NEC Laboratories America
Relentless passion for innovation

Hierarchical Feature Pooling with Structure Learning: A new method for Pedestrian Detection

Xiaoyu Wang — NEC Labs America

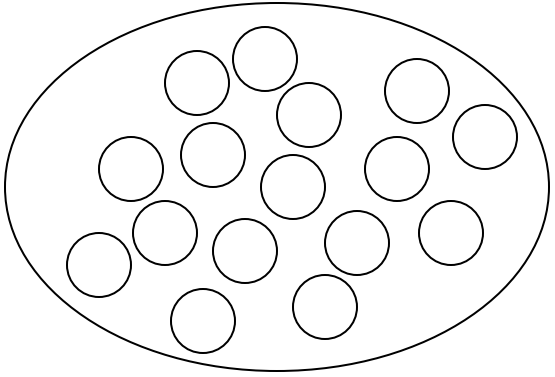
Liangliang Cao, Rogerio Feris, Ankur Data — IBM T.J. Watson Research Center

Tony X. Han — University of Missouri

Motivation

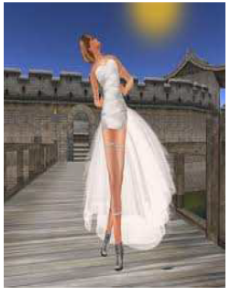
- Large intra-class variation in object detection
  - Deformation
  - Multiple viewpoints
  - Rich appearance
- Part-based models
  - Handling deformation very well
  - No scheme to deal with rich appearance

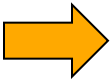
Our approach

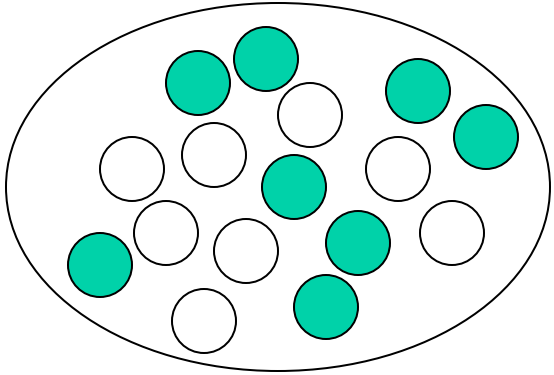


A collection of patch classifiers

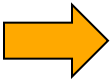
Our approach

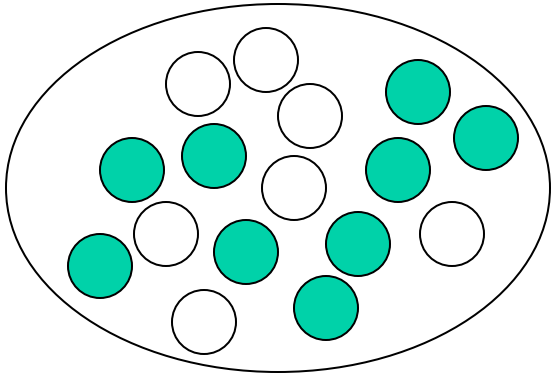




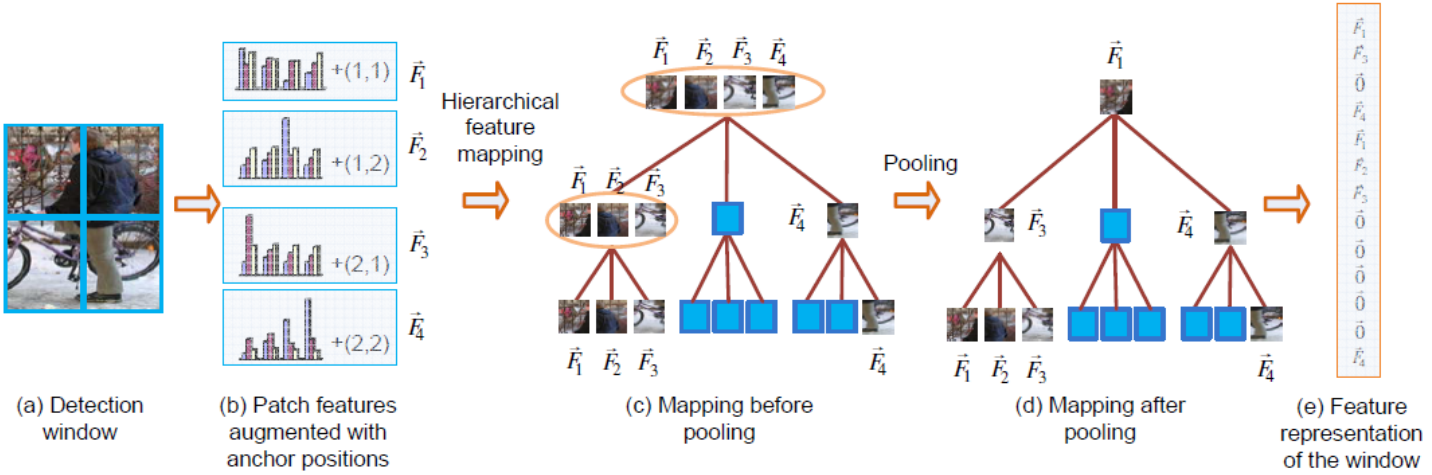


Our approach

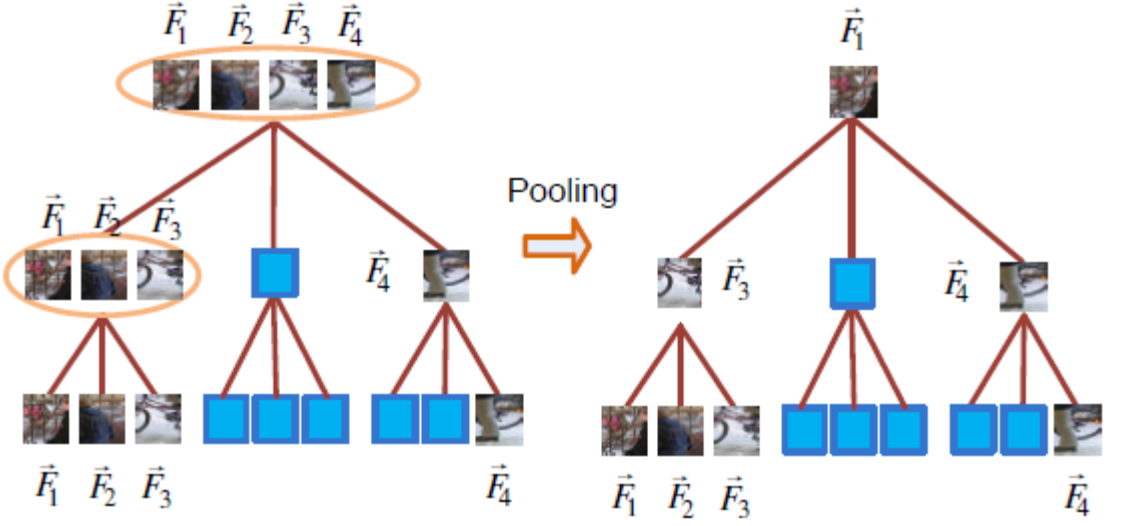




Framework


+(1,1)
+(1,2)
+(2,1)
+(2,2)
Hierarchical feature mapping
Pooling
(a) Detection window
(b) Patch features augmented with anchor positions
(c) Mapping before pooling
(d) Mapping after pooling
(e) Feature representation of the window

Our approach



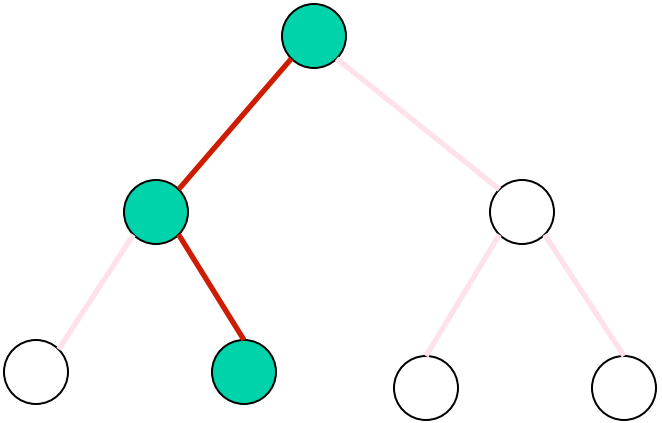
(c) Mapping before pooling

(d) Mapping after pooling

Key steps

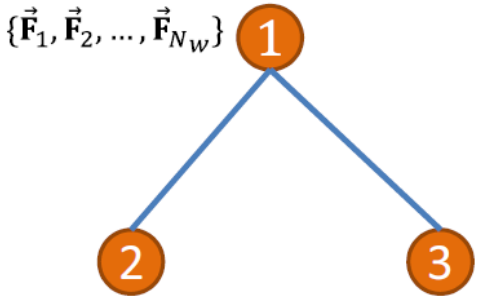
- Feature Mapping
  - Assign a patch to a node
- Pooling
  - Select one patch when multiple patches are assigned to the same node
  - Happen frequently in shallow layers

Key steps: Max-response mapping



Select the path which gives the maximum response

Key steps: Max-response pooling



$$1 \leftarrow \underset{\vec{\mathbf{F}}_j}{\operatorname{argmax}} \langle \vec{\mathbf{F}}_j, \vec{\mathbf{w}}_1 \rangle, \quad j = 1, \dots n$$

$$2 \leftarrow \underset{\vec{\mathbf{F}}_j}{\operatorname{argmax}} \langle \vec{\mathbf{F}}_j, \vec{\mathbf{w}}_2 \rangle, \quad j = 1, \dots k$$

$$3 \leftarrow \underset{\vec{\mathbf{F}}_j}{\operatorname{argmax}} \langle \vec{\mathbf{F}}_j, \vec{\mathbf{w}}_3 \rangle, \quad j = k+1, \dots n$$

Select the patch which gives the maximum response

Training: Formulation

$$\min_{w,\xi} \frac{\lambda}{2}||w||^2 + \frac{1}{N}\sum_{i=1}^{N}\xi_i$$

$$s.t. \max_{M_1}(\langle\Phi(x_i, y_i, M_1), w\rangle) - \max_{M_2}(\langle\Phi(x_i, \hat{y}_i, M_2), w\rangle)$$

$$\geq \triangle(y_i, \hat{y}_i) - \xi_i, \xi_i > 0,$$

Training: Implementation

Algorithm 1: Stochastic subgradient descent Structure SVM training with feature pooling

Input: training pairs $\{x_i, y_i\}_{i=1}^{N}, x_i \in \mathcal{X}, y_i \in \mathcal{Y}$; feature map $\Phi(x, y)$, loss function $\triangle(y, y')$, regularizer λ; number of iterations T, stepsizes η_t for t=1,...,T

1 **begin**

2 $w \leftarrow \vec{0}$

3 **for** *t=1,...,* T **do**

4 $(x_i, y_i) \leftarrow$ randomly chosen training example pair

5 $\hat{y}, \hat{M}_1, \hat{M}_2 \leftarrow \arg\max_{y \in \mathcal{Y}} \triangle(y_i, y) + \max_{M_1}(\langle w, \Phi(x_i, y, M_1)\rangle) - \max_{M_2}(\langle w, \Phi(x_i, y_i, M_2)\rangle)$

6 $w \leftarrow$

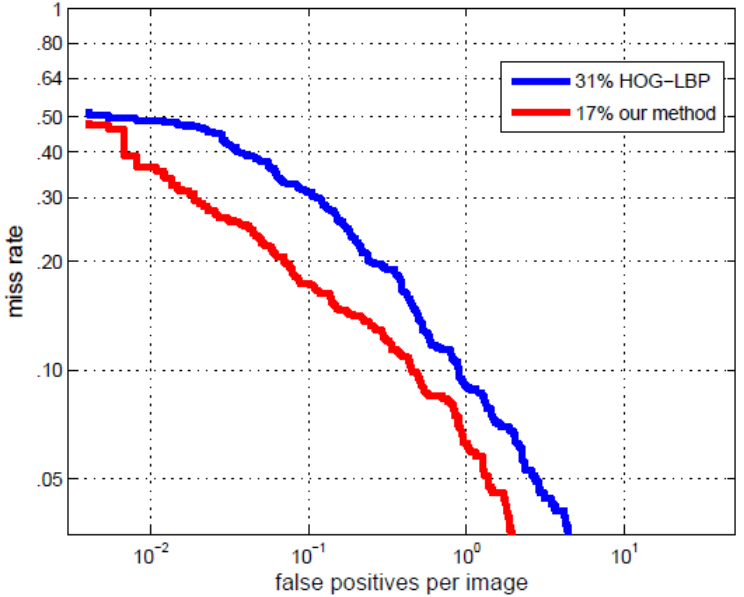
$w - \eta_t(w - \frac{1}{\lambda N}[\Phi(x_i, \hat{y}, \hat{M}_1) - \Phi(x_i, y_i, \hat{M}_2)]$

Output: prediction function $f(x) = \arg\max_{y \in \mathcal{Y}} \max_{M}(\langle w, \Phi(x, y, M)\rangle)$

Experiments

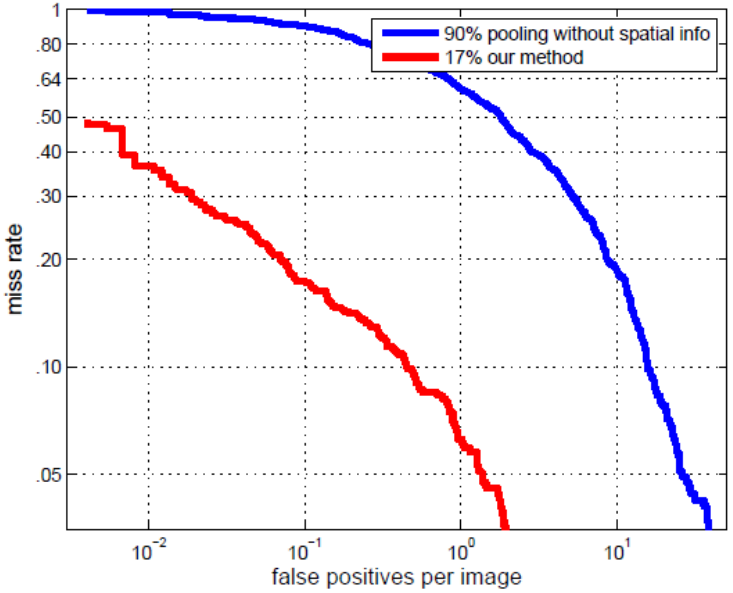
- Dataset: INRIA
- Evaluation: FPPI(False Positive Per Image)
- Baselines:
  - Detector without hierarchical mapping structure
  - Other state of the arts

Experiments



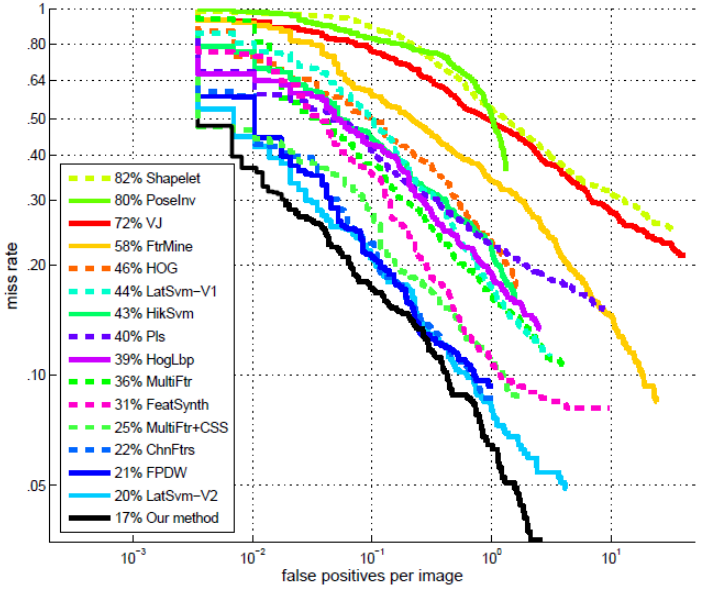
Performance comparison with the baseline

Experiments



Importance of spatial information

Experiments



Comparison with other methods

Conclusion

- An effective framework to handle large intra-class variations in pedestrian detection
- Efficiently trained by structure learning
- It produces promising results on the INRIA dataset

NEC Laboratories America

Relentless passion for innovation

Thank You!