Efficient Temporal Consistency for Streaming Video Scene Analysis

Ondrej Miksik  Daniel Munoz
J. Andrew Bagnell  Martial Hebert

Oxford Brookes Vision Group

Carnegie Mellon
THE ROBOTICS INSTITUTE
Scene Analysis/Parsing

Temporal Inconsistency
Per-frame
Munoz ECCV 2010

Temporal
This Work
Spatio-temporal Graphical Model

• Approximate inference
• Typically the MAP labeling (no uncertainty)
• Mini-batch processing

Spatio-temporal Segmentation

- Hard decisions

Grundmann CVPR 2010, Brendel & Todorovic ICCV 2011, Xu & Corso ECCV 2012
Spatio-temporal Segmentation

- Hard decisions

Grundmann CVPR 2010, Brendel & Todorovic ICCV 2011, Xu & Corso ECCV 2012
This Work

- Causal/Online/Running-average Algorithm
- Efficient and easy to implement
- Complements your favorite scene parser
- Outputs per-pixel probabilities
t-1

Temporally smoothed
Per-pixel Probabilities

\[ y_i = [P_i(\text{Tree}), P_i(\text{Sidewalk}), \ldots, P_i(\text{Car})]^T \]
Temporally smoothed

Make Consistent!

Scene Parser
Munoz 2010
Farabet 2011
Wojek 2008

Per-frame

t-1
Ideally: Exact Correspondences

\[
y_i^{(t)} \quad \hat{y}_j^{(t-1)}
\]
Ideally: Exact Correspondences

\[ \hat{y}_{i}^{(t)} \propto y_{i}^{(t)} \circ \hat{y}_{j}^{(t-1)} \]
Ideally: Exact Correspondences

\[ \hat{y}_i^{(t)} \propto y_i^{(t)} + \hat{y}_j^{(t-1)} \]
Approx. Correspondences via Optical Flow

Optical Flow
Werlberger 2009

\[ t-1 \]  \quad \Rightarrow \quad \text{Optical Flow} \quad \Rightarrow \quad \text{Werlberger 2009} \quad \Rightarrow \quad t \]
Handling No/Imperfect Correspondences
Handling No/Imperfect Correspondences
Handling No/Imperfect Correspondences
Handling No/Imperfect Correspondences
Handling No/Imperfect Correspondences
Handling No/Imperfect Correspondences
Handling No/Imperfect Correspondences
Handling No/Imperfect Correspondences
Handling No/Imperfect Correspondences
Handling No/Imperfect Correspondences
Causal/Running Update

\[
\hat{y}_i^{(t)} \propto y_i^{(t)} + \sum_{j \in N_i^{(t)}} w_{ij} \hat{y}_j^{(t-1)}
\]
\[ w_{ij} = \exp \left( - \frac{d(f_i, f_j)}{\sigma^2} \right) \]

\[ d(f_i, f_j) = (f_i - f_j)^T (f_i - f_j) \]
Metric Learning

Metric Learning

\[ d_M(f_i, f_j) \]

Metric Learning

\[ d_M(f_i, f_j) \leq d_M(f_i, f_k) + 1 \]

Learned Metric

\[ d_M(f_i, f_j) = (f_i - f_j)^T M (f_i - f_j) \]
Euclidean vs. Learned

$$w_{ij} = \exp \left( -\frac{d(f_i, f_j)}{\sigma^2} \right)$$
Summary

• Input: images and per-pixel probabilities

• Compute and apply optical flow (0.02 s)

• Compute pixel similarities and average (0.75 s)

\[
\hat{y}_i^{(t)} \propto y_i^{(t)} + \sum_{j \in N_i^{(t)}} w_{ij} \hat{y}_j^{(t-1)}
\]
CamVid Dataset (8 labels)

Per-frame
Munoz ECCV 2010

Temporal
This Work
NYU Scenes Dataset (11 labels)

Per-frame
Farabet ICML 2012

Temporal
This Work
MPI-VehicleScenes Dataset (5 labels)

Per-frame
Wojek ECCV 2008*

Temporal
This Work
## Overall Pixel Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Per-frame</th>
<th>Temporal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CamVid-05VD</td>
<td>84.60</td>
<td>86.85</td>
</tr>
<tr>
<td>CamVid-16E5</td>
<td>87.37</td>
<td>88.84</td>
</tr>
<tr>
<td>NYUScenes</td>
<td>71.11</td>
<td>75.31</td>
</tr>
<tr>
<td>MPI-Vehicle</td>
<td>93.10</td>
<td>93.55</td>
</tr>
</tbody>
</table>
“Relative” Pixel Accuracy

- Evaluate on pixels where **Per-frame** and **Temporal** predictions disagree

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Per-frame</th>
<th>Temporal</th>
</tr>
</thead>
<tbody>
<tr>
<td>CamVid-05VD</td>
<td>26.1</td>
<td>59.1</td>
</tr>
<tr>
<td>CamVid-16E5</td>
<td>28.6</td>
<td>53.0</td>
</tr>
<tr>
<td>NYUScenes</td>
<td>20.9</td>
<td>50.0</td>
</tr>
<tr>
<td>MPI-Vehicle</td>
<td>40.7</td>
<td>53.3</td>
</tr>
</tbody>
</table>
Conclusion

• Pros:
  – Efficient temporal consistency
  – Use your favorite scene parsing output
  – Easy to implement

• Cons:
  – Can’t fix major errors from the scene parser
  – Can over-smooth small objects
Thanks!
BACKUPS
CamVid (Sequence 05VD)
# CamVid

<table>
<thead>
<tr>
<th>Class</th>
<th>Per-frame</th>
<th>Temporal</th>
<th>Per-frame</th>
<th>Temporal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sky</td>
<td>.303</td>
<td>.682</td>
<td>.237</td>
<td>.639</td>
</tr>
<tr>
<td>Tree</td>
<td>.352</td>
<td>.563</td>
<td>.336</td>
<td>.518</td>
</tr>
<tr>
<td>Road</td>
<td>.197</td>
<td>.546</td>
<td>.150</td>
<td>.529</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>.277</td>
<td>.512</td>
<td>.188</td>
<td>.357</td>
</tr>
<tr>
<td>Building</td>
<td>.165</td>
<td>.232</td>
<td>.275</td>
<td>.395</td>
</tr>
<tr>
<td>Car</td>
<td>.127</td>
<td>.456</td>
<td>.304</td>
<td>.597</td>
</tr>
<tr>
<td>Pole</td>
<td>.201</td>
<td>.386</td>
<td>.252</td>
<td>.285</td>
</tr>
<tr>
<td>Person</td>
<td>.138</td>
<td>.218</td>
<td>.324</td>
<td>.193</td>
</tr>
<tr>
<td>Bicycle</td>
<td>.323</td>
<td>.165</td>
<td>.348</td>
<td>.043</td>
</tr>
<tr>
<td>Fence</td>
<td>.325</td>
<td>.712</td>
<td>.335</td>
<td>.668</td>
</tr>
<tr>
<td>Sign</td>
<td>.367</td>
<td>.029</td>
<td>.378</td>
<td>.039</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td><strong>.261</strong></td>
<td><strong>.591</strong></td>
<td><strong>.286</strong></td>
<td><strong>.530</strong></td>
</tr>
</tbody>
</table>
NYU Scenes Dataset (11 labels)

Bar chart showing the performance of per-frame and independent methods for different categories in the NYU Scenes Dataset.
# NYU Scenes Dataset (11 labels)

<table>
<thead>
<tr>
<th>Class</th>
<th>Per-frame</th>
<th>Independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Building</td>
<td>.231</td>
<td>.547</td>
</tr>
<tr>
<td>Car</td>
<td>.207</td>
<td>.630</td>
</tr>
<tr>
<td>Door</td>
<td>.046</td>
<td>.000</td>
</tr>
<tr>
<td>Person</td>
<td>.094</td>
<td>.080</td>
</tr>
<tr>
<td>Pole</td>
<td>.169</td>
<td>.139</td>
</tr>
<tr>
<td>Road</td>
<td>.152</td>
<td>.575</td>
</tr>
<tr>
<td>Sidewalk</td>
<td>.373</td>
<td>.274</td>
</tr>
<tr>
<td>Sign</td>
<td>.127</td>
<td>.000</td>
</tr>
<tr>
<td>Sky</td>
<td>.002</td>
<td>.019</td>
</tr>
<tr>
<td>Tree</td>
<td>.353</td>
<td>.630</td>
</tr>
<tr>
<td>Window</td>
<td>.102</td>
<td>.101</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td><strong>.209</strong></td>
<td><strong>.500</strong></td>
</tr>
</tbody>
</table>
MPI-VehicleScenes Dataset (5 labels)

![Bar chart showing accuracy for different labels: Background, Road, Lane-marking, Vehicle, Sky, and Accuracy. The chart compares per-frame and temporal accuracy.]
# MPI-VehicleScenes Dataset (5 labels)

<table>
<thead>
<tr>
<th>Class</th>
<th>Per-frame</th>
<th>Temporal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Background</td>
<td>.420</td>
<td>.730</td>
</tr>
<tr>
<td>Road</td>
<td>.503</td>
<td>.321</td>
</tr>
<tr>
<td>Lane-marking</td>
<td>.779</td>
<td>.319</td>
</tr>
<tr>
<td>Vehicle</td>
<td>.075</td>
<td>.276</td>
</tr>
<tr>
<td>Sky</td>
<td>.206</td>
<td>.571</td>
</tr>
<tr>
<td><strong>Accuracy</strong></td>
<td><strong>.407</strong></td>
<td><strong>.533</strong></td>
</tr>
</tbody>
</table>