Two action recognition schools

**Actions as spatio-temporal objs**

+ simple
+ reuse lessons from obj categorization
+ robust to hard imaging conditions
  (recently ;)
- not adapted for accurate localization
? multiple people ?
? scalable to many classes ?

e.g. Shuldt et al. 04; Niebles and Fei-Fei 07;
Laptev et al. 07/08; Dollar et. al. 05

**Human-centric**

+ natural representation
+ appearance-invariant
  (need few training examples)
+ focus on person, not background
+ potentially fewer false-positives
+ easy to reason about multiple people
- pose estimation is fragile

e.g. Ramanan and Forsyth 03;
Ikizler and Forsyth 07; Hong et al. 00

Others
Blank et al. 05;
Fathi and Mori 08
Our work

- advance human-centric school
- this talk: automatic pose estimation in unconstrained video
- preliminary pose retrieval results on several hours of *Buffy*
- a step towards human-centric action recognition
**Goal:** detect people and estimate 2D pose in images/video

**Pose**
spatial configuration of body parts

**Estimation**
localize body parts in \((x, y, \theta, s)\)

**Desired scope**
TV shows and feature films

**Body parts**
fixed aspect-ratio rectangles for head, torso, upper and lower arms

e.g. Ramanan and Forsyth, Mori et al.  
Felzenszwalb and Huttenlocher, Sigal and Black
The search space of pictorial structures is large

**Body parts**
- fixed aspect-ratio rectangles for head, torso, upper and lower arms
  
  = 4 parameters each \((x, y, \theta, s)\)

**Search space**
- 4P dim (a scale factor per part)
- 3P+1 dims (a single scale factor)
- \(P = 6\) for upper-body

\[720 \times 405 \text{ image} = 10^{45} \text{ configs!}\]

**Kinematic constraints**
- reduce space to valid body configs \((10^{28})\)
- efficiency by model independencies \((10^{12})\)

---

e.g. Ramanan and Forsyth, Mori et al.
Felzenszwalb and Huttenlocher, Sigal and Black
The challenge of unconstrained imagery

varying illumination and low contrast; moving camera and background; multiple people; scale changes; extensive clutter; any clothing
The challenge of unconstrained imagery

Extremely difficult when knowing nothing about appearance/pose/location
Approach overview

Human detection

Foreground highlighting

Image parsing

Learn appearance models over multiple frames

Repulsive model

reduce search space

estimate pose

every frame independently

integration over multiple frames

Ramanan NIPS 2006

new additions
Single frame
Search space reduction by human detection

**Train**

**Idea**
get approximate location and scale with a
detector generic over pose and appearance

**Building an upper-body detector**
- based on Dalal and Triggs CVPR 2005
- train = 96 frames X 12 perturbations
  (no Buffy)

**Benefits for pose estimation**
+ fixes scale of body parts
+ sets bounds on x,y locations
+ detects also back views
+ fast
- little info about pose (arms)

**Test**

detected enlarged
Search space reduction by human detection

Upper-body detection and temporal association
Search space reduction by foreground highlighting

**Idea**
exploit knowledge about structure of search area to initialize Grabcut

**Initialization**
- learn fg/bg models from regions where person likely present/absent
- clamp central strip to fg
- don’t clamp bg (arms can be anywhere)

**Benefits for pose estimation**
+ further reduce clutter
+ conservative (no loss 95.5% times)
+ needs no knowledge of background
+ allows for moving background
Search space reduction by foreground highlighting

Idea
exploit knowledge about structure of search area to initialize Grabcut

Initialization
- learn fg/bg models from regions where person likely present/absent
- clamp central strip to fg
- don’t clamp bg (arms can be anywhere)

Benefits for pose estimation
+ further reduce clutter
+ conservative (no loss 95.5% times)
+ needs no knowledge of background
+ allows for moving background
Pose estimation by image parsing

**Goal**
estimate posterior of part configuration

\[ P(L \mid I) \propto \exp \left( \sum_i \Phi(l_i) + \sum_{i,j \in E} \Psi(l_i, l_j) \right) \]

\( \Phi = \text{image evidence (given edge/app models)} \)

\( \Psi = \text{spatial prior (kinematic constraints)} \)

**Algorithm**

1. inference with \( \Phi = \text{edges} \)
2. learn appearance models of body parts and background
3. inference with \( \Phi = \text{edges + appearance} \)

**Advantages of space reduction**

+ much more robust
+ much faster (10x-100x)

Ramanan 06
Pose estimation by image parsing

Goal
estimate posterior of part configuration

\[ P(L | I) \propto \exp \left( \sum_i \Phi(l_i) + \sum_{i,j \in E} \Psi(l_i, l_j) \right) \]

\( \Phi = \) image evidence (given edge/app models)
\( \Psi = \) spatial prior (kinematic constraints)

Algorithm
1. inference with \( \Phi = \) edges
2. learn appearance models of body parts and background
3. inference with \( \Phi = \) edges + appearance

Advantages of space reduction
+ much more robust
+ much faster (10x-100x)

Ramanan 06
Failure of direct pose estimation

Ramanan NIPS 2006 unaided
Multiple frames
Transferring appearance models across frames

**Idea**
refine parsing of difficult frames, based on appearance models from confident ones (exploit continuity of appearance)

**Algorithm**
1. select frames with low entropy of $P(L|I)$
2. integrate their appearance models
3. re-parse every frame using integrated appearance models

**Advantages**
+ improve parse in difficult frames
+ better than Ramanan CVPR 2005: integrated models are richer, more robust and generalize to more frames
Transferring appearance models across frames

Idea
refine parsing of difficult frames, based on appearance models from confident ones (exploit continuity of appearance)

Algorithm
1. select frames with low entropy of $P(L|I)$
2. integrate their appearance models
3. re-parse every frame using integrated appearance models

Advantages
+ improve parse in difficult frames
+ better than Ramanan CVPR 2005: integrated models are richer, more robust and generalize to more frames
Transferring appearance models across frames

Idea
refine parsing of difficult frames, based on appearance models from confident ones (exploit continuity of appearance)

Algorithm
1. select frames with low entropy of $P(L|I)$
2. integrate their appearance models
3. re-parse every frame using integrated appearance models

Advantages
+ improve parse in difficult frames
+ better than Ramanan CVPR 2005: integrated models are richer, more robust and generalize to more frames
Transferring appearance models across frames

Idea
refine parsing of difficult frames, based on appearance models from confident ones (exploit continuity of appearance)

Algorithm
1. select frames with low entropy of $P(L|I)$
2. integrate their appearance models
3. re-parse every frame using integrated appearance models

Advantages
+ improve parse in difficult frames
+ better than Ramanan CVPR 2005: integrated models are richer, more robust and generalize to more frames
**The repulsive model**

**Idea**
extend kinematic tree with edges preferring non-overlapping left/right arms

**Model**
- add repulsive edges
- inference with Loopy Belief Propagation
  \[ \Psi = \text{kinematic constraints} \]
  \[ \Lambda = \text{repulsive prior} \]

**Advantage**
+ less double-counting
The repulsive model

Idea
extend kinematic tree with edges
preferring non-overlapping left/right arms

Model
- add repulsive edges
- inference with Loopy Belief Propagation

\( \Psi \) = kinematic constraints
\( \Lambda \) = repulsive prior

Advantage
+ less double-counting
Full-body pose estimation in easier conditions

Weizmann action dataset (Blank et at. ICCV 05)
Upper-body pose estimation in TV shows

- >70000 frames over 4 episodes of *Buffy the Vampire Slayer*  
  (>1000 shots)

- uncontrolled and extremely challenging
  
  low contrast, scale changes, moving camera and background,  
  extensive clutter, any clothing, any pose

- figure overlays = after transferring appearance models
Example estimated poses
Quantitative evaluation

*Ground-truth*
69 shots × 4 = 276 frames × 6 = 1656 body parts (sticks)

*Upper-body detector*
fires on 243 frames (88%, 1458 body parts)
Quantitative evaluation

**Pose estimation accuracy**

- 62.6% with repulsive model
- 59.4% with appearance transfer
- 57.9% with foreground highlighting (*best single-frame*)
- 41.2% with detection
- 9.6% Ramanan NIPS 2006 unaided

**Conclusions:**

- Both reduction techniques improve results
- Small improvement by appearance transfer ...
- Method good also on static images
- Repulsive model brings us further
Example video
Pose retrieval: task

**Task**

Given user-selected query frame+person ...

... retrieve shots with persons in the same pose from video database (in experiments: 4 episodes = 3 hours)
Pose retrieval: method

**Pose descriptors**
- soft-segmentations of body parts
- distributions over orient+location for parts and pairs of parts

**Similarity measures**
- dot-product (= soft intersection)
- Batthacharryya / Chi-square
What is missed?

- too small
- out of image
- low contrast
- confused by other people
- missed detections
- incorrect pose estimates

PR AUC = 0.4
The road ahead

*Further improve pose estimation*
- include background model (explain every pixel)
- multi-people reasoning (e.g. occlusion modeling)
- simultaneous spatio-temporal pose estimation

*Explore better pose descriptors*
- integrate over a small temporal neighborhood
- robustness to missed/wrong parts
- learning from a few examples

*Evolve from pose retrieval to action recognition*
Discussion

Back to schools

- human-centric: how robust can it get?
  can we do ‘hugging’ explicitly?

- human-centric: how high is the price of higher complexity?

- actions=objects: scale up to many action classes?
  at which training price?

- hybrid approaches are a promising future?
  e.g. start from actions=objects, then verify with human-centric
Questions?

(Video!)

www.robots.ox.ac.uk/~vgg

- ground-truth annotated stickmen
- upper-body detection and tracking software
- ground-truth time intervals labeled by pose class (soon ;)

(Video!)