Transductive deep hand segmentation

Natalia Neverova, PhD student, 3rd year
Université de Lyon, INSA-Lyon
LIRIS UMR CNRS 5205

Christian Wolf
Guelph University
Canada

Graham W. Taylor

Florian Nebout
Awabot

Elisa Fromont Rémi Emonet Taygun Kekec Alain Trémeau Damien Muselet
LHC UMR CNRS 5516

March 20th, 2015
Image segmentation

Segmentation into distinct regions
Features: color, grayscale, texture, depth, SIFT/HoG, learned features ...

Semantic full scene labelling

Segmentation of the human body

Background subtraction in videos

Segmentation for object detection (R-CNN, Girshick CVPR 2014)
Segmentation and spatial relationships

<table>
<thead>
<tr>
<th>Features</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_i$</td>
<td>$l_i$</td>
</tr>
</tbody>
</table>

- **No**
  - Pixelwise classification (independant)

- **Yes**
  - MRF/CRF/BN. Inference of a global solution with high computational complexity

- **Yes**
  - Auto-context models

- **Yes**
  - Pixelwise classification. The prior improves the classifier
Semantic full scene labelling

Hand segmentation (pose estimation)
Semantic full scene labelling

Goal: label each pixel of an image with a semantic class.

Joint work with LHC, St. Etienne

Elisa Fromont
Rémi Emonet
Taygun Kekec
Alain Trémeau
Damien Muselet

(a) msConvnet  (b) msAugLearner  (c) GroundTruth

[BMVC-2014]
Independant pixelwise classification

Figure by [Sermanet 2012]
Deep auto-context: Augmented learner

Context learner $f_c$:
- Learn to predict a whole patch of semantic labels.
- Integrate this patch as « features » into the direct learner $f_d$
- « Augmented learner »
- 1-in-K encoding (« hot one ») of semantic labels when used as input
# Scene labelling: results

<table>
<thead>
<tr>
<th>Learner</th>
<th>Scale</th>
<th>Stanford dataset</th>
<th>SIFT Flow Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct</td>
<td>single</td>
<td>69.7%</td>
<td>66.2%</td>
</tr>
<tr>
<td>Augmented</td>
<td>single</td>
<td>72.0%</td>
<td>66.2%</td>
</tr>
<tr>
<td>Direct</td>
<td>multi</td>
<td>75.7%</td>
<td>67.1%</td>
</tr>
<tr>
<td>Augmented</td>
<td>multi</td>
<td><strong>76.4%</strong></td>
<td><strong>68.5%</strong></td>
</tr>
</tbody>
</table>

Content

Semantic full scene labelling

Hand segmentation (pose estimation)
Hand gesture recognition

1st ranked at ECCV W. on Chalearn 2014
looking at people: gesture recognition

Estimation de la pose de la main

- A complex problem
  - Small images (according to distances between hand and sensor)
  - Large variation in hand poses
  - Real time is a challenge
- Our solution
  - Segmenting hands into parts
  - Structured deep learning
  - Semi supervised setting

PhD of Natalia Neverova
Collaboration with Graham W. Taylor
Guelph University, Canada
Results

Input: real depth videos from a Kinect sensor.
Synthetic training data

600,000 images rendered with 3D modeler (« Poser »)
Calculations distributed over 4 workstations (several weeks)
- **Testing:**
  - pure classification, no graphical model (speed!)

- **Training:**
  - Leverage structural information to get a loss on real data w/o groundtruth
Structured loss (1): local context

Loss generated from a context learner is calculated on a segmented output patch.

The context learner is trained on synthetic images.
Structured loss (2): global context

A single region is supposed to exist for each label.
Unconnected outlier regions are identified to generate loss.

Si pixel out of range

\[ Q_{glb}^+(\theta_d \mid y_d^{(i,j)}) = -F_{y_d}^{(i)} \log P \left( Y_d = y_d^{(i,j)} \mid x^{(i,j)}, \theta_d, \theta_c \right) \]

Else

\[ Q_{glb}^-(\theta_d \mid y_d^{(i,j)}) = -F_{\gamma}^{(i)} \log P \left( Y_d = \gamma x^{(i,j)} \mid x^{(i,j)}, \theta_d, \theta_c \right) \]
High resolution segmentation

(Classical) resolution reductions between layers. System resolution is kept high by keeping different shifts.
## Results: supervised vs. semi-supervised

On 50 manually annotated frames (real data)

<table>
<thead>
<tr>
<th>Training method</th>
<th>Training data</th>
<th>Test data</th>
<th>Accuracy (per pixel)</th>
<th>Accuracy (per class)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supervised</td>
<td>synth.</td>
<td>synth.</td>
<td>85.9%</td>
<td>78.5%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>real</td>
<td>47.2%</td>
<td>35.0%</td>
</tr>
<tr>
<td>Semi-supervised</td>
<td>all</td>
<td>synth.</td>
<td>75.5%</td>
<td>78.3%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>real</td>
<td><strong>50.5%</strong></td>
<td><strong>43.4%</strong></td>
</tr>
</tbody>
</table>

Average gain of a single update (stochastic gradient descent):

<table>
<thead>
<tr>
<th>Gain in % points</th>
<th>Local</th>
<th>Global</th>
<th>Loc+Glb</th>
<th>Supervised (w. Labels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No labels required</td>
<td></td>
<td></td>
<td></td>
<td>[ACCV 2014]</td>
</tr>
</tbody>
</table>
Results on real images
Register real images
Align unlabeled real images with a vast dataset of synthetic images in order to artificially create groundtruth labels.
Conclusion

- Deep learning is a powerful tool for feature and representation learning.
- Works best on large number of annotated training data.
- Including structure is a goal and a challenge:
  - Structure can translate into loss for unlabeled data.
  - Structure can add invariance.