Video-Based Event Recognition

Ian Ballard and Lane McIntosh
Neurosciences Program, Stanford University School of Medicine

Introduction

- How can we automatically extract events from video?
- Human visual systems perform this effortlessly, but state-of-the art classification results still hover around 20% accuracy. [1]
- We consider the problem of event recognition for surveillance videos taken from fixed camera positions where the problem is constrained and events are well-defined
- We develop new models for video-based event recognition based on early visual processing in vertebrates

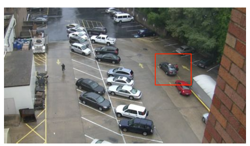
Several examples of people unloading cars. As the availablility of data increases, threre is increased need to automatically recognize events such as these.

Dataset

- Annotated VIRAT video database with fixed surveillance camera footage
- Over 40 GB of 1920x1080 resolution video at 30 frames/second
- Human-labeled by Amazon's Mechanical Turk with 12 different event classifications

Example Frames and Corresponding Labels

Unloading a car



No event

Possible Labels

- person loading an object to a vehicle
- person unloading an object from a vehicle
- person opening a trunk
- person closing a trunk
- person getting into vehicle
- person getting out of vehicle
- person gesturing
- person digging
- person carrying an object
- person running
- person entering a facility
- person exiting a facility

Evaluation

Baselines

Events are likely associated with large frame-to-frame changes

$$\text{prediction}_t = \mathbb{1}\left[\sum_{i,j} \mathbb{1}[|p_t^{(i,j)} - p_{t-1}^{(i,j)}| > \alpha] > \beta\right]$$

α smallest pixel difference we care about
β smallest number of different pixels we care about
$p_t^{(i,j)}$ pixel intensity at position (*i,j*) in frame *t*

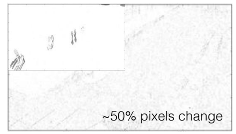
14% of events detected, 99.7% labels correct

100% of events detected, 64% labels correct after re-labeling

All 1's or all 0's baseline

0% of events detected, 99.6% labels correct

100% of events detected, 62% labels correct after re-labeling



Inset: some events are readily distinguishable from difference images. Full: Many movies also involve large pixel differences due to wind and camera instability

Issues

- labeled events are very sparse (~19/6000)
- human labeling only approximates an event's first frame

Re-labeled videos so that every frame during an event is labeled as an event; now 62% of frames are events

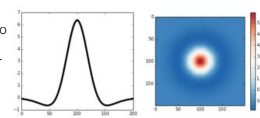
Oracle

- ideal performance is human-level event detection
- we take Mechanical Turk annotations to be ground truth

Pre-processing

"Feature extraction [is] arguably the most important part of machine learning." – Percy Liang

- convert to grayscale
- 30 fps to 1 fps
- downsample resolution to 56x100 pixels
- subtract off previous frame
- Z-score normalize
- convolve with retina-like difference of Gaussians (feature tempate)
- re-label movies with all events
- convert labels from "first frame" to "all frames during each event"
- set absorbing lower poundary for pixel intensities
- flatten into arrays
- concatenate across timepoints and movies

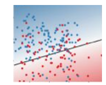


Model and Results

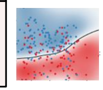
Support Vector Machine

Linear kernel

$$k(x,y) = x^T y + c$$

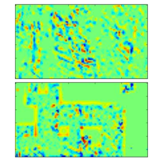


Gaussian kernel



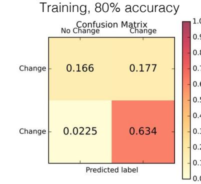
$$k(x,y) = \exp\left(-\frac{||x-y||}{2\sigma^2}\right)$$

Inputs

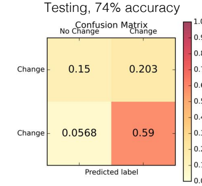


Error Analysis

Training, 80% accuracy

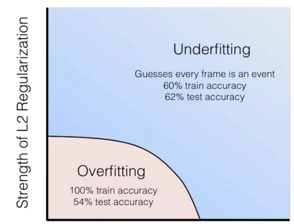


Testing, 74% accuracy



Regularization

Before augmenting feature templates



Strength of L1 Regularization

Conclusions

- Proper labeling and pre-processing was critical for achieving above chance performance
- Regularization controls trade-off between extreme overfitting and extreme underfitting, and increasing the number of feature templates was necessary for exploring the space between these extremes
- Nonlinear features are necessary to achieve above chance performance
- Convolutional networks may provide more robust results

1. Wang, Xiaoyang, and Qiang Ji. A Hierarchical Context Model for Event Recognition in Surveillance Video. *CVF* (2013)