Stixmentation – Probabilistic Stixel based Traffic Scene Labeling
Roadmap: Detecting Moving Objects

Where are the moving objects?

→ Stereo Computation
Stereo Computation

Step 1: Dense Stereo Computation using SGM\textsuperscript{1, 2}

Open Issues:

- Where are the relevant obstacles?
- Too much data!
- Stereo measurement outliers

Stixel Computation

Step 2: Computing the Stixel World

Stixel Computation

Open Issues:

- Everything is static?

Step 3: Computing the Dynamic Stixel World

Stixel Motion Estimation

Open Questions

- Decision between moving and stationary not clear yet!
- Grouping of Stixels into objects is desirable for many applications

1st Try: Simple Velocity Thresholding

Desired output
1st Try: Simple Velocity Thresholding

Motion information alone is not sufficient!
Related Work (1)

MRF based Scene Labeling for **static scenes:**

- [Brostow‘08] : Segmentation and Recognition using Structure from Motion Point Clouds, ECCV 2008

- [Brostow‘09] : Combining Appearance and Structure from Motion Features for Road Scene Understanding, BMVC 2009


Related Work (2)

Motion segmentation (dynamic scenes):

- Layered, regularised **optical flow** reconstruction:
  [Unger’12]: Joint Motion Estimation and Segmentation of Complex Scenes with Label Costs and Occlusion Modeling, CVPR 2012

- **Clustering** based:
  [Kitt’10]: Detection and Tracking of Independently Moving Objects in Urban Environments, ITSC 2010

- **CRF** pixel-based:
Better Try: Probabilistic Formulation

Best Segmentation:

$$L^* = \arg\max_{L^t \in \mathcal{L}^t} p \left( L^t \mid Z^t, L^{t-1} \right)$$

5 Classes: $l_i \in \{ \text{Forward, Left, Right, Oncoming, Stationary} \}$

CRF formulation: minimize

$$E = -\log p \left( L^t \mid Z^t, L^{t-1} \right)$$

$$\sim \sum_{i=1}^{N} \psi \left( l_i^t \mid Z^t, L^{t-1} \right) + \lambda \cdot \sum_{(i,j) \in \mathcal{N}_2} \phi \left( l_i^t, l_j^t \mid Z^t, L^{t-1} \right)$$

Features:

$$Z_i^t = \{ \dot{X}_i^t, \dot{Z}_i^t, X_i^t, H_i^t, Z_i^t \}^T$$
Better Try:
Probabilistic Formulation

\[ E = - \log p \left( l^t \mid Z^t, L^{t-1} \right) \]
\[ \approx \sum_{i=1}^{N} \psi \left( l^t_i \mid Z^t, L^{t-1} \right) + \lambda \cdot \sum_{(i,j) \in N_2} \phi \left( l^t_i, l^t_j \mid Z^t, L^{t-1} \right) \]

Model the unary terms \[ \psi \left( l^t_i \mid Z^t, L^{t-1} \right) = - \log p \left( l^t_i \mid Z^t, l^{t-1}_i \right) \] with

\[
p \left( l^t_i \mid Z^t, l^{t-1}_i \right) = p \left( l^t_i \mid z^t_i, l^{t-1}_i \right)
\approx p \left( z^t_i, l^{t-1}_i \mid l^t_i \right) \cdot p \left( l^t_i \right)
\approx p \left( z^t_i \mid l^t_i \right) \cdot p \left( l^{t-1}_i \mid l^t_i \right) \cdot p \left( l^t_i \right)
\]

- **Data Term**
- **Temporal Expectation**
- **Prior Term**

\[
p \left( z^t_i \mid l^t_i \right) = p \left( \dot{X}^t_i, \dot{Z}^t_i, X^t_i, H^t_i, Z^t_i \mid l^t_i \right)
\approx p \left( \dot{X}^t_i, \dot{Z}^t_i \mid Z^t_i, l^t_i \right) \cdot p \left( X^t_i, Z^t_i \mid l^t_i \right) \cdot p \left( H^t_i \mid l^t_i \right)
\]

- **Motion Term**
- **Position Term**
- **Height Term**
Better Try: Probabilistic Formulation

\[ E = - \log p (L^t | Z^t, L^{t-1}) \]
\[ \approx \sum_{i=1}^{N} \psi (l^t_i | Z^t, L^{t-1}) + \lambda \cdot \sum_{(i,j) \in \mathcal{N}_2} \phi (l^t_i, l^t_j | Z^t, L^{t-1}) \]

Model the unary terms

\[ \psi (l^t_i | Z^t, L^{t-1}) = -\log p (l^t_i | Z^t, l^{t-1}_i) \]

with

\[ p (l^t_i | Z^t, l^{t-1}_i) = p (l^t_i | z^t_i, l^{t-1}_i) \]
\[ \approx p (z^t_i, l^{t-1}_i | l^t_i) \cdot p (l^t_i) \]
\[ \approx p (z^t_i | l^t_i) \cdot p (l^{t-1}_i | l^t_i) \cdot p (l^t_i) \]

Data Term Temporal Expectation Prior Term

\[ p (z^t_i | l^t_i) = p (\hat{X}^t_i, Z^t_i, X^t_i, H^t_i, Z_i | l^t_i) \]
\[ \approx p (\hat{X}^t_i, Z^t_i | Z_i, l^t_i) \cdot p (X^t_i, Z^t_i | l^t_i) \cdot p (H^t_i | l^t_i) \]

motion term position term height term
Better Try: Probabilistic Formulation

\[
E = - \log p \left( \mathbf{L}^t \mid \mathbf{Z}^t, \mathbf{L}^{t-1} \right) \\
\sim \sum_{i=1}^{N} \psi \left( l_i^t \mid \mathbf{Z}^t, \mathbf{L}^{t-1} \right) + \lambda \cdot \sum_{(i,j) \in \mathcal{N}_2} \phi \left( l_i^t, l_j^t \mid \mathbf{Z}^t, \mathbf{L}^{t-1} \right)
\]

Model the unary terms \( \psi \left( l_i^t \mid \mathbf{Z}^t, \mathbf{L}^{t-1} \right) = - \log p \left( l_i^t \mid \mathbf{Z}^t, l_i^{t-1} \right) \) with

\[
p \left( l_i^t \mid \mathbf{Z}^t, l_i^{t-1} \right) = p \left( l_i^t \mid \mathbf{Z}_i^t, l_i^{t-1} \right)
\approx p \left( \mathbf{Z}_i^t, l_i^{t-1} \mid l_i^t \right) \cdot p \left( l_i^t \right)
\approx p \left( \mathbf{Z}_i^t \mid l_i^t \right) \cdot p \left( l_i^{t-1} \mid l_i^t \right) \cdot p \left( l_i^t \right)
\]

Data Term \quad Temporal Expectation \quad Prior Term

\[
p \left( \mathbf{Z}_i^t \mid l_i^t \right) = p \left( \mathbf{X}_i^t, \mathbf{Z}_i^t, \mathbf{H}_i^t, \mathbf{Z}_i^t \mid l_i^t \right)
\approx p \left( \mathbf{X}_i^t, \mathbf{Z}_i^t \mid \mathbf{Z}_i^t, l_i^t \right) \cdot p \left( \mathbf{X}_i^t, \mathbf{Z}_i^t \mid l_i^t \right) \cdot p \left( \mathbf{H}_i^t \mid l_i^t \right)
\]

motion term \quad position term \quad height term
Consider a Stixel without any measurement features. We don’t know anything about its object class.

Each Stixel has:

- an a priori class membership probability

Maximize

\[
p(L^t \mid Z^t, L^{t-1}) \propto \prod_{i=1}^{N} p\left(\dot{X}_i^t, \dot{Z}_i^t \mid Z_i^t, l_i^t\right) \cdot p\left(X_i^t, Z_i^t \mid l_i^t\right) \cdot p\left(H_i^t \mid l_i^t\right) \cdot \prod_{i=1}^{N} p\left(l_{i-1}^t \mid l_i^t\right) \cdot p\left(l_i^t\right) \cdot \prod_{(i, j) \in N_2} \left(p\left(l_i^t, l_j^t \mid Z^t\right)\right)^\lambda
\]
Consider a Stixel without any measurement features. We don’t know anything about its object class.

Each Stixel has:

- an a priori class membership probability
A priori knowledge

<table>
<thead>
<tr>
<th></th>
<th>BG</th>
<th>LEFT</th>
<th>RIGHT</th>
<th>FORWARD</th>
<th>ONCOMING</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prior (GT)</td>
<td>87%</td>
<td>1%</td>
<td>1%</td>
<td>7%</td>
<td>4%</td>
</tr>
</tbody>
</table>

Ground Truth statistics was setup from a long downtown drive (38 000 images, manually labeled)

In doubt, favor the static background class
A priori knowledge

Ground Truth statistics was setup from a long downtown drive (38 000 images, manually labeled)

In doubt, favor the static background class
Data Term: Height Feature

Each Stixel has:

- Prior knowledge stationary?
- A height [meter]

\[
p(L^t | Z^t, L^{t-1}) \propto \prod_{i=1}^{N} p(\tilde{X}_i^t, \tilde{Z}_i^t | Z_i^t, l_i^t) \cdot p(X_i^t, Z_i^t | l_i^t) \cdot p(H_i^t | l_i^t) \cdot \prod_{i=1}^{N} p(l_i^{t-1} | l_i^t) \cdot p(l_i^t) \cdot \prod_{(i,j) \in N_2} (p(l_i^t, l_j^t | Z^t))^\lambda
\]

\[H = 1.2m\]
Data Term: Height Feature
Data Term: Height Feature

H = 1.2 m

Moving?
Data Term: Height Feature

GT probability:

Moving?  

- height

$H = 1.2 \text{ m}$
Data Term: Height Feature

GT probability:

Moving?

H = 1.2 m
Data Term : Position
Feature

Each Stixel has:

✓ Prior knowledge  ➔ stationary ?
✓ a height [meter]  ➔ moving?
Data Term: Position
Feature

Each Stixel has:

✓ Prior knowledge → stationary?
✓ a height [meter] → moving?
✓ a 2D world position [meter]
Data Term: Position Feature

Each Stixel has:

- Prior knowledge → stationary?
- a height [meter] → moving?
- a 2D world position [meter]

\[
p(L^t | Z^t, L^{t-1}) \propto \prod_{i=1}^{N} p\left(\hat{X}_i^t, \hat{Z}_i^t | Z_i^t, l_i^t\right) \cdot p(X_i^t, Z_i^t | l_i^t) \cdot p(H_i^t | l_i^t) \cdot \prod_{(i, j) \in \mathcal{N}_2} (p(l_i^t, l_j^t | Z^t))^\lambda
\]
Data Term: Position Feature
Data Term: Position Feature
Data Term: Position Feature

\[ X = 5.0 \text{ m} \]
\[ Z = 18.0 \text{ m} \]
Data Term: Position Feature

- position

GT probability: High...Low
Data Term: Position Feature

GT probability:

+ position

High...Low
Data Term: Velocity Feature

Each Stixel has:

✓ Prior knowledge → stationary?
✓ a height [meter] → moving?
✓ a 3D world position [meter] → stationary?

\[
p(L^t | Z^t, L^{t-1}) \propto \prod_{i=1}^{N} p(\hat{X}^t_i, \hat{Z}^t_i | Z^t_i, l^t_i) \cdot p(X^t_i, Z^t_i | l^t_i) \cdot p(H^t_i | l^t_i) \cdot \\
\prod_{i=1}^{N} p(l^t_i \mid l^t_i) \cdot p(l^t_i) \cdot \prod_{(i,j) \in N_2} (p(l^t_i, l^t_j \mid Z^t))^\lambda
\]
Data Term: Velocity Feature

Each Stixel has:

- Prior knowledge → stationary?
- a height [meter] → moving?
- a 3D world position [meter] → stationary?
- a 2D velocity [m/s]
Data Term: Velocity Feature

\[ Z = 0-20 \text{ m} \]

\[ Z = 20-40 \text{ m} \]

\[ Z = 40-70 \text{ m} \]
Data Term: Velocity Feature

\[ v_x = +0.7 \text{ m/s} \]
\[ v_z = -2.0 \text{ m/s} \]
Data Term: Velocity Feature

- motion

GT probability: High...Low
Data Term: Velocity Feature

+ motion

GT probability: High...Low
# Experimental Results

<table>
<thead>
<tr>
<th>Features</th>
<th>BG</th>
<th>LEFT</th>
<th>RIGHT</th>
<th>FORWARD</th>
<th>ONCOMING</th>
<th>Global</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>99.5</td>
<td>85.1</td>
<td>95.7</td>
<td>79.8</td>
<td>83.2</td>
<td>98.0</td>
</tr>
<tr>
<td>w/o motion</td>
<td>92.3</td>
<td>0.0</td>
<td>0.0</td>
<td>48.0</td>
<td>56.1</td>
<td>85.4</td>
</tr>
<tr>
<td>w/o height</td>
<td>99.7</td>
<td>83.7</td>
<td>94.9</td>
<td>77.4</td>
<td>80.6</td>
<td>97.9</td>
</tr>
<tr>
<td>w/o prior</td>
<td>88.7</td>
<td>93.0</td>
<td>97.3</td>
<td>89.0</td>
<td>88.9</td>
<td>89.0</td>
</tr>
<tr>
<td>w/o position</td>
<td>99.8</td>
<td>92.7</td>
<td>95.8</td>
<td>75.7</td>
<td>66.4</td>
<td>98.0</td>
</tr>
<tr>
<td>w/o binary</td>
<td>98.1</td>
<td>80.3</td>
<td>89.7</td>
<td>75.6</td>
<td>80.9</td>
<td>96.3</td>
</tr>
<tr>
<td>w/o temporal</td>
<td>99.7</td>
<td>84.3</td>
<td>95.2</td>
<td>78.0</td>
<td>83.1</td>
<td>98.0</td>
</tr>
</tbody>
</table>

Data Set: Amount of training images : 38 000
Evaluation : 8 000
Experimental Results: Example
Conclusion

- CRF approach for Traffic Scene Labeling based on the Dynamic Stixel World

- Real time implementation: computational time on a single CPU core ~1 ms

- Training and Evaluation on sufficient data sets demonstrate its performance and its ability to generalize to unseen Traffic Scenes