Human Motion Categorization & Detection

Juan Carlos Niebles$^{1,2}$

Advisor : Prof. Fei-Fei Li$^1$
Juan Carlos Niebles$^{1,2}$
In this talk

- Where are the humans?
  - Which pixels?
- What motions are they doing?
  - Learn models and classify novel sequences
In this talk

- Where are the humans?
  - Which pixels?
- What motions are they doing?
  - Learn models and classify novel sequences
Identification of human actions in video

Challenges:
- Camera Motion
- Complex Background
- Viewpoint Change
Localizing human actions in long sequences

Multiple actions by different people in one single sequence

Camera Motion

Dynamic Background
Previous Work

1. Recognition by template correlation/matching
   - Efros et al, ICCV ‘03
   - Shechtman et al, CVPR ‘05
   - Ramanan et al, NIPS ‘03

2. Action analysis using graphical models
   - Song et al, PAMI ‘03
   - Fanti et al, ICCV ‘05
   - Boiman et al, ICCV ‘05
   - Bergler, CVPR ‘97

3. Spatial-temporal Interesting Points
   - Laptev et al, ICCV ‘03; Schuldt et al, ICPR ‘04
   - Dollar et al, ICCV VS-PETS ‘05
   - Ke et al, ICCV ‘05
Motivation

Sparse and Local representation

Spatio-temporal information

Johansson, 1973
Spatio-Temporal Interest Points

Detection

\[ R = (I * g_x * g_y * h_{ev})^2 + (I * g_x * g_y * h_{od})^2 \]

Local optimum of \( R \) define the position of features

[Dollar et al '05]
Spatial-Temporal Interest Points

- walking
- running
- jogging
- boxing
- handclapping
- handwaving
Spatial-Temporal Interest Points

Detection

\[ R = (I \ast g_x \ast g_y \ast h_{ev})^2 + (I \ast g_x \ast g_y \ast h_{od})^2 \]

Local optimum of \( R \) define the position of features

Description

Spatial-temporal cube

Concatenated brightness gradient

Experiments showed that optical flow descriptor is equally effective

[Dollar et al ’05]
Object \rightarrow \text{Bag of ‘words’}
Codebook and Representation
Codebook and Representation

[Image of a diagram depicting training videos mapped to codewords in a feature space, with an input video and codewords on the right side.]
Experiment I: Codebook

- walking
- running
- jogging
- handwaving
- handclapping
- boxing
Feature extraction and description

Learning

Feature extraction

Codebook

Video representation

Learn models

Model 1

Model N

Recognition

New Input

Input

Feature extraction

Class 1

Class N

Class 1

Class N

Feature extraction and description

Class 1

Class N

Learn

Decide on best model
Unsupervised learning using pLSA

Input

“camel spin”

Action category

Spatial-temporal word

N

W_d

d → z → w

Colgate
pLSA Model

\[ p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k) p(z_k \mid d_j) \]

**action category vectors**

Word distribution per action category

**action category weights**

Action category distribution per video
pLSA Model

\[ p(w_i \mid d_j) = \sum_{k=1}^{K} p(w_i \mid z_k) p(z_k \mid d_j) \]

Unsupervised Learning

\[ L = \prod_{i=1}^{M} \prod_{j=1}^{N} p(w_i \mid d_j)^{n(w_i,d_j)} \]
Feature extraction and description

Input

Feature extraction

Codebook

Video representation

Class 1

Class N

Class 1

Class N

Learn models

Model 1

Model N

Learn models

Decide on best model

Recognition

New Input
 Recognition

\[
P(w | d_{\text{test}}) = \sum_{k=1}^{K} P(z_k | d_{\text{test}}) P(w | z_k)
\]

action category = \text{arg max}_{k} P(z_k | d_{\text{test}})
Experiment I:

KTH dataset
[Schuldt et al., 2004]:

Walking  Boxing  Hand Waving  Running  Jogging  Hand Clapping

25 persons, indoors and outdoors, 4 long sequences per person
### Experiment I: Performance

- Leave-one person out cross validation
- Average performance: 81.50%

<table>
<thead>
<tr>
<th>Motion</th>
<th>Our Method</th>
<th>Dollar et al.</th>
<th>Schuldt et al.</th>
<th>Ke et al.</th>
</tr>
</thead>
<tbody>
<tr>
<td>walking</td>
<td>.79</td>
<td>.81</td>
<td>.88</td>
<td>.93</td>
</tr>
<tr>
<td>running</td>
<td>.01</td>
<td>.36</td>
<td>.52</td>
<td>.77</td>
</tr>
<tr>
<td>jogging</td>
<td>.11</td>
<td>.52</td>
<td>.77</td>
<td>1.00</td>
</tr>
<tr>
<td>handwaving</td>
<td>.00</td>
<td>.00</td>
<td>.93</td>
<td>.01</td>
</tr>
<tr>
<td>handclapping</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.23</td>
</tr>
<tr>
<td>boxing</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
</tr>
</tbody>
</table>

- Unsupervised training
- Handle multiple motions
Experiment I: Caltech dataset

Trained with the KTH data

Tested with the Caltech dataset

Only words from the corresponding action are shown
Experiment I: A longer sequence

- walking
- running

Trained with the KTH data
Tested with our own data
Experiment I: Multiple motions

- handclapping
- handwaving

Trained with the KTH data
Tested with our own data
Experiment II:

Figure Skating data set:
[Y.Wang, G.Mori et al, CVPR 2006]

7 persons, 3 action classes: camel spin, stand spin, sit spin
Experiment II: Examples

Figure skating actions

Camel spin  Sit spin  Stand spin
Experiment II: Long Sequences
references


• with
  – Hongcheng Wang
but...

- We need models that exploit geometrical arrangements of features
in the object recognition world

• all have same P under bag-of-words model.
• part-based models that capture geometrical information
  – Constellation model
  – Pictorial structures
  – etc
**Constellation Model**

- **P1**, **P2**, **P3**, **P4**
- Small number of features
- Strong shape representation
- No geometrical or shape information

**bags of features**

- Large number of features
- No geometrical or shape information
Constellation of bags of features

- Large number of features
- Strong shape representation

Diagram:
- Nodes P1, P2, P3, P4 connected by lines, each connected to a layer labeled Bg, and an image layer labeled \( w \).
**Constellation of bags of features**

- $P_p$: Parts
- $w$: Observed patches (shape, motion and position)

- P’s are *similar* to the parts in Constellation model
- Each Part is the parent of a “bag-of-features”

😊 Large number of features
😊 Strong shape representation
Constellation Model

Constellation of bags of features
Constellation of bags of features
Constellation of bags of features

- Use a mixture to account for data multimodality.
learned models
Constellation of bags of features

Part layer

Feature layer

Image

$\theta_{\omega}$

$P_1$

$P_2$

$P_3$

$P_4$

$B_g$

$\theta$

$w$
Mixture model

Single component

bag of words

Classification performance

Model comparison

<table>
<thead>
<tr>
<th>Model Type</th>
<th>Classification Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixture+ Motion+ Static</td>
<td>70%</td>
</tr>
<tr>
<td>NoMixture+ Motion+ Static</td>
<td>65%</td>
</tr>
<tr>
<td>Bag+ Motion+ Static</td>
<td>45%</td>
</tr>
</tbody>
</table>
Constellation of bags of features

Part layer

Feature layer

Image

θω

P1

P2

P3

P4

Bg

w
Constellation of bags of features

- Static features
  Edge map + Shape Context
  [Belongie ’03]

- Motion features
  Interest Points + ST-gradients
  [Dollar ’05]

Representation

video frame: \( w = \{x, a\} \)
\( w_i = \{x_i, a_i\} \)
\( x_i: i\)-th feature position
\( a_i: i\)-th feature appearance
Constellation of bags of features

- **Static features**
  - Edge map + Shape Context
    - [Belongie ’03]

- **Motion features**
  - Interest Points + ST-gradients
    - [Dollar ’05]
## Constellation of bags of features

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w$</td>
<td>shape, motion and location features</td>
</tr>
<tr>
<td>$Y$</td>
<td>set of possible locations for the $P$ parts (a discretized grid over the image)</td>
</tr>
<tr>
<td>$h$</td>
<td>index variable to select a specific position of the parts among all the possibilities (in the grid)</td>
</tr>
<tr>
<td>$P$</td>
<td>chooses a specific assignment of features $w$ to parts $P$</td>
</tr>
</tbody>
</table>

$$p(w, Y | \theta) \approx$$

$$\sum_{\omega=1}^{\Omega} \sum_{h \in H} \pi_\omega p(h|\theta_\omega) p(Y|h, \theta_\omega) p(w|Y, m^*, h, \theta_\omega)$$

- **Part layer**
- **Local feature layer**
**Constellation of bags of features**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(w)</td>
<td>shape, motion and location features</td>
</tr>
<tr>
<td>(Y)</td>
<td>set of possible locations for the (P) parts (a discretized grid over the image)</td>
</tr>
<tr>
<td>(h)</td>
<td>index variable to select a specific position of the parts among all the possibilities (in the grid)</td>
</tr>
<tr>
<td>(P)</td>
<td></td>
</tr>
<tr>
<td>(m)</td>
<td>chooses an specific assignment of features (w) to parts (P)</td>
</tr>
</tbody>
</table>

**Part layer term: (constellation)**

\[
p(Y \mid h, \theta_\omega) = \mathcal{N}(Y_T(h) \mid \mu_{L, \omega}, \Sigma_{L, \omega})
\]
Constellation of bags of features

<table>
<thead>
<tr>
<th>( w )</th>
<th>shape, motion and location features</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>set of possible locations for the ( P ) parts (a discretized grid over the image)</td>
</tr>
<tr>
<td>( h )</td>
<td>index variable to select a specific position of the parts among all the possibilities (in the grid)</td>
</tr>
<tr>
<td>( m )</td>
<td>chooses an specific assignment of features ( w ) to parts ( P )</td>
</tr>
</tbody>
</table>

Local feature layer term:

\[
p(w | Y, m^*, h, \theta_\omega) = \prod_{w_j \in Bg} p(x_j^r | \theta_0^X) p(a_j | \theta_0^A) \prod_{p=1}^P \prod_{w_i \in P_p} p(x_i^r | Y, h_p, \theta_p^X) p(a_i | \theta_p^A)
\]
Learning

Constellation shape: Gaussian
\[ \mu_{L,\omega}, \Sigma_{L,\omega} \]

Part shape: Zero mean Gaussian
\[ \sum_{P,\omega}^X \]

Part appearance: Multinomial
\[ \theta_{P,\omega}^A \]

Parameter learning with
\[ \mathbf{EM} = \left\{ \mu_{L,\omega}, \Sigma_{L,\omega}, \sum_{P,\omega}^X, \theta_{P,\omega}^A, \theta_0^X, \theta_0^A \right\} \]
\[ p = 1 \ldots P \]
\[ \omega = 1 \ldots \Omega \]
“Actions as Space-Time Shapes”
IEEE International Conference on Computer Vision (ICCV), Beijing, October 2005.
**experimental results**

- 9 action classes, performed by 9 subjects [Blank et al 2005]
- Leave one out cross-validation
- Video Classification performance: 72.8%

```
<table>
<thead>
<tr>
<th></th>
<th>bend</th>
<th>pjump</th>
<th>jack</th>
<th>wave1</th>
<th>wave2</th>
<th>jump</th>
<th>run</th>
<th>side</th>
<th>walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>bend</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>pjump</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>jack</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>wave1</td>
<td>0.22</td>
<td>0.11</td>
<td>0.11</td>
<td>0.44</td>
<td>0.11</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>wave2</td>
<td>0.00</td>
<td>0.11</td>
<td>0.22</td>
<td>0.67</td>
<td>0.00</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>jump</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.78</td>
<td>0.0</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>run</td>
<td>0.00</td>
<td>0.11</td>
<td>0.00</td>
<td>0.00</td>
<td>0.11</td>
<td>0.56</td>
<td>0.11</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>side</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.33</td>
<td>0.11</td>
<td>0.56</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>walk</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.11</td>
<td>0.00</td>
<td>0.33</td>
<td>0.56</td>
<td>0.00</td>
</tr>
</tbody>
</table>
```
Results

Predicted frame label: wave2(3)  Ground truth frame label: wave2

Predicted sequence label: bend, jack, jump, pjump, run, side, walk, wave1, wave2

Correct
Incorrect

Accumulated votes
Results

Predicted frame label

Ground truth frame label

Predicted sequence label

Accumulated votes

Correct

Incorrect
references

- Juan Carlos Niebles and Li Fei-Fei. A Hierarchical Model of Shape and Appearance for Human Action Classification. *CVPR*, 2007
In this talk

• Where are the humans?
  – Which pixels?

• What motions are they doing?
  – Learn models and classify novel sequences
• detect moving people in YouTube videos
• extract the spatio-temporal volume that contains each person
Fast and robust algorithms for face detection
real-world sequences

- Compression artifacts & low quality
- Videos contain multiple shots
- Unknown number of humans
- Arbitrary human motion and poses
- Unknown camera parameters and motion
- Background clutter, motion and occlusions

Balan et al ’07

Wu&Nevatia ’07

…

Wu&Nevatia ’07

…

extraction detail
System Overview

Input: Clustering of people

Extract: Spatio-temporal volume

Output:
pedestrian detection

- Upright human detection is somewhat successful
  - Dalal & Triggs 05, Sabzmeydani & Mori 07, Laptev 06, and more
clustering ped detections
clustering ped detections

- **Must-link**: detections with similar appearance and consistent spatial locations

- **Cannot-link**: detections that occur within the same frame

Ø Clustering with constraints, Klein et al ‘02
First clustering step

- Clustering with constraints [Klein et al ‘02]
- Use simple & cheap descriptors: global color histogram
- Gives ‘over-segmented’ clusters
  Conservative clustering threshold
Second clustering step

- Clustering with constraints [Klein et al '02]
- More expensive descriptor: head & torso estimates
- Reject tracks Poor head torso estimates
Detection & clustering result

Output = One cluster per person in the sequence
System Overview

input

people clustering

extract spatio-temporal volume

output
extracting the human region in a single frame [Ramanan 06]

• Pictorial structure representation
  – each part represents local visual properties
  – ‘springs’ capture spatial relationships
  – stretch & fit to find right configuration

images in this slide from: Deva Ramanan
• N-part model
• A configuration is given by
  \[ L = \{ l_i \}; \ i = 1 \ldots N, \ \text{with} \]
  \[ l_i = \{ x_i, y_i, \theta_i \} \]

• We are interested in
  \[ p(L | I, \theta) \propto p(I | L, \theta)p(L | \theta) \]
  \[ P(L | I, \theta) \propto \left( \prod_{i=1}^{\text{Image evidence}} p(I | l_i, u_i) \right) \prod_{(v_i, v_j) \in E} p(l_i, l_j | c_{ij}) \]

• efficient inference
• measurement is the bottle-neck
  expensive convolutions with non-separable filters
reducing measurement computation

• Benefits
  – can be limited to a smaller space, reducing computation time
  – avoid distracting background observations
  – final estimation accuracy can be improved

• Use temporal consistency
  – Compute full model at first frame
  – Propagate part estimations to next frame

input image  original search space for torso (full frame)  reduced search space for torso
GMM approximation is done via Kernel Density Approximation (KDA) [Han et al. 'PAMI08]

- Densities are propagated in a Bayesian filtering framework

\[
p(X_t | Z_{1:t}) \propto p(Z_t | X_t)p(X_t | Z_{1:t-1}) = \left( \sum_{i=1}^{N_1} \mathcal{N}(k_i, x_i, P_i) \right) \left( \sum_{j=1}^{N_2} \mathcal{N}(\tau_j, y_j, Q_j) \right)
\]
extracting the human region

Projecting the part posteriors into the image gives a rough segmentation of the body region
results

frame 1  frame 7  frame 18  frame 24
results

frame 1
frame 10
frame 60
frame 136
results

frame 11  frame 74  frame 86  frame 300
experiments

Precision Recall Comparison

<table>
<thead>
<tr>
<th>Rate</th>
<th>Detection only</th>
<th></th>
<th>Detection &amp; Clustering</th>
<th></th>
<th>Full model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Prec</td>
<td>Rec</td>
<td>F</td>
<td>Prec</td>
<td>Rec</td>
<td>F</td>
</tr>
<tr>
<td>0.89</td>
<td>0.90</td>
<td>0.92</td>
<td>0.93</td>
<td>0.94</td>
<td>0.89</td>
<td>0.31</td>
</tr>
<tr>
<td>0.31</td>
<td>0.25</td>
<td>0.19</td>
<td>0.16</td>
<td>0.13</td>
<td>0.31</td>
<td>0.39</td>
</tr>
<tr>
<td>0.46</td>
<td>0.39</td>
<td>0.32</td>
<td>0.27</td>
<td>0.24</td>
<td>0.46</td>
<td>0.39</td>
</tr>
<tr>
<td>0.89</td>
<td>0.90</td>
<td>0.92</td>
<td>0.93</td>
<td>0.94</td>
<td>0.89</td>
<td>0.31</td>
</tr>
<tr>
<td>0.90</td>
<td>0.92</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>0.90</td>
<td>0.31</td>
</tr>
<tr>
<td>0.92</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.90</td>
<td>0.31</td>
</tr>
<tr>
<td>0.92</td>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.90</td>
<td>0.31</td>
</tr>
<tr>
<td>0.93</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.90</td>
<td>0.31</td>
</tr>
<tr>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.94</td>
<td>0.90</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Computation Time

~ 1 order of magnitude speed up
more results
references

• Juan Carlos Niebles, Bohyung Han, Andras Ferencz and Li Fei-Fei. Extracting Moving People from Internet Videos. **ECCV** 2008

• With:

  Bohyung Han  Andras Ferencz
Conclusions

• Spatio-temporal & spatial features + statistical classifier
  – Bag of words (unsupervised)
  – Constellation of bags of features

• Human Motion extraction
  – Real world sequences from YouTube
Interesting Research Issues

• Motion taxonomy/vocabulary
• higher level activities
• ...
In this talk

• Where are the humans?
  – Which pixels?

• What motions are they doing?
  – Learn models and classify novel sequences
Thank you