Supervised and unsupervised event detection with local spatio-temporal features

Lorenzo Seidenari

Overview

- Event classification
  - Local space-time descriptors [ICIP’09, ICIAP ’11]

- Effective codebooks [VOEC@ICCV’09, TMM’12]

- Anomaly detection
  - Real-time anomaly detection [MM’10]
  - Anomaly localization [CVIU’12]

- Semantically driven adaptive video coding [ISM’11]
Event detection and human actions

- Most videos recorded and downloadable from the web contain people; the semantic is therefore defined by people behavior.

- Third generation video-surveillance systems benefit from automatic interpretation of human actions and behaviors.

**Definition 1:** physical *body motion*.

**Definition 2:** interaction with environment (objects or people) on a specific purpose.

Anomaly detection in video-surveillance

- Surveillance cameras deployed everywhere. Automatic video analysis is needed in order to scale.
- Events of interest are not always defined.
- Instead of training several discriminative classifiers to recognize a finite set of events we train a model on the available “normal” data and attempt to recognize abnormal samples.
Overview

• Event classification
  • Local space-time descriptors [ICIP’09, ICIAP ’11]

• Effective codebooks [VOEC@ICCV’09, TMM’12]

• Anomaly detection
  • Real-time anomaly detection [MM’10]
  • Anomaly localization [CVIU’12]

• Semantically driven adaptive video coding [ISM’11]

Space-time feature detection

Dollár Detector [VSPETS05]
  • Temporal Gabor filter.

Extended to multi-scale
  • Spatial and temporal scale-space sampling.
  • Dense extraction of interesting locations.
Histogram based space-time descriptors

3D Gradient
- Appearance variation.
- Polar coordinates **quantized separately.**
- Computationally **efficient.**
- Simpler representation – **no descriptor parameter tuning.**

Optical Flow
- Local apparent motion field.
- Orientation histograms
- **No-motion bin.**

Local positional statistics are computed for both image measurement
- Robust with respect to small viewpoint variation.

Moment based space-time descriptors

- To reduce descriptor dimensionality and complexity we employ 3D Zernike moments.
- Zernike moments are proven to be superior in term of **robustness to noise, information redundancy , reconstruction error [Teh&Chin98].**
- Given a function \( f(\xi) \) Zernike moments are:
  \[
  A_{nm}(\xi_0) = \frac{3}{4\pi} \int_{\|\xi - \xi_0\| \leq 1} f(\xi) \left[ v_{nm}^l \left( \frac{\xi - \xi_0}{\sigma} \right) \right]^l d\xi
  \]
  where \( \sigma \) tunes the pixel region size and \( \xi_0 = [x,y,t] \) is the point in which the moment is computed.
- They have a hierarchical structure.
- Higher order moments are the most informative but suffer from noise.
Pyramid Kernel descriptors

• To exploit information of higher order moments and alleviating the effect of noise we propose a pyramidal kernel to match 3D Zernike moment based descriptors at multiple levels of detail.

• An effective strategy to perform multi-resolution matching of features has been proposed both in the feature space [Grauman05] and in the image plane [Lazebnik06].

• We build on this idea and propose a pyramid kernel descriptor.

• The kernel between two descriptors is the following:

\[ K(d_i, d_j) = \sum_{k=0}^{n} 2^{k-n} \frac{s_i^k \cdot s_j^k}{\|s_i^k\| \cdot \|s_j^k\|}. \]

• Descriptor is decomposed using the structure imposed by the orders of moments.

• Weights increase with the order of moments.

• Satisfies Mercer conditions by construction.

Classification Framework Overview

- Interest points
- Bag-of-features
- Visual Dictionary
- SVM classifier
- Bag-of-words

- running
- walking
- jogging
- handwaving
- handclapping
- boxing
**Descriptors combination**

- **ST-Volume** → 3DGrad, HoF → **3DGrad_HoF** → Visual Dictionary → Action Representation → BoW
- **ST-Volume** → 3DGrad, HoF → **3DGrad + HoF BoW**

**Effective codebooks**

- **K-means** equally distributes feature points among all clusters – flat distribution.
- A flatter distribution of the features might imply a less discriminative visual vocabulary [Quelhas TPAMI07].
- **Radius-based** clustering provides a better coding of the mid-frequency features (supposedly more informative) and more discriminative codebooks.
Effective codebooks

- To improve classification time codebook can be compressed with non-linear dimensionality reduction.
- We employed Deep Belief Networks (DBN) fed with bag-of-words vectors.
- A DBN is composed of several Restricted Boltzmann Machines (RBM) building blocks that encode levels of non-linear relationships of the input vectors.
- The compression procedure proved effective wrt linear dimensionality reduction techniques and improved the computation time.

Results: descriptor performance

<table>
<thead>
<tr>
<th>Method</th>
<th>Size</th>
<th>Computation time(s)</th>
<th>Accuracy(%) KTH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyramid Zernike 3D</td>
<td>84</td>
<td>0.0300</td>
<td>91.3</td>
</tr>
<tr>
<td>3DGrad</td>
<td>432</td>
<td>0.0400</td>
<td>90.4</td>
</tr>
<tr>
<td>Dollar’05</td>
<td>100</td>
<td>0.0060</td>
<td>81.2</td>
</tr>
<tr>
<td>Scovanner ’07</td>
<td>640</td>
<td>0.8210</td>
<td>82.6</td>
</tr>
<tr>
<td>Mattivi ‘10</td>
<td>100</td>
<td>0.1000</td>
<td>91.2</td>
</tr>
<tr>
<td>Willems’08</td>
<td>384</td>
<td>0.0005</td>
<td>84.3</td>
</tr>
</tbody>
</table>
### Results: descriptor performance

![Graph showing accuracy vs order]

<table>
<thead>
<tr>
<th>Method</th>
<th>Weizmann ['05] (accuracy)</th>
<th>KTH ['05] (accuracy)</th>
<th>Hollywood2 ['09] (mean Average Precision)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our Method</td>
<td>95.41</td>
<td>92.66</td>
<td>45.1</td>
</tr>
<tr>
<td>Scovanner ['07']</td>
<td>82.60</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Kläser ['08']</td>
<td>84.3</td>
<td>91.4</td>
<td>43.7</td>
</tr>
<tr>
<td>Laptev ['08']</td>
<td>-</td>
<td>91.80</td>
<td>43.5</td>
</tr>
<tr>
<td>Sun ['09']</td>
<td>90.3</td>
<td>89.80</td>
<td>-</td>
</tr>
<tr>
<td>Gao ['10']</td>
<td>-</td>
<td>91.14</td>
<td>-</td>
</tr>
<tr>
<td>Yu ['11']</td>
<td>-</td>
<td>91.80</td>
<td>-</td>
</tr>
</tbody>
</table>

### Results: comparison with the state of the art

![Image showing comparison with state of the art]
Overview

• Event classification
  • Local space-time descriptors [ICIP’09, ICIAP ’11]

• Effective codebooks [VOEC@ICCV’09, TMM’12]

• Anomaly detection
  • Real-time anomaly detection [MM’10]
  • Anomaly localization [CVIU’12]

• Semantically driven adaptive video coding [ISM’11]

Video Model

• Existing approaches rely on: trajectories, optical flow, space-time descriptors.

• Often all of these systems do not have a real-time constraint.
  [Mahadevan CVPR 2010]

• Existing approaches model the feature statistics with mixture models.
  [ Mahadevan CVPR 2010, Kim CVPR 2009]

• Spatio-temporal features for joint appearance and motion modelling.

• Fully non-parametric approach to ease our system deployment:
  • almost no training time.
  • easy system adaptation to different scenes.
  • testing time reduced via fast a-nn structures.
**Video Model**

- Given a set of triples \((d, l, s)\) of descriptors, their locations and scales extracted in the past \(T\) frames we would like to evaluate the likelihood \(p(d, l, s | d, l, s)\) wrt to previous observed triples.
- We do not pose any prior on the locations (any area of the frame is equally likely to generate anomalous patterns).
- We assume models of overlapping and neighboring regions independent.
- For a set of query triples we can evaluate the likelihood with the following:

\[
p(d_q, l_q, s_q | d, l, s) \propto \sum_i \prod_j \prod_k p(d_{qj}, l_{qj}, s_{qj} | d^j, l^j, s^j)
\]

where \(O_i\) indicates the set of patches overlapping region \(i\) and \(N_j\) represents the set of represents the set of neighboring locations at the same scale. These probabilities are evaluated through non-parametric tests.

**Non-parametric model**

- Store “normal” samples in fast approximate NN structures.
- A pattern is anomalous if its “neighbourhood” is empty.
- Need to define a meaningful neighbourhood in feature space.
Optimal radius estimation

- An optimal radius for each location is estimated
  \[ \hat{r}_i = CDF_{i}^{-1} (1 - p_a) \]
- Compute CDF of distances from nearest neighbour.
- Probability \( p_a \) of an anomalous event to happen (e.g. \( 10^{-4} \))

Model update

- An anomaly list for each tree.
- Periodically estimate an intra-list optimal radius \( r_{a,i} \).
- Test each pattern in list with \( r_{a,i} \).
- Add it to the normal dataset if it has neighbours.
- Re-estimate \( r_i \) for each location.
System evaluation

- System is evaluated on UCSD dataset:
  - ~30 minutes of videos from a fixed camera.
  - Normal frames contains pedestrian on a walk-way.
  - Abnormal frames contains bikers, skaters and carts.
  - Ground truth provided at frame and pixel level (on a subset).

Detection examples

System evaluation

- Qualitative comparison with MDT [Mahadevan 2010]

Our approach

MDT Mahadevan 2010
System evaluation

- Qualitative comparison with MDT [Mahadevan 2010]

Our approach

MDT Mahadevan 2010

System performances

- Our method outperforms all other real-time approaches.
- The system is able to process 20fps (240x160) on a standard machine.
- The top performing method requires 25 s to process a frame!
Overview

• Event classification
  • Local space-time descriptors [ICIP’09, ICIAP ’11]

• Effective codebooks [VOEC@ICCV’09, TMM’12]

• Anomaly detection
  • Real-time anomaly detection [MM’10]
  • Anomaly localization [CVIU’12]

• Semantically driven adaptive video coding [ISM’11]

Semantic adaptive video coding

• Modern surveillance systems are distributed camera networks. Typically IP cameras or wireless cameras are employed.
• In this context selectively compressing video streams depending on the semantic content of each frame can result in significant bandwidth savings.
Semantic adaptive video coding

- Visual interest maps can be obtained by motion detection [Huang10], with a trained object detector [Dalal05,Felzenswab10] or using anomaly detection [Seidenari12].
- Supervised or unsupervised methods often rely on image features based on gradients or texture.
- We propose an alternative method to generate visual interest maps based on efficient low level features (FAST corners and Sobel edges).
- Frame is selectively blurred in regions without details (white) and compressed afterwards.

Semantic adaptive video coding

- We measure two performance figures:
  - Structural similarity index
    \[ SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{\mu_X^2 + \mu_Y^2 + C_1}(\sigma_X^2 + \sigma_Y^2 + C_2) \]
  - File size.

- CRF: Constant rate factor. Bitrate corresponds approximately to that of constant quantizer.
- Size is measured in % with respect to original file size.
Publications

International Journals


International Conferences and Workshops


Bibliometric indices

H-index = 3
Total number of citations = 36 (source: Google Scholar on April 3, 2012).