Temporal Models

Greg Mori

CMPT 888
“Bag-of-Words” Models

- **Text document models**
  - “It was the best of times, it was the worst of times.”

- **Bag of Words + Topic Models in Computer Vision**
  - Scenes: Fei-Fei & Perona CVPR’05
  - Objects: Sivic et al. ICCV’05, Fergus et al. ICCV’05, Russell et al. CVPR’06
  - Actions: Niebles et al. BMVC’06
  - Human Poses: Bissaco et al. NIPS’06
Role of Temporal Information

- No temporal info
  - Classify each video frame independently
  - e.g., Efros et al. 03, Shechtman & Irani 05, Fathi & Mori 08
Role of Temporal Information

• **Strong temporal info**
  – Use hidden Markov Model or grammar on top of video frames
  – e.g. Bobick & Ivanov CVPR98, Yamato et al. CVPR92
Role of Temporal Information

- Y. Wang et al. HUMO/PAMI is somewhere in between
  - Use bag of frames representation
  - Capture some temporal structure (co-occurrences of actions)
  - Simpler than full temporal models
HIDDEN MARKOV MODELS FOR ACTION RECOGNITION
**HMMs**

- **Sensor Markov assumption:** $p(x_t|z_{1:t}, x_{1:t-1}) = p(x_t|z_t)$
- **Stationary process:** transition model $p(z_t|z_{t-1})$ and sensor model $p(x_t|z_t)$ fixed for all $t$ (separate $p(z_1)$)
- **HMM special type of Bayesian network, $z_t$ is a single discrete random variable:**

![Diagram of HMM sequence with nodes $Z_i$ and $X_i$](image)

- **Joint distribution:**

$$p(z_{1:t}, x_{1:t}) = p(z_1) \prod_{i=2:t} p(z_i|z_{i-1}) \prod_{i=1:t} p(x_i|z_i)$$
Using HMMs for Action Recognition (Yamato et al.)

• Each frame is mapped to a discrete symbol
  – Visual word

• For each action category, learn HMM model parameters
  – Transition matrix, emission matrix, prior

• Recognition: compute likelihood of observed symbols under each HMM
  – Choose action category that produces highest likelihood
Features (Yamato et al.)

- **Mesh features**
  - Looks like HOG on foreground mask!
- **Some form of vector quantization (VQ) is used**
  - Manual/random(?) selection of prototypes
  - K-means?
- **1992 vs. 2010? 😊**
Experiments

- Limited data/compute available
- 6 actions, more variability than KTH/Weizmann
- Good results
  - Likely would be quite accurate with more training data / parameter CV
Other work

- Bobick & Ivanov CVPR 98
  - Recognize hand gestures
  - Grammar to describe a gesture

$$G_{\text{square}} :$$

<table>
<thead>
<tr>
<th>SQUARE</th>
<th>RH</th>
<th>[0.5]</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LH</td>
<td>[0.5]</td>
</tr>
<tr>
<td>RH</td>
<td>TOP UD BOT DU</td>
<td>[1.0]</td>
</tr>
<tr>
<td>LH</td>
<td>BOT DU TOP UD</td>
<td>[1.0]</td>
</tr>
<tr>
<td>TOP</td>
<td>LR</td>
<td>[0.5]</td>
</tr>
<tr>
<td></td>
<td>RL</td>
<td>[0.5]</td>
</tr>
<tr>
<td>BOT</td>
<td>RL</td>
<td>[0.5]</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>[0.5]</td>
</tr>
<tr>
<td>LR</td>
<td>left-right</td>
<td>[1.0]</td>
</tr>
<tr>
<td>UD</td>
<td>up-down</td>
<td>[1.0]</td>
</tr>
<tr>
<td>RL</td>
<td>right-left</td>
<td>[1.0]</td>
</tr>
<tr>
<td>DU</td>
<td>down-up</td>
<td>[1.0]</td>
</tr>
</tbody>
</table>
Other work

- Gupta et al. CVPR 2009, others
- “Storyline” model explaining video
BAG OF FRAMES MODEL
Role of Temporal Information

- Y. Wang et al. HUMO/PAMI is somewhere in between
  - Use bag of frames representation
  - Capture some temporal structure (co-occurrences of actions)
  - Simpler than full temporal models
Bag-of-Words Sequence Model
Codebook Formation

Feature Representation

Vector Quantization
Semi-Latent Dirichlet Allocation

Learning is easier due to decoupling of model parameters
cf. Blei et al. JMLR 2003
Experiments: KTH dataset

- Benchmark dataset
  - 6 actions
  - 25 subjects
  - 4 scenarios

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (sLDA)</td>
<td>91.2%</td>
</tr>
<tr>
<td>Liu &amp; Shah CVPR08</td>
<td>94.2%</td>
</tr>
<tr>
<td>Jhuang and Poggio ICCV07</td>
<td>91.7%</td>
</tr>
<tr>
<td>Niebles &amp; Fei-Fei BMVC06</td>
<td>81.5%</td>
</tr>
<tr>
<td>Schuldt &amp; Laptev ICPR04</td>
<td>71.7%</td>
</tr>
</tbody>
</table>
Experiments: Soccer Dataset

- Real actions, moving camera, poor video
- 8 classes of actions
- 4500 frames of labeled data

<table>
<thead>
<tr>
<th>Action</th>
<th>Our method (sLDA)</th>
<th>Efros et al. (k-NN)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Run left 45</td>
<td>0.64</td>
<td>0.67</td>
</tr>
<tr>
<td>Run left</td>
<td>0.77</td>
<td>0.58</td>
</tr>
<tr>
<td>Walk left</td>
<td>1.00</td>
<td>0.68</td>
</tr>
<tr>
<td>Walk in/out</td>
<td>0.86</td>
<td>0.79</td>
</tr>
<tr>
<td>Run in/out</td>
<td>0.81</td>
<td>0.59</td>
</tr>
<tr>
<td>Walk right</td>
<td>0.86</td>
<td>0.68</td>
</tr>
<tr>
<td>Run right</td>
<td>0.71</td>
<td>0.58</td>
</tr>
<tr>
<td>Run right 45</td>
<td>0.66</td>
<td>0.66</td>
</tr>
</tbody>
</table>
Experiments: Irregularity detection

• sLDA is full probabilistic model
• Can detect most unusual sequences via likelihood
  – Sequences with lowest likelihood under model shown
CAMERA NETWORKS
Multiple Cameras

• In many situations, we have a set of cameras views of a scene
  – Not necessarily overlapping
• Need models for activities that span these different views
• Loy et al. ICCV09 paper is an example of this type of work
Underground Scenario
Detect Abnormal Events

• Approach
  – Build model of “normal”
    • Incorporate time delayed relationships over scene
    • Learn structure of these relationships
  – Score clips by likelihood under this model
Camera View Regions (CVPR09)

• Decompose each camera view into regions
Features

• Divide camera view into 10x10 pixel blocks
• Count:
  – #foreground pixels in block
  – #moving pixels in block
• Aggregate over time
• Compute correlation between time series for pairs of blocks
• Perform spectral clustering on blocks
Time Delayed Analysis

• Given all the regions in all camera views, what are the relationships between them?
  – Describe each region with Gaussian mixture model on same foreground/moving features
  – Compute mutual information between pairs of time series
  – Search for best temporal offset between pairs of time series
    • Offset that maximizes mutual information
Structure Learning

- Given all regions, and MI between each pair, build a Bayesian Network
  - Use Chow-Liu algorithm
  - Finds tree-structured BN that keeps most important edges
Other details

- Bayesian parameter learning
- Aggregate likelihoods over time to smooth out noise (Cumulative Abnormality Score)