Context-Aware CNNs for Person Head Detection

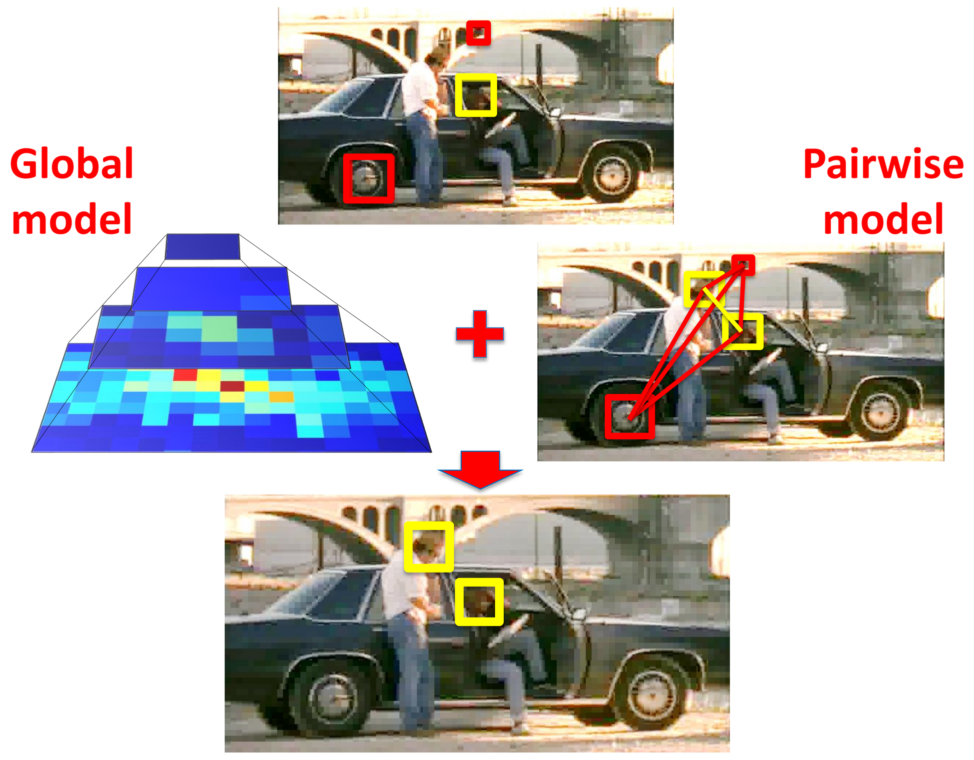
Tuan-Hung Vu, Anton Osokin, Ivan Laptev

INRIA/ENS, Paris, France

Willow

Goal

Improve region-proposal-based CNN [1] with contextual CNNs for human head detection



Contributions

- Propose two **context-aware CNN-based models:** Global and Pairwise models
- **HollywoodHeads dataset** with 369,846 head bounding-box annotations in 224,740 movie frames

Motivation

- For person detection, face detectors are insufficient and full/upper body detectors often fail in close-up views
- Success of Convolutional Neural Net in object detection
- Image context embeds constraints on the global and relative positions of objects in the image
- Local region-based models do not capture the context

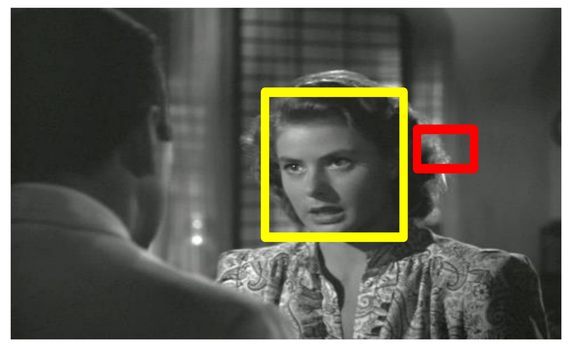
HollywoodHeads dataset

- Collected from 21 Hollywood movies of different genres from different time periods
- In total: 2,380 clips with 3,872 human tracks spanning over 224,740 frames
- Bounding-box annotation for heads on key frames
- Linear interpolation and manual verification on all frames
- Training: 216,719 frames from 15 movies; validation: 6,719 frames from 3 movies; test: 1,302 frames from 3 movies

Context-aware CNNs

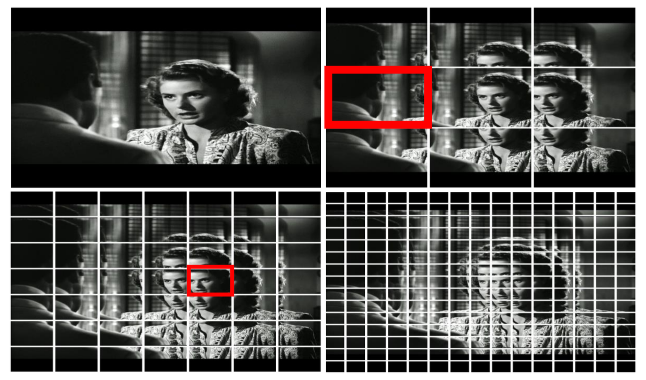
Local model

- CNN-based detector, trained on Selective Search object proposals (similar to R-CNN [1])
- Pre-training on ImageNet [2]
- Fine-tuning on HollywoodHeads dataset, minimizing the sum of independent log-losses using SGD.



Global model

- Predicts positions and scales of objects given the whole image as input
- The target is defined over a coarse **multi-scale grid** of image regions (cells)
- Label each cell as positive if its region has sufficient overlap with a ground-truth bounding box



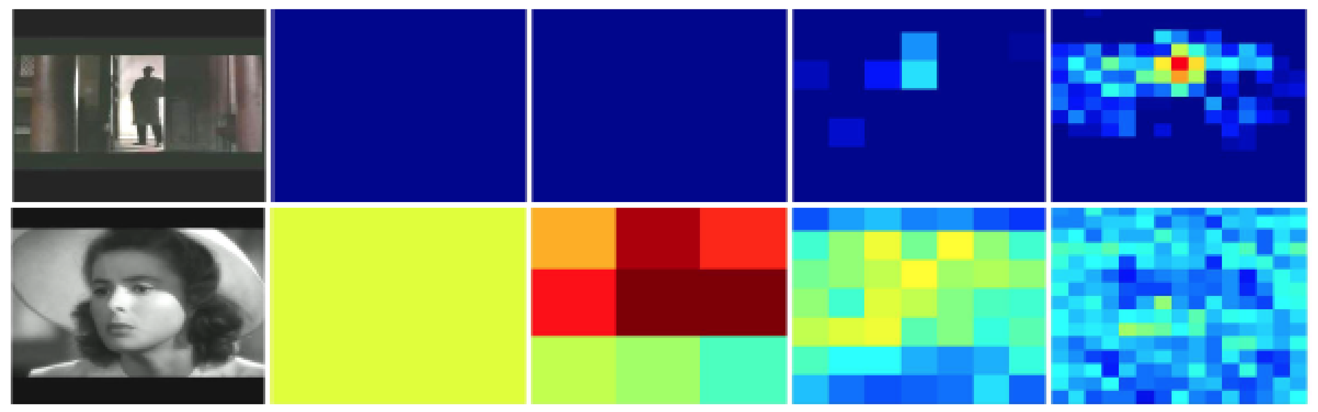
Multi-scale grids

- Training: minimizing the sum of C log-loss functions

$$\ell(\boldsymbol{f}_c(\mathbf{x}), y_c) = \sum_{y \in \{0,1\}} \log(1 + \exp((-1)^{y_c + y + 1} f_{c,y}(\mathbf{x})))$$

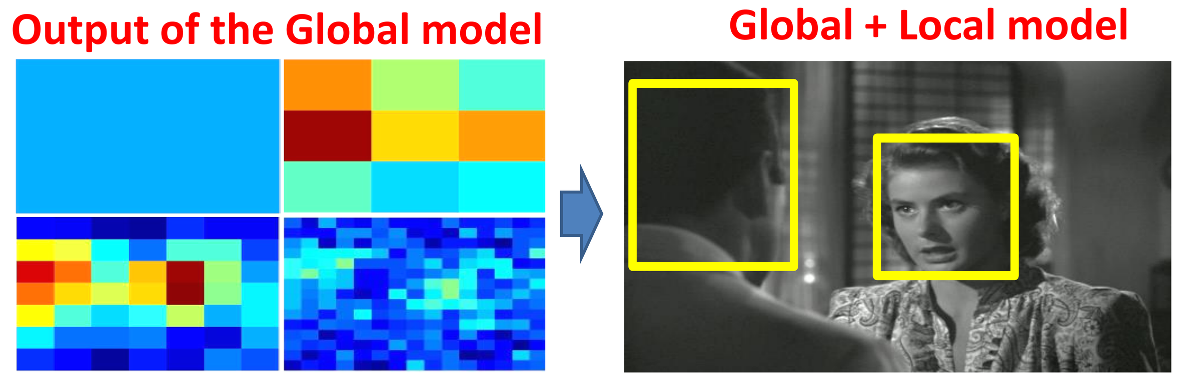
Here $y_c \in \{0, 1\}$ are ground-truth labels for cells $c \in \{1 \cdots \mathrm{C}\}$

- **Multi-scale grid** of confidence



- Combine the scores of the local and global model by matching object candidates with the grid cells of the global model.

Output of the Global model — Global + Local model



Pairwise model

Similar to [3], we construct the **joint score function** for object candidates in a given image:

$$S(\boldsymbol{y}; \boldsymbol{w}) = \underbrace{\sum_{i \in \mathcal{V}} \theta_i^U(y_i)}_{\text{Unary potential}} + \underbrace{\sum_{(i,j) \in \mathcal{E}} \theta_{ij}^P(y_i, y_j, \overset{\text{Pair cluster}}{k_{ij}})}_{\text{Pairwise potential}}$$

Here $\mathcal{V}$ is the set of all examined candidates, and $\boldsymbol{y} = (y_i)_{i \in \mathcal{V}}$ are the corresponding label assignments, $\boldsymbol{w}$ – trainable parameters

- For each candidate *i*, a **score** is computed as the difference of the max-marginals of the joint-score

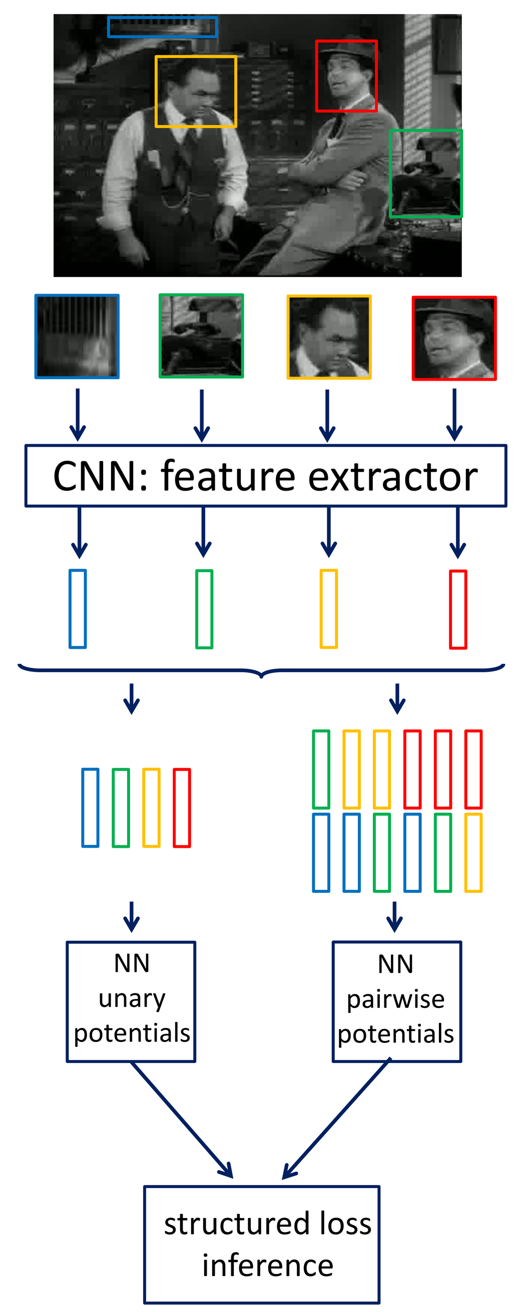
$$s_i(\boldsymbol{w}) = \max_{\boldsymbol{y}: y_i = 1} S(\boldsymbol{y}; \boldsymbol{w}) - \max_{\boldsymbol{y}: y_i = 0} S(\boldsymbol{y}; \boldsymbol{w})$$

- **Structured surrogate loss** – logistic loss on the structured scores

$$\ell(\boldsymbol{w}, \hat{\boldsymbol{y}}, \boldsymbol{x}) = \sum_{i: \hat{y}_i = 1} v(s_i(\boldsymbol{w})) + \sum_{i: \hat{y}_i = 0} v(-s_i(\boldsymbol{w}))$$

with $v(t) = \log(1 + \exp(-t))$

- Training step:
  1. Construct a set of candidates using local model
  2. Perform forward pass to compute potentials
  3. Perform inference to compute structured loss and its gradient
  4. Back-propagate the gradient

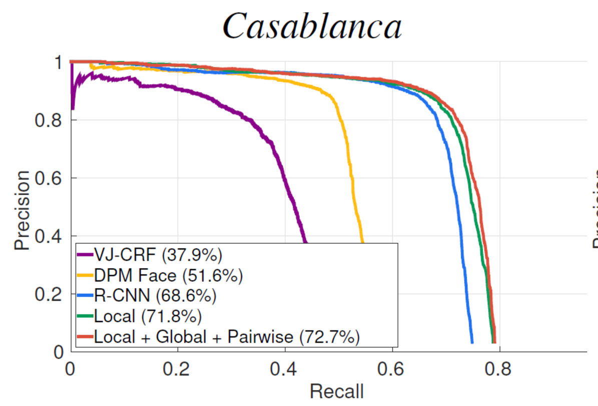


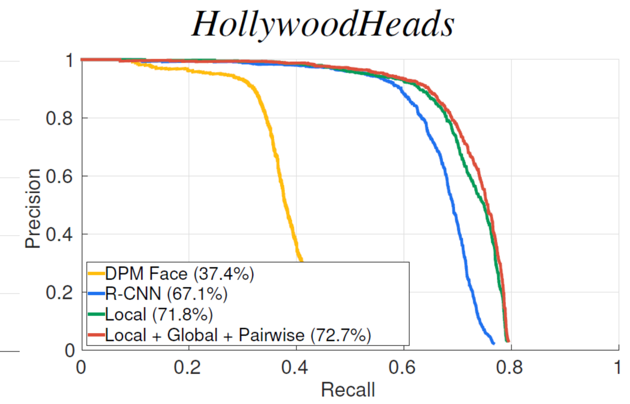
Results

We validate the method on the new **HollywoodHeads** dataset, **TVHI** dataset [4] and **Casablanca** dataset [5]. For each dataset we evaluate

- Local, Local+Global, Local+Pairwise, Local+Pairwise+Global models
- R-CNN detector [1] trained on Hollywood-Head dataset
- DPM Face detector [6]

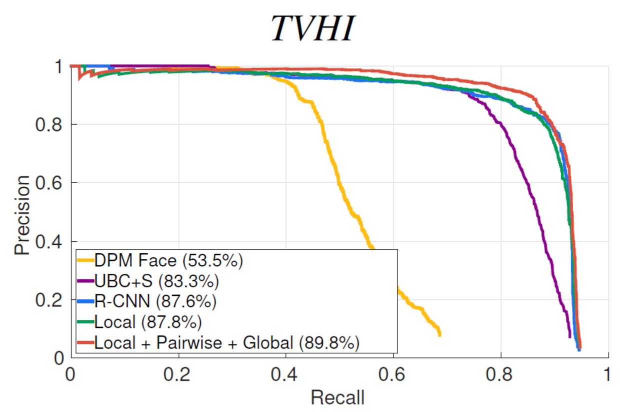
Test set	Local	Local Global	Local Pairwise	Local Pairwise Global
Casablanca	71.8	72.1	72.5	**72.7**
HH	71.8	72.5	71.9	**72.7**
TVHI	87.8	89.5	89.2	**89.8**





Training set size:

Test set	4 movies	8 movies	15 movies
Casablanca	51.2	62.5	**72.7**
HollywoodHeads	63.3	67.7	**72.7**
TVHI	88.6	88.8	**89.8**



Base architectures:

	AlexNet	Oquab	VGG-S	verydeep-16
AP	76.3	76.7	77.2	**78.5**
Train speed	**445**	284	147	30
Test speed	**1490**	980	510	74

Complexity reduction: performance with different candidate-left ratio after filtering using Global Model

% left	100	30	20	10	6	4
R-CNN	67.1	65.0	63.9	59.0	53.7	48.9
Local	71.8	68.3	66.8	60.2	53.4	48.8

Related work

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik: Rich feature hierarchies for accurate object detection and semantic segmentation. *In Proc. CVPR, 2014*

[2] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, Learning and transferring mid-level image representations using convolutional neural networks. *In Proc. CVPR, 2014.*

[3] C. Desai, D. Ramanan, and C. C. Fowlkes, Discriminative models for multi-class object layout. *IJCV, vol. 95, no. 11, pp. 1–12, 2011*.

[4] M. Hoai and A. Zisserman, Talking heads: Detecting hu- mans and recognizing their interactions. *In Proc. CVPR, 2014*

[5] X. Ren, Finding people in archive films through tracking. *In Proc. CVPR, 2008*

[6] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, Face detection without bells and whistles. *In Proc. ECCV, 2014*