Unsupervised Object Discovery and Segmentation in Videos

Samuel Schulter, Christian Leistner,
Peter M. Roth, Horst Bischof
Graz University of Technology – Institute for Computer Graphics and Vision
Microsoft Austria
What is Unsupervised Object Discovery?

- *Given:* Set of unlabeled images
What is Unsupervised Object Discovery?

• *Given*: Set of unlabeled images

• *Goal*: Discover common visual concepts
What is Unsupervised Object Discovery?

• **Given:** Set of unlabeled images

• **Goal:** Discover common visual concepts
What is Unsupervised Object Discovery?

• **Given**: Set of unlabeled images

• **Goal**: Discover common visual concepts
Typical Approach

• Collection of **still images**
• Topic modelling or clustering methods
• Rely on **prior information**
  – Arbitrary image segmentations
  – Objectness
  – etc.
• Reliable discovery without priors is difficult!
Use Videos instead of Still Images

- Motion is a strong and physically valid prior for objects
- Advantages of using videos
  - Objects can be segmented from the background
  - High variability of object appearance
  - Huge amount of data easily available
UOD in Videos

• **Given:** Videos capturing some objects
• **Goal:** Discover objects and assign them a semantic label
Outline

• Our approach for UOD from Videos
  – Overview
  – Building blocks
  – Outcome

• Experiments
  – Object discovery in videos
  – Object detection in still images
Building Blocks
Building Blocks

Unsupervised Object Discovery and Segmentation from Videos
Building Blocks
Building Blocks

[Diagram showing the process of Unsupervised Object Discovery and Segmentation from Videos]
Building Blocks
Building Blocks

Unsupervised Object Discovery and Segmentation from Videos
Motion Segmentation

- CRF-based segmentation
- Large optical flow vectors indicate objects
Motion Segmentation

• CRF-based segmentation
• Large optical flow vectors indicate objects
Motion Segmentation

- CRF-based segmentation
- Large optical flow vectors indicate objects

Input video  Optical flow  Motion segmentation
Object Proposals from Motion

Object proposal = Motion segment

• Proposals are typically noisy
  – Filter via motion constraints
  – Smooth trajectories through space and time
  – Not possible for still images
Object Proposal Clustering

• Feature vector for each remaining proposal bounding box
  – Bag-of-Words on Dense SIFT (300d codebook)
  – Spatial pyramid

• Choose the number of objects $k$
  – Only supervision required!

• Apply a spectral clustering algorithm
  – $\chi^2$ distance
Clustering Result
Clustering Result
Training Object Models

• Train classifier for each cluster
  – Allows for discovering static objects

• Random Forests on two abstraction levels
  – Superpixel level (standard RF on superpixels)
  – Object level (Hough Forests [Gall & Lempitsky, 09])
Applying Object Models

Superpixel level

Object level
Recap
Recap
Recap
CRF-based Semantic Segmentation

- Graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$
- Nodes on superpixels $S_l$
  - Regular grid
  - Fast computation
- Edges link spatially and temporally
- Label space size: $k+1$ ($k$ categories and background)
CRF-based Semantic Segmentation

- Linear combination of unary potentials
  - Optical flow fields
  - 2 semantic appearance maps
- Contrast-sensitive pairwise potentials
  - RGB color and optical flow vectors
- Standard Graph-Cut for minimization

Details in the paper
CRF-based Semantic Segmentation

- Output: Labeled video frames
Experiments

• Experiments with **video data**
  – Unsupervised object discovery

• Experiments on **still images**
  – Object detection

• **Videos from** [Ommer & Buhmann, 07]
  – 96 videos, > 7000 frames, 4 categories
  – Captured with non-static hand-held camera
Object Discovery in Videos

• **Intention**: Successful discovery of moving and static objects, requiring only the parameter $k$

• Accuracy measure is **purity**

• Frame correctly classified if largest segment is correctly labeled

• Evaluation of different parts of our approach and comparison to [Russel et al., 06]
## Quantitative Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Purity [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (full)</td>
<td>75.1</td>
</tr>
<tr>
<td>Ours (superpixel only)</td>
<td>72.3</td>
</tr>
<tr>
<td>Ours (holistic only)</td>
<td>69.4</td>
</tr>
<tr>
<td>Ours (no outlier rem.)</td>
<td>62.2</td>
</tr>
<tr>
<td>[Russel et al. 06] k=4</td>
<td>52.0</td>
</tr>
<tr>
<td>[Russel et al. 06] k=5</td>
<td>55.0</td>
</tr>
</tbody>
</table>

Results of UOD task as purity

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
</tr>
</thead>
<tbody>
<tr>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>c1</td>
<td>65</td>
<td>05</td>
<td>12</td>
<td>06</td>
</tr>
<tr>
<td>c2</td>
<td>06</td>
<td>88</td>
<td>02</td>
<td>06</td>
</tr>
<tr>
<td>c3</td>
<td>13</td>
<td>06</td>
<td>80</td>
<td>04</td>
</tr>
<tr>
<td>c4</td>
<td>13</td>
<td>00</td>
<td>04</td>
<td>84</td>
</tr>
</tbody>
</table>

Confusion matrix of the 4 categories:
c1 = bicycle
c2 = car
c3 = pedestrian
c4 = streetcar
## Quantitative Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Purity [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (full)</td>
<td>75.1</td>
</tr>
<tr>
<td>Ours (superpixel only)</td>
<td>72.3</td>
</tr>
<tr>
<td>Ours (holistic only)</td>
<td>69.4</td>
</tr>
<tr>
<td>Ours (no outlier rem.)</td>
<td>62.2</td>
</tr>
<tr>
<td>[Russel et al. 06] k=4</td>
<td>52.0</td>
</tr>
<tr>
<td>[Russel et al. 06] k=5</td>
<td>55.0</td>
</tr>
</tbody>
</table>

Results of UOD task as purity

<table>
<thead>
<tr>
<th></th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
</tr>
</thead>
<tbody>
<tr>
<td>c1</td>
<td>65</td>
<td>05</td>
<td>12</td>
<td>06</td>
</tr>
<tr>
<td>c2</td>
<td>06</td>
<td>88</td>
<td>02</td>
<td>06</td>
</tr>
<tr>
<td>c3</td>
<td>13</td>
<td>06</td>
<td>80</td>
<td>04</td>
</tr>
<tr>
<td>c4</td>
<td>13</td>
<td>00</td>
<td>04</td>
<td>84</td>
</tr>
</tbody>
</table>

Confusion matrix of the 4 categories:
c1 = bicycle
c2 = car
c3 = pedestrian
c4 = streetcar
Qualitative Results

Moving objects

Also non-moving objects
(parking cars, pedestrian)

Failure cases
Result Videos
Recognition in Still Images

- **Intention:** Show the generalization capability of the unsupervised learned models on still images
Recognition in Still Images

• *Intention*: Show the generalization capability of the unsupervised learned models on still images
Recogntion in Still Images

• Holistic appearance models can be directly applied on still images [Gall & Lempitsky, 09]

• TUD-pedestrian and ETHZ-cars data sets [Andriluka et al., 08], [Leibe et al., 07]

• Compare 3 models
  – Unsupervised (train images only from videos)
  – Supervised (original train images)
  – Combined (both image sets)
Results on TUD-pedestrian

- Combined model slightly worse than fully supervised
- Only little additional information, as TUD-pedestrian mainly shows side-view pedestrians

Unsupervised HF (ap=0.477)
Supervised HF (ap=0.631)
Combined HF (ap=0.615)
Results on ETHZ-cars

- Combined model significantly outperforms fully-supervised model
- Unlabeled data helps and comes for free!
- Motivating result

Unsupervised HF (ap=0.707)
Supervised HF (ap=0.770)
Combined HF (ap=0.844)
Conclusion

• Unsupervised Object Discovery from videos
• Motion is a strong object indicator
• Include both motion and appearance cues in a joint CRF formulation

• Successful discovery of objects in videos
• Model can even be applied on still images
Thank you!

Samuel Schulter
schulter@icg.tugraz.at
Institute for Computer Graphics and Vision
Graz University of Technology, Austria

References


Conclusion

• Unsupervised Object Discovery from videos
• Include both motion and appearance cues in a joint CRF formulation
• Successful discovery of objects in videos
Conclusion

• Unsupervised Object Discovery from videos
• Include both motion and appearance cues in a joint CRF formulation
• Successful discovery of objects in videos

Take-Home message:
• Motion is a strong prior for objects
• Appearance models also generalize well to still images
• Applicable to object detection
Discussion

• Discuss the pipeline
• Benefits and limitations
• Influence of $k \rightarrow$ scalability with $k$
• Better performance when going pixel-wise and learning some CRF parameters
• Denote this slide as future work? Rather at the end of the presentation?!
Additional Slides

- Camera motion suppression
- Shot boundary detection
- Filtering via line fitting, e.g., x-y-coordinates of bounding box center through space and time
Additional Slides

• Random Forest training
  – 2 Hough Forests (1 without offset vectors)
  – Superpixel double the size → 16x16 patches
  – Object: bounding box → 100px height → 16x16 random patches

• Why holistic model? Only vote for object center? Usefull?
Additional Slide

• CRF segmentation
  – In the first iteration, label space is the same but we spread the motion potentials to all semantic labels equally (and to background in the correct relation)
  – Appearance probabilities are normalized (from Hough Forests)

• Weighting factors are hand-tuned
• Add constant fg-probability to motion!
Unsupervised Object Retrieval

• Learn categories from unlabeled videos
• Predict the correct label for unseen test frames
• Illustration of the generalization capability

• Split the videos into train and test set (3:1)
• Accuracy metric
  – Retrieval rates per frame and video
Results

- Our model has less supervision and no shape information
- Our unsupervised „appearance only“ is 13% better than the weakly supervised „appearance only“ model

<table>
<thead>
<tr>
<th>Model</th>
<th>Frame</th>
<th>Video</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (full)</td>
<td>65.9</td>
<td>73.9</td>
</tr>
<tr>
<td>[Ommer &amp; Buhmann 07]</td>
<td>74.3</td>
<td>87.4</td>
</tr>
<tr>
<td>[Ommer et al. 09] Appear</td>
<td>53.0</td>
<td>58.9</td>
</tr>
<tr>
<td>[Ommer et al. 09] Shape</td>
<td>74.4</td>
<td>88.4</td>
</tr>
<tr>
<td>[Ommer et al. 09] Combination</td>
<td>81.4</td>
<td>94.5</td>
</tr>
</tbody>
</table>
Motion Segmentation

- CRF-based motion segmentation
- Superpixels $s_l$
  - Regular grid
  - Fast computation

- Unary potential based on optical flow vectors
  - Large optical flow vectors indicate objects

$$\Phi(s_l) = -\log \left( \eta + \frac{\text{med}(\|v(s_l)\|)}{\max_l \text{med}(\|v(s_l)\|)} \right)$$