Seeing the Objects Behind the Dots: Recognition in Videos from a Moving Camera

Bjorn Ommer, Theodor Mader, Joachim M. Buhmann
Present by: Alisa Kunapinun
Abstract

- This paper also significantly contributes to the systems design aspect—showing how all of these subtasks can be combined in a computer vision system so that they mutually benefit from another.
Keywords

- Object Recognition
- Segmentation
- Tracking
- Videos Analysis
- Compositionality
- Visual Learning
Process on this paper
Problem are investigated

- Category-level recognition
- Reducing supervision during learning
- Segmentation of videos from a moving camera
- Tracking without manual interaction
- Object models and shape representation
Process Pipeline

[Diagram showing process pipeline with steps such as interest points, optical flow, tracking, segmentation, and recognition.]
Region Tracking and Object Segmentation

- Tracking Object Regions
  - Compositions as Spatial Groupings of Parts
  - A composition represents all its constituent interest points
- Tracking Compositions
- Temporal Grouping of Composition

\[ \Gamma^t(j) = \{ i : \text{IP } i \text{ in neighborhood of } j\text{-th comp.} \}. \]  \hspace{1cm} (1)

\[ g^t_j := \frac{1}{|\Gamma^t(j)|} \sum_{i \in \Gamma^t(j)} d_i^t. \]  \hspace{1cm} (2)

\[ x_{j}^{t+1} := x_{j}^{t} + \frac{1}{|\Gamma^t(j)|} \sum_{i \in \Gamma^t(j)} d_i^t. \]  \hspace{1cm} (3)

\[ h_{j}^{t} = \eta g_{j}^{t} + (1 - \eta) h_{j}^{t-1}. \]  \hspace{1cm} (4)
Joint Tracking and Segmentation of Objects Based on Floating Image Regions

- Problem: assemble the object regions into the different objects and into background.
- Solving: Expectation-Maximization approach (EM)
- After use EM-Algorithm: Using Segmentation to Refine Object Region Tracking Algorithm
EM-Algorithm

\textsc{CompSegmentation}(\{h^t_j\}_j, \{T^{t-1}_v\}_{v=1,...,K})

1. Initialization: \( \forall v : T^t_v \leftarrow T^{t-1}_v \)
2. repeat
3. \hspace{1em} E-Step: \( \triangleright \) update assignments:
4. \hspace{2em} \( M^t_{j,v} \leftarrow 1 \{ v = \arg\min_v \mathcal{R}(T^t_v, h^t_j, x^t_j) \} \)
5. \hspace{1em} M-Step: \( \triangleright \) update segments:
6. \hspace{2em} for \( v = 1, \ldots, K \)
7. \hspace{3.5em} do Solve with Levenberg-Marquardt (start with \( \hat{T}^t_v \leftarrow T^t_v \)):
8. \hspace{4.5em} \( T^t_v \leftarrow \arg\min_{\alpha, \delta, \delta_x, \delta_y} \sum_j M^t_{j,v} \mathcal{R}(\hat{T}^t_v, h^t_j, x^t_j) \)
9. \hspace{2em} until convergence of \( M^t_{j,v} \)
10. return \( M^t, \{T^t_v\}_{v=1,...,K} \)
Using Segmentation to Refine Object Region Tracking Algorithm

\textsc{CompositionTracking}((h^{t-1}_j, x^t_j)_j, \{T^{t-1}_v\}_v=1,...,K)

1. Detect interest points $i$ in frame $t$
2. for all compositions $j$ \textit{▷} update comps with IP flow:
3. do $\Gamma'(j) \leftarrow \{i: \|x^t_j - \bar{x}^t_i\| \leq w\}$
4. \hspace{1cm} $g^t_j \leftarrow \frac{1}{|\Gamma'(j)|} \sum_{i \in \Gamma'(j)} d^t_i$
5. \hspace{1cm} $h^t_j \leftarrow \eta g^t_j + (1 - \eta)h^t_{j-1}$
6. $M^t, \{T^t_v\}_v \leftarrow \text{CompSegmentation}((h^t_j)_j, \{T^t_v\}_v)$
7. for all compositions $j$ \textit{▷} update comps with segmentation:
8. do $\Gamma'(j) \leftarrow \{i: i \in \Gamma'(j) \wedge$
9. \hspace{3cm} $1 = M^t_{j, \text{argmin}_R R(T^t_v, d^t_i, \bar{x}^t_i)}$
10. \hspace{1cm} $g^t_j \leftarrow \frac{1}{|\Gamma'(j)|} \sum_{i \in \Gamma'(j)} d^t_i$
11. \hspace{1cm} $h^t_j \leftarrow \eta g^t_j + (1 - \eta)h^t_{j-1}$
12. \hspace{1cm} $x^{t+1}_j \leftarrow x^t_j + \frac{1}{|\Gamma'(j)|} \sum_{i \in \Gamma'(j)} d^t_i$
13. return $(h^t_j, x^{t+1}_j)_j, \{T^t_v\}_v$
Object Representations for Category-Level Recognition

- Compositional, Appearance-Based Model: Use multi-class SVM
- Recognition Using the Motion of Dot Patterns: Use SVM
- Global Shape and Local Appearance Combined
- Processing Pipeline for Training: Use SVM
Process Output(2)
Experiments

- Recognition Performance on Videos with Substantial Camera Motion
  - use 10-fold cross-validation and train on 16 randomly
  - Object models are learn on a randomly drawn subset of 15 frames per train video
Experiment Output

- Baseline Performance of Appearance w/o Compositions and Shape—Bag-of-Parts
  - 53.0 . 5.6% of all frames correctly.

- Compositional Segmentation and Recognition w/o Shape Model
  - 64.9 . 5.4% per frame.
# Experiment Output(2)

- Comparing the Different Algorithm

<table>
<thead>
<tr>
<th>Object model</th>
<th>Per frame</th>
<th>Per video</th>
</tr>
</thead>
</table>

*Dataset of Ommer and Buhmann (2007)*
(car, bicycle, pedestrian, streetcar):

- Approach of Ommer and Buhmann (2007): 74.3 ± 4.3
- Compositional motion (13): 52.6 ± 1.1
- Appearance-only: bag-of-parts: 53.0 ± 5.6
- Segment. w/o shape: bag-of-comps: 64.9 ± 5.4
- Shape: $P(e^{b,v} | V^{t,v})$ (11): 74.4 ± 5.3
- Compositional appear + location (12): 79.6 ± 5.5
- Combined shape + appear (14): 81.4 ± 2.9

*Dataset (Ommer and Buhmann 2007) plus additional category “cow” from (Magee and Boyle 2002):*

- Compositional appearance (12): 76.5 ± 2.4

88.4 ± 2.3
### Comparing Different Object Models

<table>
<thead>
<tr>
<th>True classes →</th>
<th>Bicycle</th>
<th>Car</th>
<th>Pedest</th>
<th>Streetcar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bicycle</td>
<td>74.3</td>
<td>3.2</td>
<td>13.7</td>
<td>2.9</td>
</tr>
<tr>
<td>Car</td>
<td>7.8</td>
<td>84.1</td>
<td>4.2</td>
<td>5.9</td>
</tr>
<tr>
<td>Pedestrian</td>
<td>13.3</td>
<td>2.5</td>
<td>80.0</td>
<td>3.9</td>
</tr>
<tr>
<td>Streetcar</td>
<td>4.7</td>
<td>10.2</td>
<td>2.2</td>
<td>87.3</td>
</tr>
</tbody>
</table>
**Computational Demands**

- recognizes objects in videos of 768x576 pixel
- using the combined shape and appearance model at the order of 1 fps on a 3 GHz Pentium 4 desktop PC.

<table>
<thead>
<tr>
<th>Processing step</th>
<th>Comp. demand</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tracking and segmentation, Algorithm 2:</td>
<td></td>
</tr>
<tr>
<td>IPs $i$, flow $d_i^j$ (Algorithm 2, line 1)</td>
<td>27.7%</td>
</tr>
<tr>
<td>Updating comps (Algorithm 2, line 2–5)</td>
<td>5.2%</td>
</tr>
<tr>
<td>EM estimation Algorithm 1, i.e. (Algorithm 1, line 6)</td>
<td>4.9%</td>
</tr>
<tr>
<td>Updating comps with segm. (Algorithm 2, line 7–11)</td>
<td>0.3%</td>
</tr>
<tr>
<td>Feature extraction and recognition:</td>
<td></td>
</tr>
<tr>
<td>Computing loc feat hists to represent $a_i^j$ (Sect. 3.2)</td>
<td>36.5%</td>
</tr>
<tr>
<td>Computing all individual probs in (14)</td>
<td>12.3%</td>
</tr>
<tr>
<td>Eval. GM of Fig. 5, i.e. calc. product in (14)</td>
<td>0.09%</td>
</tr>
<tr>
<td>Video stream ops, writing of results, etc.</td>
<td>12.9%</td>
</tr>
</tbody>
</table>
Action Recognition using KTH

<table>
<thead>
<tr>
<th>True classes →</th>
<th>Box</th>
<th>Hclp</th>
<th>Hwav</th>
<th>Jog</th>
<th>Run</th>
<th>Walk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boxing</td>
<td>84.5</td>
<td>0.0</td>
<td>5.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Hand clapping</td>
<td>1.0</td>
<td>87.0</td>
<td>16.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Hand waving</td>
<td>12.5</td>
<td>13.0</td>
<td>75.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Jogging</td>
<td>0.0</td>
<td>0.0</td>
<td>0.5</td>
<td>93.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Running</td>
<td>2.0</td>
<td>0.0</td>
<td>0.0</td>
<td>3.0</td>
<td>92.3</td>
<td>5.0</td>
</tr>
<tr>
<td>Walking</td>
<td>0.0</td>
<td>0.0</td>
<td>2.0</td>
<td>4.0</td>
<td>7.7</td>
<td>95.0</td>
</tr>
</tbody>
</table>