Action recognition

Deva Ramanan

UC Irvine

(Many slides from Ivan Laptev, David Forsyth)
What is the problem we’re trying to solve?

Surprisingly hard to be precise about....
Computer vision grand challenge: Dynamic scene 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…
Computer vision grand challenge: Dynamic scene 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
Computer vision grand challenge: Dynamic scene understanding

Objects: cars, glasses, people, etc…

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

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

Geometry: street, wall, field, stair, etc…

constraints
Why focus on people/actions?
Lots (most?) of training data comes in video form

Data:

- >34K hours of video uploads every day
- ~30M surveillance cameras in US => ~700K video hours/day
Lots (most?) of test data comes in video form

- Wearable: where did I leave my keys?
- Predicting crowd behavior
- Counting people
- Education: How do I make a pizza?
- Motion capture and animation
Why focus on people/actions?

How many person-pixels are in the video?
Why focus on people/actions?

How many person-pixels are in the video?
Why focus on people/actions?

How many person-pixels are in the video?

Movies: 35%

TV: 34%

YouTube: 40%
One approach

Generalize ‘object detection’ techniques to spacetime windows

Spacetime (XYT) template

Grab-Cup Event

Ke et al, ICCV05

Thursday, August 8, 2013
One approach

Generalize ‘object detection’ techniques to spacetime windows

Spacetime (XYT) template

Grab-Cup Event

Ke et al, ICCV05
One approach

Generalize ‘object detection’ techniques to spacetime windows

Spacetime (XYT) template

Grab-Cup Event

Ke et al, ICCV05
Spacetime correlation

Shechtman & Irani, CVPR05
Spacetime correlation

Shechtman & Irani, CVPR05
Spacetime correlation

Shechtman & Irani, CVPR05
Flexible spacetime part templates

- Sit
- Run
- Walk
- Stand
- Catch
- Trip

Scalable object and action recognition for automated image understanding

Use or disclosure of data contained on this sheet is subject to the restriction on the title page of this proposal white paper.

3. Variables - duration action detection by tracking pictorial structures

We propose to develop spatiotemporal part representations that combine DPM/FMP representations with object tracking and reference pose markers to recognize activities that are independent to action duration.

We will leverage SRI's capabilities in real-time video-based object tracking and spatiotemporal feature representations to efficiently track parts of a person over the duration of the video. A cartoon illustration of a person drinking from a bottle is shown in Figure 6.

In the proposed approach, we will capture kinematic constraints between parts on the person's body in a single frame. In the figure, this is shown by the white skeleton connecting the head to parts on the left arm (green), right arm (red) and the torso (dark blue). We will then track the detected object between successive frames. Within the tracked region, object parts are extracted to minimize drift. Dynamic constraints are captured across consecutive frames for each part. We will also use connectivity between parts on the person and the interacting object, since our experience from [3] indicates that it improves object detection performance. We will encode part-specific trajectories and the object track as features that capture salient temporal motion and use them to classify component parts of an activity.

5. Recent Technical Breakthroughs

Our proposed effort will build up on several recent technical breakthroughs in object and action recognition, and pose estimation. These include pioneering work on modeling objects as a collection of deformable parts and a latent SVM framework to train object appearances and deformations [1]; state-of-the-art frameworks that extend [1] for pose estimation and person-object composite detections [2], [3]; approaches to model a large number of views and shapes of cars using a small number of view-based templates [4]; and significant algorithm speedups [5] [1]: "Object Detection with Discriminatively Trained Part-Based Models". PAMI 2008 [2]: "Articulated Human Detection with Flexible Mixtures of Parts". CVPR 2009 [3]: "Detecting Actions, Poses, and Objects with Relational Phraselets". ECCV 2012 [4]: "Analyzing 3D Objects in Cluttered Images". NIPS 2012 [5]: "Steerable Part Models". CVPR 2012

6. Funding plan showing requested funding per fiscal year

<table>
<thead>
<tr>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>$250,000</td>
<td>$250,000</td>
<td>$250,000</td>
</tr>
</tbody>
</table>

Figure 6: Illustration of parts detection on a person and parts tracking across video frames for a person drinking from a bottle.

Thursday, August 8, 2013
But what’s the desired output here?
But what’s the desired output here?

Will we have a “throw cat in the trash bin” template?
But what’s the desired output here?

Will we have a “throw cat in the trash bin” template?
Long-tail distributions

Challenge: actions seem to follow an extremely heavy tail distribution
Complicates dataset collection and annotation
Roadmap

Data/benchmark analysis

Spatiotemporal features

Spatiotemporal models

Walk → Stand → Crouch
Like it or not, crucial for advances in the field

Large-scale annotated video datasets are more rare - why?
Action recognition benchmarks

1) Video is cumbersome to label (difficult to define natural categories outside sports)
2) Collecting interesting but natural video is surprisingly hard
3) Most current work focuses on K-way classification (similar to image recognition 10 years ago)
KTH

Classification performance around 100%
“Outdated”
“Woodworking” action

board-trick, feeding animal, fishing, wedding, woodworking, birthday, changing vehicle tire, flash mob, vehicle unstuck, grooming an animal, sandwich making, parade, parkour, repairing appliance and sewing,...

State-of-the-art is around 5-10% accuracy
Challenge 1: how we do know what to label?

Look for cues in language (how do people describe images/videos?)

... Why weren't you honest with me? Why did you keep your marriage a secret?

Rick sits down with Lisa.

Oh, it wasn't my secret, Richard. Victor wanted it that way. Not even our closest friends knew about our marriage.

Mining movie scripts
Everingham et al. BMVC
Laptev et al 08.

Ask people on turk for descriptions
Farhadi et al ECCV10
Challenge 1: how we do know what to label?

Look for cues in medical literature on “activities of daily living” (ADLs)

Pirsiwash & Ramanan CVPR12
Challenge 1: how we do know what to label?

Actions vs goal-directed behaviors

Chase vs follow
Challenge 2: how we do obtain interesting data?

Script it, using actors

Use real but “boring” data
Challenge 2: how we do obtain interesting data?

Egocentric/wearable cameras

Pirsivash & Ramanan CVPR12

“Functional” ADLs
Easy to capture variety-rich data
Challenge 2: how we do obtain interesting data?

Egocentric/wearable cameras

Pirsivash & Ramanan CVPR12

“Functional” ADLs

Easy to capture variety-rich data
Challenge 2: how we do obtain interesting data?

Egocentric/wearable cameras
Pirsivash & Ramanan CVPR12

“Functional” ADLs
Easy to capture variety-rich data
Challenge 3: how we do produce detailed annotations?

Crowdsourced labeling

Vondrick et al “VATIC” ECCV10, NIPS11, IJCV 13

Lessons:
- Interface design matters
- Use experts, not the crowd
- Active annotation helps
Roadmap

Data/benchmark analysis

Spatiotemporal features

Spatiotemporal models

Walk  Stand  Crouch
Spacetime features

Simple approach: just use spatial features

Surprisingly (and annoyingly) effective

Build a bank of static-image detectors (of poses, objects, ....)
Exploiting motion

Spatiotemporal interest points (STIPS)

[Laptev 2005]
## XYT descriptor evaluation

<table>
<thead>
<tr>
<th>Descriptors</th>
<th>Harris3D</th>
<th>Cuboids</th>
<th>Hessian</th>
<th>Dense</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOG3D</td>
<td>43.7%</td>
<td>45.7%</td>
<td>41.3%</td>
<td>45.3%</td>
</tr>
<tr>
<td>HOG/HOF</td>
<td>45.2%</td>
<td>46.2%</td>
<td>46.0%</td>
<td><strong>47.4%</strong></td>
</tr>
<tr>
<td>HOG</td>
<td>32.8%</td>
<td>39.4%</td>
<td>36.2%</td>
<td>39.4%</td>
</tr>
<tr>
<td>HOF</td>
<td>43.3%</td>
<td>42.9%</td>
<td>43.0%</td>
<td>45.5%</td>
</tr>
<tr>
<td>Cuboids</td>
<td>-</td>
<td>45.0%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>E-SURF</td>
<td>-</td>
<td>-</td>
<td>38.2%</td>
<td>-</td>
</tr>
</tbody>
</table>

[Wang, Ullah, Kläser, Laptev, Schmid, 2009]
Capturing the “right” temporal motion

Image motion confounds camera translation, object translation, and nonrigid deformations
Capturing the “right” temporal motion

Image motion confounds camera translation, object translation, and nonrigid deformations

Stabilized camera
Stabilized object
Stabilized camera + object
Motion features for detection in videos

Caltech Pedestrian Benchmark; reduce miss rate from 48% to 36%

Park et al CVPR13
Roadmap

Data/benchmark analysis

Spatiotemporal features

Spatiotemporal models

Walk  Stand  Crouch

Motion Features
Tracking
Why do we need to track?

Spacetime window maybe “shearing”
Why do we need to track?

Spacetime window maybe “shearing”
Tracking

Immense literature

\[ S(x,z) = \text{local templates} + \text{spatial relations} + \text{temporal relations} \]

Historically, last term has been focus of tracking community

Given \( z_t \), predict \( z_{t+1} \) with \( P(z_{t+1}|z_t) \)

e.g., particle filtering, Isard & Blake
Extreme form of problem: multi-object tracking
Estimate number of tracks and their extent

Do not assume manual initialization
Estimate birth and death of each track
Tracking by detection

Detect candidates
Link detections over time into tracks
Multi-object tracking as integer/linear programming

View as combinatorial problem of what detections to turn on/off

Jiang et al CVPR07
Zhang et al CVPR08
Berclaz et al PAMI2011
Andriyenko and Schindler ECCV10
Pirsiavash, Ramanan, Fowlkes CVPR11
Butt and Collins CVPR13
Local cost of window
Pairwise cost of transition

Use dynamic programming (DP) to find single track
(e.g., Viterbi algorithm)
Trellis Graph

Shortest path from S to T = best variable-length track

Still can use DP
Min-cost flow problem

(generalization of min-cut / max-flow)

Cost of a K-unit flow = sum of flow along each edge * cost

1) Capacity along each edge is 1
2) Sum of flow into a node = sum of flow out
(ensures non-overlapping tracks)
Exact solution for $K > 1$

Problem: once we instantiate a track, we cannot edit it

Solution: compute shortest path on residual graph augmented with reserve edges

New tracks can "suck flow" out of existing tracks

Keep repeating until next instantiated track increases cost
Okay... so what about tracking articulations?

Which one is correct?
What should a single-image pose estimation alg. output?
N-best decoding

Generate N high-scoring candidates with simple (tree) model, and evaluate with complex model

Popular in speech, but why not vision?
N-best decoding

Generate N high-scoring candidates with simple (tree) model, and evaluate with complex model

Popular in speech, but why not vision?
N-best maximal decoding

N-best with “NMS” or “mode-finding”

Park and Ramanan, ICCV11
Yadollahpour et al. ECCV12
N-best maximal decoding

Intuition: backtrack from all part "max-marginals", not just root
(can we done without any noticeable increase in computation)

Park and Ramanan, ICCV 2011

Pixel locations

head torso leg

• Each part model: all locations.
• "Springs" models part relations.
• Joints: no hard constraint of parts. No initialization of part scores.

1. Initialize nodes with match cost
2. Initialize edges with spring cost
3. Find lowest-cost path from left to right with dynamic programming

If we have n parts and k pixel locations, what is the complexity?

What is complexity when we truncate spring cost (e.g., there are only valid eye offsets for each nose)?

"Secret": In practice, truncation can reduce computation so that local match cost dominates head torso leg

Thursday, August 8, 2013
N-best maximal decoding

Park & Ramanan, “N-best decoders for part models” ICCV 2011

Find N-best “modes” rather than N-best poses

Philosophy: Delay hard decisions as much as possible

Candidate interest points
Candidate parts
Candidate poses
Maximal poses from a single frame

Correct one picked out by temporal context (tracker)
Aside: other ways of representing uncertainty

\[ P(z|x) \propto e^{S(x,z)} \]

Log-linear conditional models
Ramanan NIPS 06
Tracking by articulated detection

Problem: linking up these detections won’t work
Recall: Why is finding people difficult?

- variation in appearance
- variation in pose and viewpoint
- occlusion & clutter

Classic “nuisance factors” in image recognition
Recall: Why is finding people difficult?

- variation in appearance
- variation in pose and viewpoint
- occlusion & clutter

Classic “nuisance factors” in image recognition
Tracking by repeated detection

Generic Person Template

‘Lola’ Template

Thursday, August 8, 2013
Tracking as model-building

A generic object template must be \textit{invariant}

We want to build a model of the object as we track it.
Discriminative clothing models

2002 World Series

motion blur & interlacing
Track long footage (10,000 frames)

Michelle Kwan, 1998 Olympics

extreme pose  motion blur  fast movement

Thursday, August 8, 2013
Olympic woes

silver, not gold
Olympic woes

silver, not gold →

Kwan led after the short program. In the long program, skating to Lyra Angelica by the British composer William Awyn, the 17-year-old turned in a clean, if cautious, effort. Kwan didn't make a major error -- with only one slight wobble on a triple jump -- earning her a solid row of 5.9s on presentation from the judges. As flowers rained upon the ice from her fans, the gold medal, it seemed, was hers. Still, her conservative routine earned five 5.7s for technical merit, and the door was opened, however slight, for Lipinski.

http://espn.go.com/classic/biography/s/Kwan_Michelle.html
The culprit
The culprit

Unexpected/unlikely motions often very important
The motion prior $P(z_{t+1}|z_t)$ may smooth out such subtleties
Tracking multiple people

Independently track each figure

Clothing appearance is no longer a nuisance

Ramanan & Forsyth

CVPR 03
Roadmap

Data/benchmark analysis

Spatiotemporal features

Spatiotemporal models

Walk → Stand → Crouch
Spatiotemporal models

Data-driven

Model-driven

- Run
- Catch
- Trip
- Walk
- Stand
- Sit
Data-driven action recognition

Annotations
{run, walk, wave, etc.}

user

3D motion library

Motion Synthesizer

original video

2D track

annotated video

StandWave
Data-driven action recognition

Annotations
\{run, walk, wave, etc.\}

user
+

3D motion library

Motion Synthesizer

match 1/2 second clips of motion

StandWave

annotated video

original video

2D track

Motion Synthesizer

stand wave

original video

2D track

3D motion library

StandWave

annotated video

original video

2D track

3D motion library

StandWave

annotated video

original video

2D track
Results

(Low-level) tracking

(High-level) spatiotemporal models

Pipeline surprisingly rare (e.g., doesn’t work on TrecVid)
Recognizing structured actions

Making tea from a wearable camera

- Start boiling water
- Do other things (while waiting)
- Pour in cup
- Drink tea

- Start boiling water
- Wait
- Steep tea leaves

Thursday, August 8, 2013
How do we capture long-term structure?

Walk ←→ Stand ←→ Crouch

Markov models
What’s magic behind semi-Markov models?

Walk ↔ Stand ↔ Crouch

Walk
Stand
Crouch

Walk
Stand
Crouch

t1  t2  t3
Add counting states and force sparse transitions (Walk0 to Walk1)
Counting state costs can model arbitrary priors over segment lengths
How do we capture long-term structure?

Exploit models for language

"The hungry rabbit eats quickly"

Context-free grammar
Example grammar

Clean&Jerk action = [red, green, blue]
Snatch action = [red, blue]

\[
S \rightarrow [\text{red}]
\]
\[
S \rightarrow S [\text{red}, \text{green}, \text{blue}, \text{black}]
\]
\[
S \rightarrow S [\text{red}, \text{blue}, \text{black}]
\]
Example parse

Zhu et al

Bobick et al

time
A look back

Data/benchmark analysis

Spatiotemporal features

Spatiotemporal models

Walk  Stand  Crouch