Frame Perfect: Film Director Classification using Convolutional Neural Networks

Christopher Naughton (cwnaught@stanford.edu)
Daniel Thomlinson (thomlins@stanford.edu)

Department of Aeronautics and Astronautics, Stanford University

Introduction

Auteur theory, or the idea that the director of a film is its ”author” and has a personal, recognizable artistic style, has been a significantly influential yet controversial paradigm within the film community since its inception in the 1950s.

Goal: Can a CNN successfully recognize and classify the director from the visuals of a movie scene?

Data Processing

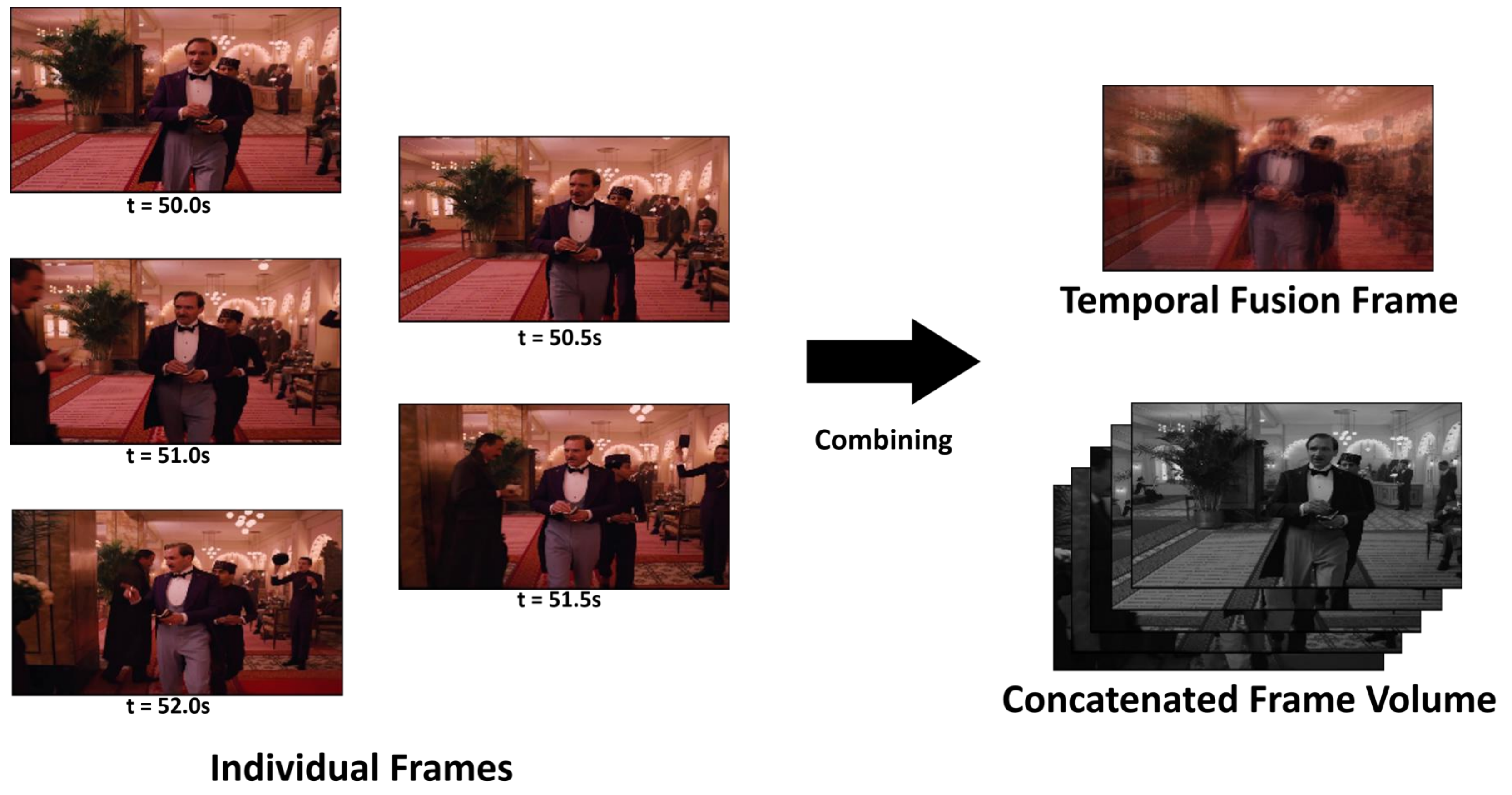
Raw data = Movie clips consisting of ~28,800 frames per director

Goal: Combine frames to incorporate temporal data

Methods: 1) Average frames in a Temporal Fusion Frame (TFF)

2) Stack frames to create a ‘volume’ of pixels

-Similar to Andrej Karpathy’s approach in [1]



Combined 5 frames separated by 0.5s to create these ‘temporal frames’. The concatenated frame was converted to grayscale to meet memory constraints. The final number of temporal frames for each director is:

1) Coen Brothers (495 frames)	3) Quentin Tarantino (470 frames)	5) Wes Anderson (451 frames)
2) Michael Bay (471 frames)	4) Stanley Kubrick (496 frames)	6) Zack Snyder (465 frames)

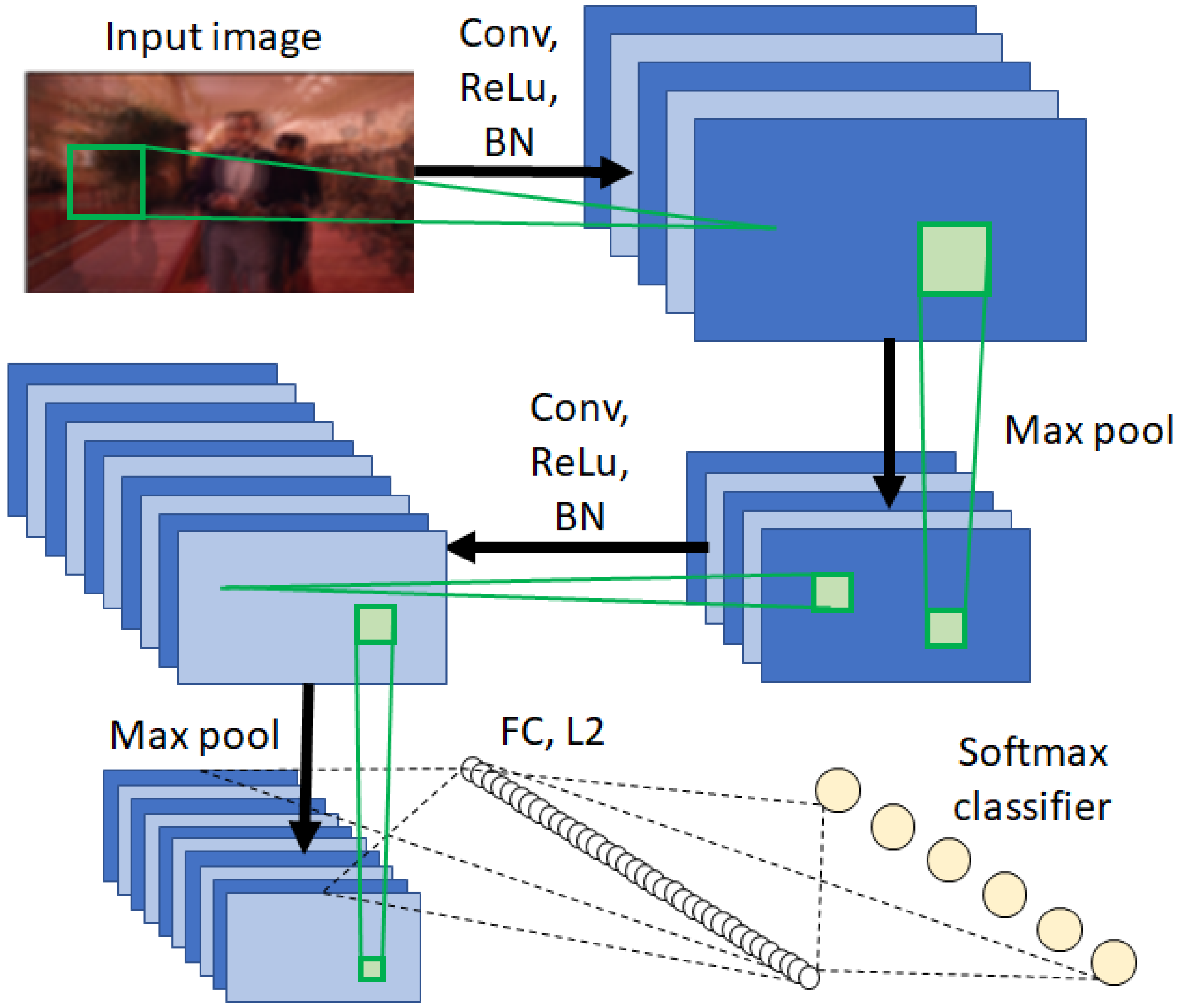
Predictor Architecture

Same architecture but differently tuned hyperparameters for each of the following inputs:

1) Color TFF

2) Grayscale TFF

3) Grayscale volume



$$Loss(p,y,w) = -\underbrace{\sum_{c=1}^{N_{dir}} y_{o,c}\log(p_{o,c})}_{\text{Cross Entropy Loss}} + \underbrace{\frac{\lambda}{2}\sum_{i=1}^{L}\|w_i\|^2}_{\text{L2-Regularization}}$$

Tried to help the model generalize by adding an L2 regularization term to the loss function, using Batch Normalization to shift and scale the data between layers [2], and by using Dropout with a rate of 20% to randomly remove neuron connections from the CNN [3].

Results

Baseline = SVC using a feature vector consisting of average color values

Oracle = Human labeling of test-set frames sent to the CNN

Model	Training Set Accuracy	Test Set Accuracy
Baseline (SVM)	78.1%	32.6%
Color TFF	79.8%	18.1%
Grayscale TFF	41.7%	22.5%
Frame Volume	43.0%	22.1%
Oracle (Human)	~	86.1%

Predicted Director

Actual Director

6	14	11	0	0	10
1	8	8	16	6	10
8	0	25	6	8	5
0	41	1	1	0	0
0	0	6	2	48	0
6	23	5	3	0	4

Baseline

33	2	1	13	0	0
0	12	0	26	21	0
0	14	7	1	41	0
10	12	3	24	0	2
43	0	2	20	0	2
0	1	1	0	47	0

Grayscale TFF

48	0	0	1	0	0
4	38	0	1	9	7
0	0	60	0	0	3
1	0	0	47	0	3
0	0	0	1	66	0
1	14	0	2	0	32

Oracle

Discussion

TFF with large frame spacing were unclassifiable (16%). Slightly better with closer spacing

Predictors did not generalize well

- Temporal frames may not include enough information to differentiate
- Dataset size was limited due to memory constraints

Future work:

- Improve system to allow more data for each director
- Using transfer learning to fine tune a pre-existing image classifier

References

[1] Andrej Karpathy et al. Large-scale video classification with convolutional neural networks. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1725-1732, 2014

[2] Sergey Ioffe and Christian Szegedy. Batch Normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015

[3] Srivastava et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *2014 Journal of Machine Learning Research*, pages 1929-1958