Statistical and Structural Recognition of Human Actions

Structural Methods
POSE ESTIMATION AND ACTION RECOGNITION
Pose Estimation for Action Recognition

- Pose seems sufficient for certain action categories
- Remove effects of clothing, lighting variation from representation

Pose Estimation – Problem Definition
Problem
Models vs. Exemplars

- Two broad classes of approaches
  - Match templates (exemplar-based)
  - Fit a human body model
Exemplar Matching For Pose Estimation

Mori & Malik PAMI 2005
Shakhnarovich, Viola & Darrell ICCV 2003
Bourdev & Malik ICCV 2009

Database of Exemplars
Human Body Models for Pose Estimation

Pictorial Structures model

\[
Pr(L|I, \Theta) \propto \exp \left( \sum_{(i,j) \in E} \psi(l_i - l_j) + \sum_{i=1}^{K} \phi(l_i) \right)
\]

spatial prior  
part appearance

Felzenszwalb & Huttenlocher CVPR 2000
Ramanan NIPS 2006
Ferrari, Marin & Zisserman CVPR 2008
Action from Pose I: Model Likelihood

- Detect corners in images/video
- Assess likelihood under action-specific pose model
- Discriminate between walking directions, bicycle riding

Song, Goncalves & Perona NIPS 2001, PAMI 2003
Action from Pose II: Key frame Templates

- Key frame matching to test sequence to find similar poses
  - Shape matching on edge maps using order structure

Sullivan & Carlsson ECCV 2002
• Automatic person detection-tracking
• Compare quantized pose to labeled training poses
  – Smooth over time

M is quantized 3d pose
T is root orientation

Ramanan & Forsyth NIPS 2003
Action from Pose III: Pose Search

• Video shot retrieval from pose
  – Either *query-by-example* or classification
  – Focus on upper body pose
    • Pictorial structures model

Ferrari, Marin & Zisserman CVPR 2009
• SVM on descriptors of absolute & relative part locations, segmentations
  – Include short tracks for robustness

CODE AVAILABLE ONLINE
Focus on discriminative elements of pose for action classification
Use exemplar-based “poselet” representation

Yang, Wang & Mori CVPR 2010
Successful classification examples

Unsuccessful classification examples
Action from Pose V: Poses and Objects

\( A: \)
- Tennis forehand
- Croquet shot
- Volleyball smash

\( O: \)
- Tennis racket
- Croquet mallet
- Volleyball

\( H: \)
- Intra-class variations
  - More than one \( H \) for each \( A \);
  - Unobserved during training.

\( P: \)
- \( l_p \): location; \( \theta_p \): orientation; \( s_p \): scale.

\( f: \)
- Shape context. [Belongie et al, 2002]

Yao & Fei-Fei CVPR 2010
Learning Results

Cricket defensive shot

Cricket bowling

Croquet shot
Analyzing Image Collections

• Build action models from web search results

Ikizler-Cinbis, Cinbis, Sclaroff ICCV 2009
Clustering Actions

- Find repeated poses in a dataset

Wang, Jiang, Drew, Li, Mori CVPR 2006
Dataset: PASCAL VOC Action Classification

- Person location given
- Classify into one of 9 categories

Riding horse  Reading book  Taking photo

Riding bike  Play instrument  Running

Phoning  Use computer  Walking
Summary

• Pose as representation for action recognition
  – Captures much information about action
  – Invariance to clothing / lighting effects
  – Model and exemplar based representations

• New direction: Action recognition from still images
  – Image retrieval and analysis
  – An important cue for video-based action recognition
  – Pose seems essential
SCENE MODELS
Getting the Whole Picture
Scene Model Ingredients

• Describe low-level components
  – Actions of individual people
  – Movement of pixels

• Identify key objects or locations in scene
  – Buildings, roads, etc.

• Model interactions between people, objects, and locations
Scene Models I: Rule-based System

• Detect and track moving objects
• Manually identify key regions in scene
  – E.g. road, checkpoint
• Scenarios describe relative arrangements of objects in scene
  – E.g. proximity of car to checkpoint
  – Notions of scene context

Medioni, Cohen, Bremond, Hongeng, Nevatia PAMI 2001
Scene Models II: Bayes Nets

- Detect and track players, ball
- Low-level action detectors for individual players
- Hand-constructed Bayes net for each activity
  - Spatial and temporal relations between low-level actions

Intille & Bobick CVPR 1999
Scene Models III: Unsupervised Learning of Unusual Events

- Global, frame-level feature
  - Bag-of-words representation
- Detect unusual events by clustering
  - Isolated, varied clusters are unusual

Zhong, Shi & Visontai CVPR 2004
• Cheating detection in simulated card game

• Real-world highway dataset
  – Cars pulling off road, backing up, U-turns
Scene Models IV: Unsupervised Hierarchical Scene Model

- Describe moving pixels by location and motion direction
  - No object detection
- Use as visual words in Latent Dirichlet Allocation (LDA) type model
  - Infer low-level actions from words

Wang, Ma, Grimson PAMI 2009

Blei, Ng, Jordan JMLR 2003
• Higher-level activity analysis
  – Distribution of low-level actions over entire scene

• Applications
  – Temporal segmentation by activity
  – Abnormality detection
Scene Models V: Hierarchical with Temporal Dependencies

- Hierarchical Dirichlet Process model
  - Learn number of activities automatically

Kuettel, Breitenstein, van Gool & Ferrari CVPR 2010
traffic light controlled scene

- continuous video
- annotated with states and history
- 3x speed
Scene Models VI: Multi-Camera Scene Decomposition

Loy, Xiang & Gong CVPR, ICCV 2009
• Consider time-delayed correlations between regions
  – Applications to irregularity detection
Scene Models VII: Person-Person Context

Choi, Shahid, & Savarese VS 2009
Lan, Wang, Yang, & Mori SGA 2010, NIPS 2010
Scene Models VIII: Storyline Model

- Captioned baseball videos in training
- Build AND-OR graph representation of activities
  - AND specifies elements of an activity that must occur
  - OR allows variation in how an element appears
- Describe low-level tracks using STIPs
- Match tracks to actions in AND-OR graph

Gupta, Srinivasan, Shi, Davis CVPR 2009
• Scene modeling to look at the big picture
• Feature representations
  – Holistic: describe entire scene, irrespective of individuals
  – Local: describe actions of individuals
• Structure of activities
  – Model free: clustering-type approaches
  – Strong models: grammars, probabilistic models
CONCLUSIONS
Computer vision grand challenge: Video understanding

Objects:
cars, glasses, people, etc...

Actions:
drinking, running, door exit, car enter, etc...

Scene categories:
indoors, outdoors, street scene, etc...

Geometry:
Street, wall, field, stair, etc...

Constraints
Future Directions I: Problem Definitions

- Reading book
- Play instrument
- Riding horse
- Riding bike
Datasets & Baselines

• Standardization of datasets for field
  – Allow comparison of algorithms
    • E.g. KTH for low-level features, atomic actions
  – Fair tuning of model parameters

• New algorithms compare to baselines
  – Bag-of-words on densely sampled STIPs
  – Pose estimation (Ferrari et al. code)
  – HOG SVM (Dalal & Triggs code, Ramanan code)
Datasets & Baselines

• Standardization of datasets for field
  – Don’t feel constrained by the existing problem definitions
  – Do make your new dataset available
    • Should clearly specify separate training and test sets

• New algorithms compare to baselines
  – Do use reasonable variant of standard baselines for your new problem
Future Directions II: Back to Basics
Future Directions II: Back to Basics

• Even atomic low-level actions are very difficult to detect reliably
  – Far more work needed on representations for the action of a single person
  – Features
  – Temporal representation, smoothing
  – Tracking
  – ...


Future Directions III: Obtaining Data

1. Cameras and bandwidth are cheap
2. Lots of training data is potentially available

+ = Training data

Potential for the huge progress … if we can get the data
## Readily available video annotation

<table>
<thead>
<tr>
<th></th>
<th>Aligned with video</th>
<th>Describes visual content</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subtitles</td>
<td>Yes</td>
<td>No</td>
<td>DVD, Internet</td>
</tr>
<tr>
<td>Scripts for TV series, movies and sport games</td>
<td>No</td>
<td>Yes</td>
<td>Internet, e.g. <a href="http://www.dailyscript.com">www.dailyscript.com</a></td>
</tr>
<tr>
<td>Plot summaries and synopses</td>
<td>No</td>
<td>Yes, sparsely</td>
<td>Internet (e.g. IMDB)</td>
</tr>
<tr>
<td>Instruction videos</td>
<td>No</td>
<td>Yes</td>
<td>Internet, e.g. <a href="http://www.videojug.com">www.videojug.com</a></td>
</tr>
<tr>
<td>Descriptive Video Service</td>
<td>Yes</td>
<td>Yes</td>
<td>DVD, rare</td>
</tr>
<tr>
<td>Word tags</td>
<td>No</td>
<td>Yes, sparsely</td>
<td>Internet (e.g. YouTube)</td>
</tr>
<tr>
<td>Manual labelling, Human Computation</td>
<td>??</td>
<td>??</td>
<td>Mechanical Turk, ESP Game, Grad undergrad students</td>
</tr>
</tbody>
</table>
Open questions:

- How to benefit from the structure of the human body in complex situations, e.g. heavy occlusions, uniformly colored clothing?
- Will action classification generalize over different video domains: Movies, TV, YouTube, Surveillance video?
- What is the useful action vocabulary? Are we trying to solve the right problem? How can we visualize/display the results?

Interesting novel directions:

- Use actions for recognizing functional and physical object properties, e.g. "sitable", "eatable", "heavy", "solid" objects…
- Action prediction, i.e. what can happen in the given situation: e.g. is it dangerous to cross this road?
- Explore more sources of strong and weak supervision: Manual surveillance, Descriptive Video Service (DVS), YouTube tags; Transcripts of sports games; Instruction videos.
References - Preliminaries

References – Motion/shape Templates

• E. Shechtman and M. Irani. Space-time behavior based correlation. In CVPR 2005
References – Local Features

• P. Matikainen, M. Hebert, R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. ICCV workshop on Video-oriented Object and Event Classification, 2009
References – Human Pose and Actions

References – Periodic motion


References – View invariance

Workshop materials available:
https://sites.google.com/site/humanactionstutorialeccv10/