Discriminative Methods

Greg Mori
CMPT 888
ACTION RECOGNITION USING MID-LEVEL MOTION FEATURES
Spatial Motion Descriptor

Image frame

Optical flow

$F_x^-, F_x^+, F_y^-, F_y^+$

blurred

$F_x^-, F_x^+, F_y^-, F_y^+$
Spatio-temporal Motion Descriptor

Temporal window $w$

\[ \sum \]

Sequence A

Sequence B

frame-to-frame similarity matrix

motion-to-motion similarity matrix
Template Matching

• Efros et al. template matching is very slow
  – Could use fast nearest neighbour search
    algorithm / data structure

• Distance measure doesn’t focus on important
  parts of motion

• Instead, discriminative learning to figure out
  which pieces of the motion are most
  important
Which Motions are Important?

- Use AdaBoost algorithm (Viola and Jones IJCV 2004)
1) Calculate Optical Flow

2) Normalize / remove noise

\[ v'_{i,j} = \frac{v_{i,j}}{\|v\| + \epsilon}, \quad \epsilon = 0.5 \]

3) Separate components

4) Blur components

5) Compute zero component

\[ F_0 = F_{x+} + F_{x-} + F_{y+} + F_{y-} \]

Similar to Efros et al. ICCV03
Template matching too slow
$h_t(v, l) = \begin{cases} 
1 & p_{t,l}v_{\tau(t)} > p_{t,l} \theta_t \\
0 & \text{otherwise} 
\end{cases}$

$H(v, l) = \sum_{t=1}^{N} \alpha_t h_t(v, l)$

N = 1500 weak classifiers
Mid-level Features

• Two-level learning with AdaBoost
  – First train classifiers in spatio-temporal cuboids
  – Use the output of these classifiers in another round of AdaBoost
• Non-linearity in between layers, so not equivalent to single run
• Comparison to single-run lacking in this paper
  – (Unpublished) experiments indicate it performs better empirically
    • Theoretical arguments missing, though intuition about squeezing more discrimination out of each subwindow, non-linearity providing larger hypothesis space for classifier
Learned Motion Features

Learned classifier outperforms nearest-neighbour on standard datasets in accuracy and efficiency
Example – Real-time Gesture Recognition
Wang and Mori NIPS 2008, CVPR 2009

MAX-MARGIN HIDDEN CONDITIONAL RANDOM FIELDS FOR HUMAN ACTION RECOGNITION
Previous Work

Large-scale feature
[e.g. Efros, Berg, Mori, Malik, ICCV03]

Local patches
[e.g. Laptev & Perez, ICCV07]
Large vs. Small Scale Features

Challenge: How to combine in a principled manner?
Hidden Conditional Random Field

\[ p(y, h | x) \propto \exp(\Psi(y, h, x)) \]
Learning hCRF Parameters

- **Conditional likelihood**
  - Integrate out latent part labels $h$

- **Max-margin**
  - Examine best setting for latent part labels $h$
  - Latent-SVM (Felzenszwalb et al. CVPR08), MI-SVM (Andrews et al. NIPS03)
Conditional Likelihood

- Choose parameters to make likelihood on ground-truth labels as large as possible

\[
\ell = \sum_t \log p(y^t | x^t) = \sum_t \log \left( \sum_h p(y^t, h | x^t) \right)
\]
Max-Margin

Choose parameters to make score on ground-truth label higher than any competing label

\[
\max_{h} p(Y = y^t, h \mid x^t) > \max_{h} p(Y \neq y^t, h \mid x^t)
\]
Finding Parts

learn a root model

learn final model
Experiments: Weizmann dataset

- Benchmark dataset
  - 9 actions
  - 9 subjects

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (MM-hCRF)</td>
<td>100%</td>
</tr>
<tr>
<td>Ours (CL-hCRF)</td>
<td>97.2%</td>
</tr>
<tr>
<td>Jhuang &amp; Poggio ICCV07</td>
<td>98.8%</td>
</tr>
<tr>
<td>Niebles &amp; Fei-Fei BMVC06</td>
<td>72.8%</td>
</tr>
</tbody>
</table>
Inferred Part Labels
Visualization of Learned Model
## Conditional Likelihood vs. Max-Margin

### Weizmann dataset

| Method     | $|H| = 6$ | $|H| = 10$ | $|H| = 20$ |
|------------|---------|---------|---------|
| hCRF-CL    | 91.7    | 97.2    | 94.4    |
| hCRF-MM    | 97.2    | 100     | 97.2    |

### KTH dataset

| Method     | $|H| = 6$ | $|H| = 10$ | $|H| = 20$ |
|------------|---------|---------|---------|
| hCRF-CL    | 78.5    | 87.6    | 75.1    |
| hCRF-MM    | 84.8    | 92.5    | 89.7    |

CL: \[
\log \sum_{h} p(Y = y^t, h|x^t) \text{ vs. } \log \sum_{h} p(Y \neq y^t, h|x^t)\]

MM: \[
\max_{h} p(Y = y^t, h|x^t) > \max_{h} p(Y \neq y^t, h|x^t)\]
RECOGNIZING HUMAN ACTIONS FROM STILL IMAGES WITH LATENT POSES

Yang, Wang, and Mori CVPR 2010
Goal

- Action recognition from still images
  - News/sports image retrieval and analysis
  - An important cue for video-based action recognition
Previous work

• Global template-based representation
e.g. Wang et al. CVPR06, Ikizler-Cinbis et al. ICCV09

• Pose estimation + action recognition
e.g. Ramanan and Forsyth NIPS03, Ferrari et al. CVPR09
Discriminative Pose

- Not all elements of pose are equally important
- Develop integrated learning framework to estimate pose for action recognition
Pose Representation

- We use a coarse non-parametric pose representation
  - An action-specific variant of the *poselet* [Bourdev & Malik ICCV09]

- A *poselet* is a set of patches not only with similar pose configuration, but also from the same action class.
• Poselets obtained by clustering ground-truth joint positions of body parts for each action
Model Formulation

• Develop a scoring function $H(I, Y; \Theta)$
  – Should have high score for correct action label $Y$
  – Low score for other action labels
  – Model parameters $\Theta$
Model Formulation

\[ H(I, Y; \Theta) = \max_L \Theta^T \Psi(I, L, Y) \]
Model Formulation

Action Label

Pose

Image

Large score for $H(I, Y = \text{Running}; \Theta)$
Model Formulation

Action Label

Pose

Image

Small score for $H(I, Y = Sitting; \Theta)$
Model Details I

Action Label

Pose

Image

Relative body part locations
Model Details II

- **Action Label**: $Y$
- **Pose**: Image
- **Poselet matching**: $I$
Model Details III

Action Label

Pose

Image

Canonical poses for an action
Full Model

Model parameters learned using max-margin
Experiments

- Still image action dataset
  - Five action categories
  - 2458 images total
  - Train using 1/3 of images from each category

Baseline – HOG/SVM: 52% per class accuracy

Ours – Latent Pose: 62% per class accuracy
Visualization of latent pose

Successful classification examples

Unsuccessful classification examples