Activity recognition in videos

Christian Wolf
A.P. Ta, G. Taylor, G. Lavoué, A. Baskurt, J.M. Jolion

Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69622, France
http://liris.cnrs.fr/christian.wolf

April 30th, 2010
What happens in a video?

- On an individual level?
- Group behavior?

Applications

- Video surveillance
- Robotics
- Video indexing and macro segmentation
- Intelligent VCR
- Human-computer interaction
Keypoints
Spatial (2D)
- Broadcast video: Recurrent networks
  - Baccouche, Mamalet, Wolf and Baskurt
  - Work in progress

Keypoints
Spatio–t. (3D)
- Bags of Pairwise features
  - Ta, Wolf, Lavoue, Baskurt and Jolion
  - ICPR 2010

Segment.
BG subtraction
- Learning activities from binary shape sequences:
  - Condit. Boltzmann machines
  - Wolf, Taylor and Jolion
  - Work in progress

Trajectory clustering
- Pop Scuturici and Miguet
  - ICPR 2008

Hypergraph matching
- Ta, Wolf, Lavoue and Baskurt
  - AVSS 2010

C. Wolf — http://liris.cnrs.fr/christian.wolf
Activity recognition in videos 3 / 37
### Graph matching

<table>
<thead>
<tr>
<th>Keypoints</th>
<th>Broadcast video: Recurrent networks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial (2D)</td>
<td>Baccouche, Mamalet, Wolf and Baskurt</td>
</tr>
<tr>
<td></td>
<td>Work in progress</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Keypoints</th>
<th>Bags of Pairwise features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatio–t. (3D)</td>
<td>Ta, Wolf, Lavoue, Baskurt and Jolion</td>
</tr>
<tr>
<td></td>
<td>ICPR 2010</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Segment. BG subtraction</th>
<th>Learning activities from binary shape sequences: Condit. Boltzmann machines</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wolf, Taylor and Jolion</td>
</tr>
<tr>
<td></td>
<td>Work in progress</td>
</tr>
</tbody>
</table>

- Work in progress
- Hypergraph matching
  - Ta, Wolf, Lavoue and Baskurt
  - AVSS 2010
- Condit. Boltzmann machines
  - Wolf, Taylor and Jolion
  - Work in progress
- Learning activities
  - Condit. Boltzmann machines
  - Wolf, Taylor and Jolion
  - Work in progress

- Broadcast video: Recurrent networks
  - Baccouche, Mamalet, Wolf and Baskurt
  - Work in progress

- Bags of Pairwise features
  - Ta, Wolf, Lavoue, Baskurt and Jolion
  - ICPR 2010

- Hypergraph matching
  - Ta, Wolf, Lavoue and Baskurt
  - AVSS 2010

- Learning activities from binary shape sequences: Condit. Boltzmann machines
  - Wolf, Taylor and Jolion
  - Work in progress
Space-time interest points

Points which change in space and in time
Motivation

- Classify individual human actions (per human)
- BG-subtraction may fail and is difficult when the camera moves
- Exploit the robustness of space-time interest points
- Match the ST-points of several “model” frames against the ST-points of a new sequence

Authors

A.-P. Ta, C. Wolf, G. Lavoué, A. Baskurt
Graph matching for video classification

Hyper-edges (triangles) are created from close ST-points.

The points in a triangle are ordered.

Matching takes into account features on the ST-points as well as the ST-geometry of the triangles.
Hyper-graph matching [Duchenne et al., 2009]
## Results on entire sequences

<table>
<thead>
<tr>
<th>Method</th>
<th>KTH</th>
<th>Weizmann</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Our method</strong></td>
<td>91.2</td>
<td>100.0</td>
</tr>
<tr>
<td>[Schindler and van Gool, 2008]</td>
<td>92.7</td>
<td>100.0</td>
</tr>
<tr>
<td>[Fathi and Mori, 2008a]</td>
<td>90.5</td>
<td>100.0</td>
</tr>
<tr>
<td>[Gorelick et al., 2007a]</td>
<td>-</td>
<td>99.6</td>
</tr>
<tr>
<td>[Sun et al., 2009a]</td>
<td>94.0</td>
<td>97.8</td>
</tr>
<tr>
<td>[Niebles et al., ]</td>
<td>83.3</td>
<td>90.0</td>
</tr>
<tr>
<td>[Kim and Cipolla, 2009]</td>
<td>95.3</td>
<td>-</td>
</tr>
<tr>
<td>[Liu and Shah, 2008]</td>
<td>94.2</td>
<td>-</td>
</tr>
<tr>
<td>[Ryoo and Aggarwal, 2009]</td>
<td>93.8</td>
<td>-</td>
</tr>
<tr>
<td>[Zhang et al., 2008]</td>
<td>91.3</td>
<td>-</td>
</tr>
<tr>
<td>[Ballan et al., 2009]</td>
<td>92.1</td>
<td>92.4</td>
</tr>
<tr>
<td>[Gilbert et al., 2008]</td>
<td>89.9</td>
<td>-</td>
</tr>
<tr>
<td>[Savarese et al., 2008]</td>
<td>86.8</td>
<td>-</td>
</tr>
<tr>
<td>[Dollar et al., 2005]</td>
<td>81.2</td>
<td>-</td>
</tr>
</tbody>
</table>
Results on individual people (1) — consecutive frames
Results on individual people (2) — consecutive frames
Learning shape sequences

Keypoints
- Spatial (2D)
  - Broadcast video: Recurrent networks
    - Baccouche, Mamalet, Wolf and Baskurt
    - Work in progress

Keypoints
- Spatio–t. (3D)
  - Bags of Pairwise features
    - Ta, Wolf, Lavoue, Baskurt and Jolion
    - ICPR 2010
  - Hypergraph matching
    - Ta, Wolf, Lavoue and Baskurt
    - AVSS 2010

Segment.
- BG subtraction

Learning activities
- from binary shape sequences:
  - Condit. Boltzmann machines
    - Wolf, Taylor and Jolion
    - Work in progress
Video classification with binary shape sequences

Authors
Christian Wolf, Graham Taylor, Jean-Michel Jolion

- Segmentation of the scene into moving objects
- Classification decisions are taken on a per object basis

⇒ We learn the evolution of binary shapes over time

Two steps:
- Segmentation (background subtraction)
- Modeling motion through binary shape sequences
A motion model of binary shape sequences

- Extract robust shape features for each shape
- Model the temporal dependencies

Decomposition into complex Zernike moments
Restricted Boltzmann machine (RBM)

- A model for static data (*NOT* sequences!)
- Each observed variable is connected to each hidden variable and vice versa
- All hidden variables are binary labeled
Restricted Boltzmann machine (RBM)

Autoencoder: iterative learning algorithm

- The observed variables are updated from the hidden $v$.
- The hidden variables are updated from the observed $v$.

The case of binary observations:

\[
p(v_i = 1 | h) = \frac{1}{1 + \exp(-a_i - \sum_j W_{ij} h_j)}
\]

\[
p(h_j = 1 | v) = \frac{1}{1 + \exp(-b_i - \sum_i W_{ij} v_i)}
\]

The case of real valued observations:

\[
p(v_i | h) = \mathcal{N} \left( a_i + \sigma_i \sum_j W_{ij} h_j \sigma_i^2 \right)
\]
Dynamic data: Conditional RBMs (CRBM)

- [Taylor et al., NIPS 2007]
- Mixture between directed and undirected connections
- Connections between the past and the present: dynamic bias
- No change of the inference algorithm
Conditional deep belief network (CDBN)

- Deep belief network: [Hinton et al., Science, 2006]
- Learning of high level features
- The hidden layer $H^0$ is considered as observed for hidden layer $H^1$
Generated sample data

- Input data: Zernike moments up to order 40
- Sample data generated from 5 input frames
- DBN; layers 1&2: order 2, 150 hidden variables
- Input was NOT part of the training data

The model is capable of modeling the dynamics of 2D shape rather than simply storing sequences of repeating frames.
The CRBM model serves as a mid- and high-level feature extractor

“Intelligent” data reduction taking into account the dynamics of the data

Outperforms PCA by a large margin

Possibility to “validate” the model by generating data

Feature extraction is fast: a single bottom-up pass

⇒ A discriminative step is needed for classification

⇒ SVM !!
Overview

Test

- 7 consecutive frames
- Low level feature extr.
- Zernike moments & Co.
- Calculate posterior of hidden variables
- High level features
- SVM Classification
- Activity

Training

- CRBM Training (Contrastive divergence)
- SVM Training
- CRBM Model
- SVM Model

C. Wolf — http://liris.cnrs.fr/christian.wolf

Activity recognition in videos
BG subtraction: Weizmann database
BG subtraction: ANR-Canada database

- Run
- Sitdown
- Jump vert.
Recognition results: ANR-Canada database
Weizmann dataset: entire sequences

Classification accuracy on entire sequences: comparison with established methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed method [Schindler and van Gool, 2008]</td>
<td>100.0</td>
</tr>
<tr>
<td>[Fathi and Mori, 2008b]</td>
<td>100.0</td>
</tr>
<tr>
<td>[Gorelick et al., 2007b]</td>
<td>99.6</td>
</tr>
<tr>
<td>[Sun et al., 2009b]</td>
<td>97.8</td>
</tr>
<tr>
<td>[Bregonzio et al., 2009]</td>
<td>96.7</td>
</tr>
<tr>
<td>[Ballan et al., 2009]</td>
<td>92.4</td>
</tr>
<tr>
<td>[Niebles et al., 2008]</td>
<td>90.0</td>
</tr>
<tr>
<td>[Kläser et al., 2008]</td>
<td>84.3</td>
</tr>
<tr>
<td>[Scovanner et al., 2007]</td>
<td>82.6</td>
</tr>
<tr>
<td>[Niebles and Li, 2007]</td>
<td>72.8</td>
</tr>
</tbody>
</table>
Weizmann dataset: entire sequences

Classification accuracy on entire sequences: different mid- and high-level features

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zernike+CRBM+SVM (Prop. meth.)</td>
<td>100.0</td>
</tr>
<tr>
<td>Zernike+SVM</td>
<td>89.3</td>
</tr>
<tr>
<td>Zernike+PCA (3000 dim.)</td>
<td>89.3</td>
</tr>
<tr>
<td>Zernike+PCA (1000 dim.)</td>
<td>68.8</td>
</tr>
<tr>
<td>Zernike+PCA (150 dim.)</td>
<td>16.1</td>
</tr>
</tbody>
</table>
Video

Demo-video
Conclusion

- Segmentation (BG-subtraction) allows to take decisions on individual objects (humans)
- The generative model takes into account the temporal dependencies and extracts useful features
- The results compare very favorably to the state of the art
Current big issue: classifying individual actions

Two types of methods: segmentation based, or interest point based

Interest points are easy to extract from moving cameras

Shape information is easier to use from segmented frames

Most powerful statistical learning techniques are hard to apply on structural data (interest points)


References VII

