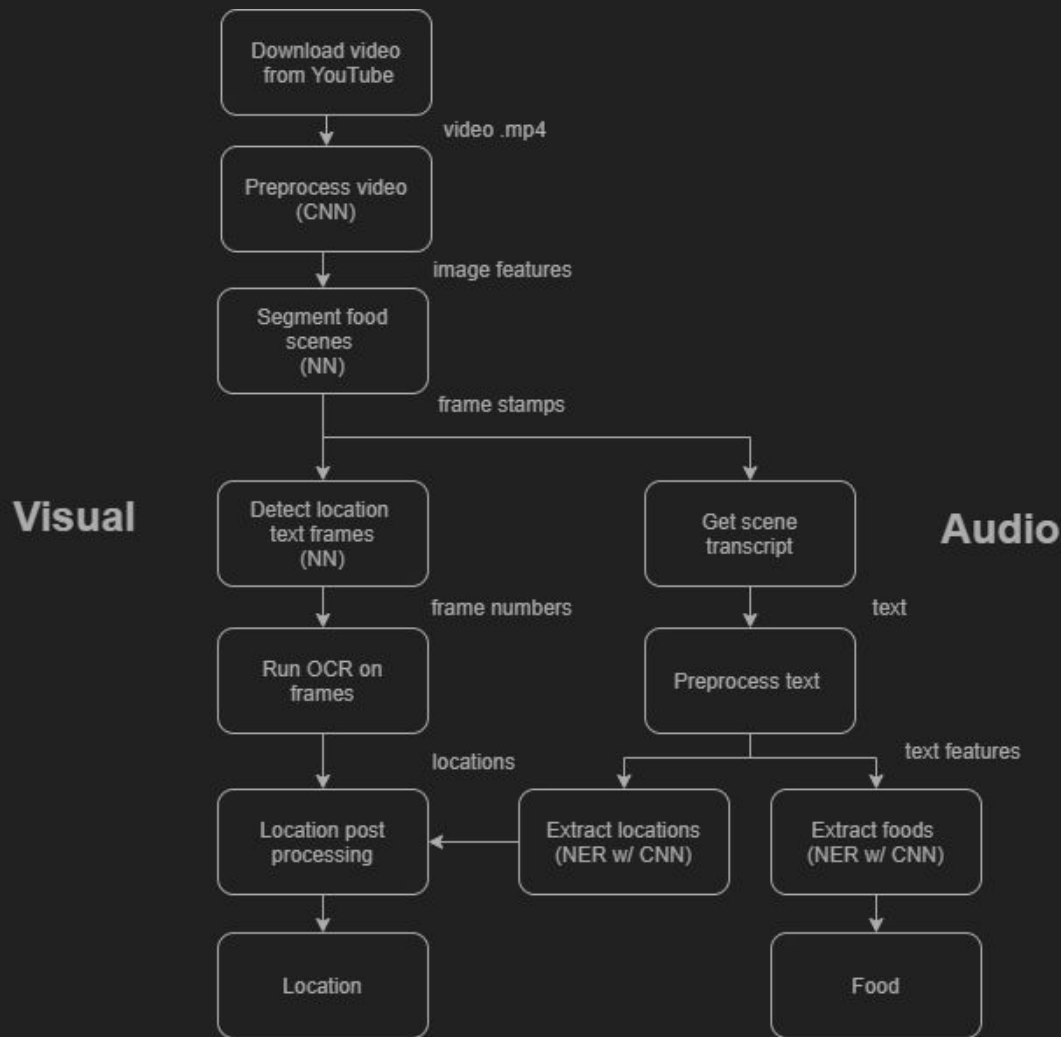


Data

- BuzzFeed series “Worth It” from YouTube, 60 videos
- 180 total food segments
- Mix of transition scenes and food scenes
- In each scene location, the hosts visit a restaurant and try a unique food
- Each video around 15 min, each location around 3.3 min



Pipeline



Food scene segmentation

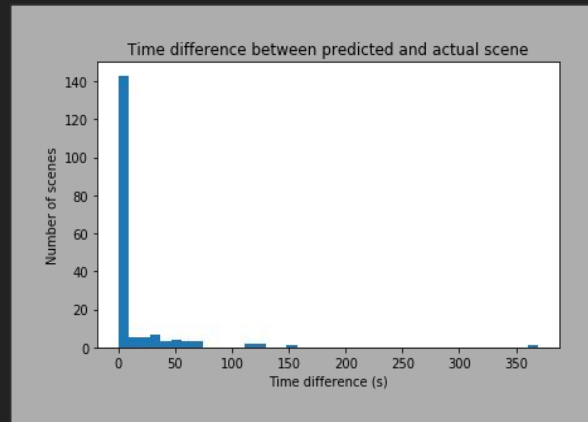
- ResNet CNN features (2048) extracted for every 10th frame
- 5-fold Cross validation
- Train on 48 videos (~96000 frames), validate on 12 videos (~21200 frames)

Method	Cross validation accuracy
NN 6 layers 64 units with dropout and batch norm	93.825%
NN 3 layers 64 units with dropout and batch norm	93.721%
NN 6 layer 64 units	93.033%
SVM (C=10, RBF kernel)	~90% on smaller dataset



Food scene segmentation

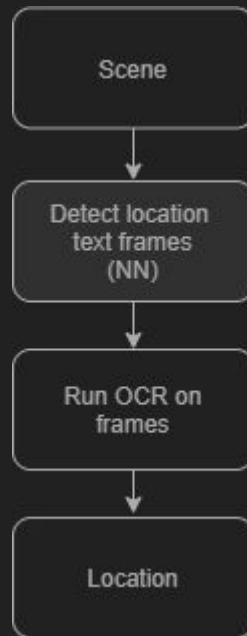
- Post processing to go from per frame labels to scenes
 - Treat 10 keyframes in a row as transition
 - Filter out scenes < 1 min
 - Filter out transitions < 10s
- Identified 179 out of the 180 total scenes (1 FN)



	Time		
True positive	34979.83 s	F1	96.76%
False positive (misclassified)	1351.88 s	Recall	97.25%
False negative (missed)	990.17 s	Precision	96..28%

Localization (Visual)

- Classify if the frame has location information or not
- Can be in scene or transition before scene
- Don't have to run OCR on every frame (OCR output is very noisy)
- Use ResNet CNN features
- Optimize for high recall
- Best model: NN 3 layer 64 unit with dropout, batch norm
- Cross validation: **94.4% Recall**



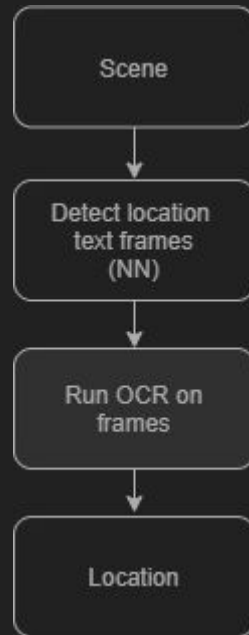
OCR Problems

- Video frames are noisy
- Text does not have flat background
- Text can be different colors
- SOTA OCR (Tesseract, EAST) has problems dealing with text in raw video frames

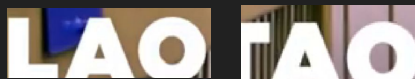
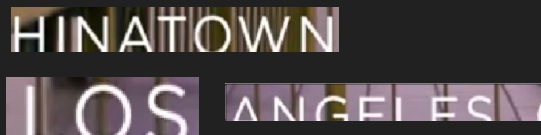
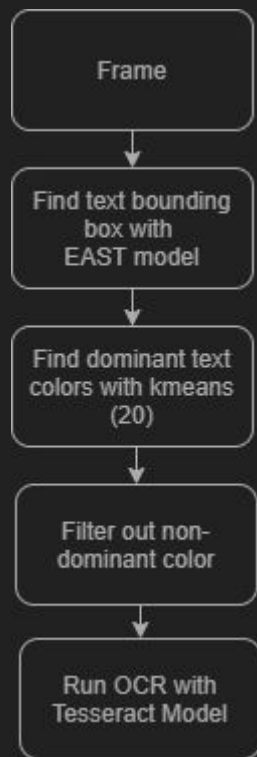


Example output without preprocessing:

I'L- .'
C—I:fl-
#WW *
I|| WAN-W "fig-m,
.141;
.-.....i'q|| 5



OCR



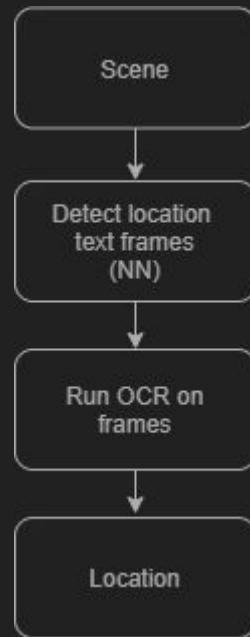
Example output with preprocessing:

\$91!:
LAO TAO
CHINATOWN | LOS ANGELES, CA

OCR Results

- OCR results are noisy
 - ['BIGMISTA'S BARBECUE a. SAMMICH SHOP', 'LONG BEACH, CA'],
 - ["fi'd' ', .7 V ' A V ', '+1 // ~ — // r', 'BIGMISTA'S BARBECUE & SAMMICH snop', 'LONG BEACH, CA .']
- Filtering by using heuristics:
 - Containing US state, containing country, filter out non alphanumeric symbols
- Locations evaluated by creating Google Maps query with output
 - <https://www.google.com/maps/?q=BIGMISTAS+BARBECUE++SAMMlch+shop,+LONG+BEACH,+CA>

Method	Recall	Precision
OCR on every scene frame	47.78%	84.3%
OCR on scene and transition	55%	81.81%
OCR with detected location frames	71.67%	83.16%

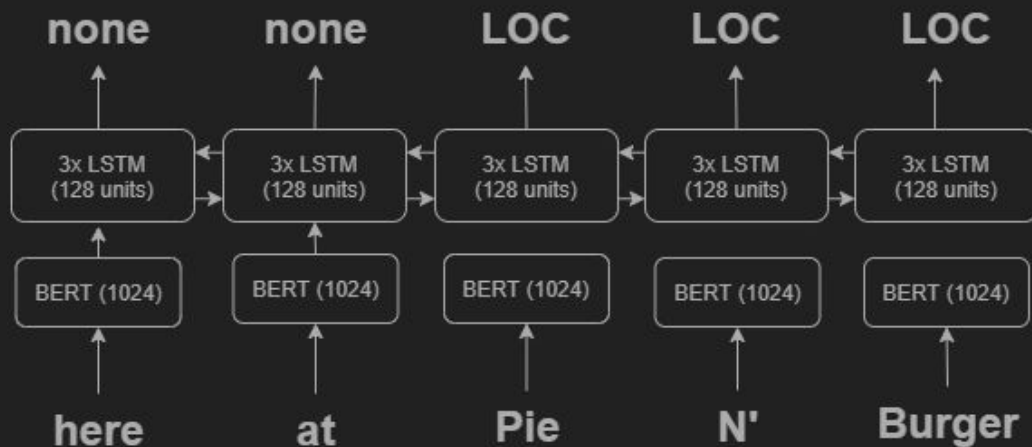


Localization and Food Recognition (Audio)

- 45 videos with manual subtitles, 15 with low quality autogenerated (miss entity names)
- Preprocessing removes audio specific text
 - (Jazz music)
 - [Steven laughs]
- Locations mentioned in audio
 - "My name's Michael Osborn, I'm the owner here at Pie 'N Burger."
 - "We had the line outside the door for a year and a half, and it gave us the opportunity to expand Pasta Sisters to our second location."
- Foods mentioned in audio:
 - We are gonna be trying several of our classic fruit pies today, and we'll feature apple and boysenberry.
 - So let's talk about the lasagna.
- 2 methods explored: NER with LSTM and CNN

NER with LSTM

- Tokenize sentence and get BERT embeddings for each token
- Train Bidirectional LSTM to recognize locations and foods
- Max length of 20 tokens, use first 20 lines of transcript



I'm the owner here at Pie N' Burger

NER LSTM Results

- Cross validation with k=5 for 45 videos
- Downsample to deal with NER tag sparsity

Arch	Location Recall	Location Precision	Food Recall	Food Precision
2 layer BLSTM 32 units	17.89%	--	21.19%	--
3 layer BLSTM 32 units	30.38%	--	26.61%	--
3 layer BLSTM 64 units	35.07%	52.27%	26.48%	35.14%
3 layer BLSTM 128 units	33.27%	50.18%	29.83%	39.18%

Spacy NER CNN

- Fine tune on NER model from Spacy (Bloom embeddings + Residual CNN) by adding LOCATION and FOOD tags
- Use first 20 lines in transcript
- Cross validation k=5

Model	Location Recall	Location Precision	Food Recall	Food Precision
Default Spacy NER (First ORG)	5.55%	25.26%	--	--
Fine tuned Spacy NER	32.07%	48.11%	23.4%	50%
Best LSTM	35.07%	52.27%	29.83%	39.18%





Localization Fusion (Visual + Audio)

- Audio alone does not have enough information about city, state, country
- Use audio location to choose ambiguous output from OCR
 - Audio results are not good enough compared to visual results
- Use audio location as backup
- Future work: explore better ways of fusing

Method	True positive	False positive	Recall	Precision
Best visual localization	128	26	71.11%	83.11%
Visual + LSTM Audio	129	29	71.67%	81.64%
Visual + CNN Audio	132	29	73.33%	81.96%

Issues

- OCR struggling on noisy images in videos
- Foreign languages (not in transcript, hard to OCR)
 - 50 DEUL NYUK jig1, SEOCHO, SEOCHO DSTRCT SEOUL | M23 3
 - So Duel Nyuk, Seocho, Seocho District Seoul
- Small dataset, manually annotate data and bootstrap

	<p>\$3 Ramen Vs. \$79 Ramen • Japan BuzzFeedVideo • 28M views • 2 years ago</p> <p>Worth It is in Japan taking on the most requested dish: Ramen! Food lovers Steven Lim and Andrew, along with their cameraman ...</p> <p>CC:</p>
	<p>\$1 Sushi Vs. \$133 Sushi • Japan BuzzFeedVideo • 17M views • 1 year ago</p> <p>"It's like the Tinder of dining." Credits: https://www.buzzfeed.com/bfmp/videos/67312 Check out more awesome videos at ...</p> <p>CC:</p>
	<p>Buzzfeed Worth It Japan Terumi Ueno</p> <p>\$1,977 Japanese Grapes • 5:52 I Went To Japan To Make The Most Difficult Omelet • 14:34</p> <p>VIEW FULL PLAYLIST</p>
	<p>\$1 Eggs Vs. \$89 Eggs • Japan BuzzFeedVideo • 18M views • 1 year ago</p> <p>"Compliments to the chef, who is the chicken." Credits: https://www.buzzfeed.com/bfmp/videos/67313 Check out more awesome ...</p>

Future work

- Visible way to recognize food
- Better way to fuse audio + visual results
- Text model to predict valid locations from noisy OCR output
- Use autogenerated subtitles