Incremental Learning of Latent Structural SVM for Weakly Supervised Image Classification

Thibaut Durand (1), Nicolas Thome (1), Matthieu Cord (1), David Picard (2)

(1) Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6
(2) ETIS/ENSEA, University of Cergy-Pontoise, CNRS, UMR 8051

ICIP 2014
Outline

1. Context
2. Model
3. Experiments
Outline

1. Context
2. Model
3. Experiments
• Address the problem of weakly supervised object classification
• The goal is to predict the label of the image using object position as latent variable
• Training data only provides image-level annotation (presence/absence of each category)
Context

- Model the (unknown) object location using latent variables
- Desired output during test time: predicted image label
Learning a **joint model** for both localization and classification

- Widely-used approach:
  - Latent SVM (LSVM) [PAMI10]
  - Latent Structural SVM (LSSVM - extension to structured output) [ICML09]
- Excellent performances for detection tasks
- Performances for categorization are less impressive
- 2 limitations:
  - Computation is very demanding
  - Optimization problem is hard (non-convex)

[PAMI10: Felzenszwalb, Girshick, McAllester, Ramanan. *Object detection with discriminatively trained part based models*]

[ICML09: Yu, Joachims. *Learning structural svms with latent variables*]
• Popular approach for modeling object positions: sliding window
• Popular approach for modeling object positions: sliding window
Context

- Popular approach for modeling object positions: sliding window
• Popular approach for modeling object positions: sliding window
• Popular approach for modeling object positions: sliding window
- Popular approach for modeling object positions: sliding window
Popular approach for modeling object positions: sliding window
• Popular approach for modeling object positions: sliding window
Context

- Popular approach for modeling object positions: sliding window
Context

• Popular approach for modeling object positions: sliding window
• Popular approach for modeling object positions: sliding window
Contributions

• Propose an original evolution of the latent parameter space based on **cropping**
• Explore only some boxes at each iteration
• Speed-up training and inference
• Incremental Latent Structural SVM (ILSSVM)
Outline

1. Context
2. Model
3. Experiments
ILSSVM model

- Multi-class problem
- LSSVM formalism
- Training data: \( n \) labeled images \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \)
- \( x_i \) is an image
- \( y_i \in \mathcal{Y} = \{1, 2, \ldots, K\} \) is a label
- Latent variable \( h_i = (h_{i1}, h_{i2}, h_{i3}, h_{i4}) \) represents the bounding box of the predictive object location

![Diagram of an image with a bounding box labeled by latent variables](image-url)
ILSSVM model

- Evolution of the latent parameter space based on cropping
- Explore only some boxes at each iteration
- **Coarse to fine approach**
- Gradually remove the background
- Evolution w.r.t. the previous latent value

![Initialization](image1)

... t ...

![t](image2)

... t+4
ILSSVM model

Initialization ... t ... t+n
ILSSVM model

(a) no crop  
(b) crop left  
(c) crop right  
(d) crop top  
(e) crop down  
(f) crop 4 sides

**Figure:** Examples of possible cropping (blue boxes) for a current bounding box (red)
ILSSVM model

- Learn a LSSVM discriminant function of the form:

\[(y, h) = \arg \max_{y \in \mathcal{Y}, h \in \mathcal{H}} \langle w, \Psi(x_i, y, h) \rangle\]  \hspace{1cm} (1)

where \(\Psi(x_i, y, h)\) is the joint feature.

- Objective function at the iteration \(t\)

\[
P_t(w) = \frac{1}{2} \|w\|^2 + \frac{C}{n} \sum_{i=1}^{n} \left( \max_{(y, h) \in \mathcal{Y} \times \mathcal{H}_i^t} [\Delta(y_i, y) + \langle w, \Psi(x_i, y, h) \rangle] \right) 
- \frac{C}{n} \sum_{i=1}^{n} \max_{h \in \mathcal{H}_i^t} \langle w, \Psi(x_i, y_i, h) \rangle \] \hspace{1cm} (2)
ILSSVM model

1. Fast in inference:
   - 6 windows/image (sliding window > 1000)

2. Better generalization (curriculum learning)
   - Easy examples = large regions
   - Start with large regions, and gradually cropped these regions

3. No require knowledge on the size and the ratio of objects → adapt itself to objects.
Image representation

- Foreground-background feature representation
- Foreground region: spatial pyramid $1 \times 1, 3 \times 3 \rightarrow$ spatial structure of the object
- Background region $\rightarrow$ strong context for classification
- BoW models using SIFT descriptors

[ECCV12: Russakovsky, Lin, Yu, Fei-Fei. Object-centric spatial pooling for image classification]
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Image classification

- Use the same coarse-to-fine approach
- Start with a box initialized on the whole image
- Crop it until convergence
Outline

1. Context
2. Model
3. Experiments
Mammal dataset

(a) bison  
(b) deer  
(c) elephant  
(d) giraffe  
(e) llama  
(f) rhino

Figure: Images of the different categories of Mammal dataset
## Results

<table>
<thead>
<tr>
<th>split</th>
<th>SW (6 scales)</th>
<th>ILSSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22.58</td>
<td>12.90</td>
</tr>
<tr>
<td>2</td>
<td>29.03</td>
<td>25.81</td>
</tr>
<tr>
<td>3</td>
<td>22.58</td>
<td>22.58</td>
</tr>
<tr>
<td>4</td>
<td>16.13</td>
<td>25.81</td>
</tr>
<tr>
<td>5</td>
<td>45.16</td>
<td>38.71</td>
</tr>
<tr>
<td>6</td>
<td>25.81</td>
<td>22.58</td>
</tr>
<tr>
<td>7</td>
<td>35.48</td>
<td>16.13</td>
</tr>
<tr>
<td>8</td>
<td>25.81</td>
<td>12.90</td>
</tr>
<tr>
<td>9</td>
<td>35.48</td>
<td>22.58</td>
</tr>
<tr>
<td>10</td>
<td>35.48</td>
<td>32.26</td>
</tr>
<tr>
<td>mean</td>
<td>29.35 ± 8.53</td>
<td>23.23 ± 8.16</td>
</tr>
</tbody>
</table>

**Table:** Classification error for the 10 splits for multi-scale sliding window (SW) and our method (ILSSVM)
Computation time

<table>
<thead>
<tr>
<th>method</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>ILSSVM</td>
<td>1 h</td>
</tr>
<tr>
<td>one-scale sliding window</td>
<td>3 h</td>
</tr>
<tr>
<td>multi-scale sliding window (6 scales, 1 ratio)</td>
<td>30 h</td>
</tr>
<tr>
<td>multi-scale sliding window (6 scales, 6 ratios)</td>
<td>250 h</td>
</tr>
</tbody>
</table>

**Table:** Time Comparisons for the ten splits on 1 CPU
Parameter of evolution of the latent variables

- Influence of the crop step
- Step proportional to the maximum of the width or height

<table>
<thead>
<tr>
<th>crop step (%)</th>
<th>13</th>
<th>17</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>classification error (%)</td>
<td>25.81</td>
<td>23.87</td>
<td>23.23</td>
<td>23.55</td>
<td>25.16</td>
</tr>
</tbody>
</table>

Table: Evolution of classification error with respect to the crop step

- Robust to this parameter (small variation <3%)
- More robust than the scale in sliding window (variation > 20%)
Qualitative results of predicted boxes

Figure: Examples of predicted boxes for a step of 5% (left) and 20% (right) at different iterations. The green box is the final box.
Qualitative results of predicted boxes

Figure: Examples of predicted boxes
Conclusion

- Original **coarse-to-fine approach** for weakly supervised image classification based on Latent Structural SVM formulation
- **Small and incremental** latent parameter space
- Find better optimum with small computation time
Thank you for your attention!

Questions?

Thibaut Durand\textsuperscript{(1)} \hspace{1cm} thibaut.durand@lip6.fr
Nicolas Thome\textsuperscript{(1)} \hspace{1cm} nicolas.thome@lip6.fr
Matthieu Cord\textsuperscript{(1)} \hspace{1cm} matthieu.cord@lip6.fr
David Picard\textsuperscript{(2)} \hspace{1cm} picard@ensea.fr

\textsuperscript{(1)} Sorbonne Universités, UPMC Univ Paris 06, UMR 7606, LIP6
\textsuperscript{(2)} ETIS/ENSEA, University of Cergy-Pontoise, CNRS, UMR 8051

Java code available on demand