Context-Aware CNNs for Person Head Detection
Tuan-Hung Vu, Anton Osokin, Ivan Laptev
INRIA/ENS, Paris, France

Goal

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

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

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

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

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

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
- DPM Face detector [6]

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

Related work

Training set size:
- Hollywood: 63.3% / 63.3% / 72.3%
- TVHI: 80.0% / 80.5%
- Casablanca:
  - 72.7% / 72.5%
  - 69.8% / 69.2%

Base architectures:

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

Output of the Global model

Global + Local model