Beyond detection: GANs and LSTMs to pay attention at human presence

Rita Cucchiara

Imagelab, Dipartimento di Ingegneria «Enzo Ferrari»
University of Modena e Reggio Emilia, Italy
Agenda

Beyond Human detection:

1) See humans
2) See what humans see

Use of GANs, Iterative and Recurrent neural architectures in Vision
Beyond (People) detection

✓ 10 years pedestrian detection [S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele CVPR2016] , about 70% accuracy on Caltech

✓ Many deep networks for pedestrian detection: CNNS+ handcraft feature: 9% miss rate on Caltech reasonable dataset *

✓ Object detector: SSD**, YOLO, YOLOv2***..
  YOLOv2 78.6% mAP on VOC2007-12 at 40fps

Still a margin of improvement..
CHALLENGES IN NEW ENVIRONMENTS

Real-time detection of people and AGVs in working areas on embedded NVIDIA boards at Imagelab

Standard networks are not enough. Embedded vision solutions with bckg sub and CNNs
If People Detection Solved...

GANS FOR UNDERSTANDING HUMAN PRESENCE UNDER EXTREME CONDITIONS

(thanks to Matteo Fabbri, and Simone Calderara)

thanks to PANASONIC
Now CNNs can classify more than 50 attributes

Problems with
• Low resolution
• Occlusions and self-occlusions
Generative Adversarial Networks

“..a generative model $G$ captures the data distribution, ..a discriminative model $D$ estimates the probability that a sample came from the training data rather than $G$.

The training procedure for $G$ is to maximize the probability of $D$ making a mistake”
[I.Goodfellow.. Y.Bengio 2014]

A **conditional generative** model $p(x | c)$ can be obtained by adding $c$ as input to both $G$ and $D$
**RAP:** A Richly Annotated Dataset for Pedestrian Attribute Recognition

Dataset dimension:
- 41,585 pedestrian samples
- 33,268 for training
- 8,317 for testing

Dataset image resolution:
- from 36x92 to 344x554
Generative Adversarial Network for De-occlusion (or Super-Resolution)

RAP

occRAP by Imagelab

lowRAP by Imagelab

Cross Entropy

Generator

Encoder

Decoder

Discriminator

Compare (SSE)

De-occluded (Fake)

Original image (Real)

occRAP

RAP
Selected Architecture

Encoder

Decoder

Generator

Discriminator

Input image

Classification
(fake or real)
RESULTS

De-occlusion

Super Resolution
The Complete Approach: De-occlusion and Super-resolution For Aspect Recognition

**Attribute Classification Network details**
- batch size: 8
- GPU: 1080ti
- Training time: 24 hours

**Reconstruction GAN (for deocclusion) details**
- batch size: 256
- GPU: 1080ti
- Training time: 48 hours

**Super Resolution GAN (for image resolution) details**
- batch size: 128
- GPU: 1080ti
- Training time: 72 hours
Attribute classification

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age 17-30</td>
<td>71.40</td>
<td>90.08</td>
</tr>
<tr>
<td>Age 31-45</td>
<td>81.16</td>
<td>83.10</td>
</tr>
<tr>
<td>Age Less 16</td>
<td>66.45</td>
<td>90.08</td>
</tr>
<tr>
<td>Backpack</td>
<td>80.09</td>
<td>69.17</td>
</tr>
<tr>
<td>Bald Head</td>
<td>60.87</td>
<td>84.67</td>
</tr>
<tr>
<td>Black Hair</td>
<td>96.87</td>
<td>97.31</td>
</tr>
<tr>
<td>Body Fat</td>
<td>64.27</td>
<td>74.96</td>
</tr>
<tr>
<td>Body Normal</td>
<td>88.12</td>
<td>79.37</td>
</tr>
<tr>
<td>Body Thin</td>
<td>85.74</td>
<td>63.16</td>
</tr>
<tr>
<td>Boots</td>
<td>82.43</td>
<td>68.04</td>
</tr>
<tr>
<td>Box</td>
<td>62.39</td>
<td>82.52</td>
</tr>
<tr>
<td>Calling</td>
<td>71.81</td>
<td>83.35</td>
</tr>
<tr>
<td>Carry By Arm</td>
<td>69.84</td>
<td>91.37</td>
</tr>
<tr>
<td>Carry By Hand</td>
<td>78.60</td>
<td>76.04</td>
</tr>
<tr>
<td>Casual Shoes</td>
<td>64.29</td>
<td>84.64</td>
</tr>
<tr>
<td>Clerk</td>
<td>70.12</td>
<td>85.83</td>
</tr>
<tr>
<td>Cloth Shoes</td>
<td>74.39</td>
<td>78.76</td>
</tr>
<tr>
<td>Cotton</td>
<td>72.95</td>
<td>81.19</td>
</tr>
<tr>
<td>Customer</td>
<td>98.68</td>
<td>97.24</td>
</tr>
<tr>
<td>Dress</td>
<td>89.91</td>
<td>70.83</td>
</tr>
<tr>
<td>Female</td>
<td>88.33</td>
<td>90.84</td>
</tr>
<tr>
<td>Gathering</td>
<td>77.20</td>
<td>67.20</td>
</tr>
<tr>
<td>Glasses</td>
<td>70.07</td>
<td>63.16</td>
</tr>
<tr>
<td>Hand Bag</td>
<td>66.76</td>
<td>85.04</td>
</tr>
<tr>
<td>Hand Trunk</td>
<td>79.21</td>
<td>88.51</td>
</tr>
<tr>
<td>Hat</td>
<td>63.33</td>
<td>62.48</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Holding</td>
<td>63.39</td>
<td>90.31</td>
</tr>
<tr>
<td>Jacket</td>
<td>83.30</td>
<td>78.25</td>
</tr>
<tr>
<td>Jeans</td>
<td>86.29</td>
<td>83.37</td>
</tr>
<tr>
<td>Leather Shoes</td>
<td>70.98</td>
<td>81.75</td>
</tr>
<tr>
<td>Long Trousers</td>
<td>95.46</td>
<td>89.78</td>
</tr>
<tr>
<td>Muffler</td>
<td>66.95</td>
<td>72.52</td>
</tr>
<tr>
<td>Other Attachment</td>
<td>70.41</td>
<td>81.06</td>
</tr>
<tr>
<td>Other Trousers</td>
<td>76.91</td>
<td>83.39</td>
</tr>
<tr>
<td>Paper Bag</td>
<td>79.94</td>
<td>65.09</td>
</tr>
<tr>
<td>Plastic Bag</td>
<td>81.18</td>
<td>64.97</td>
</tr>
<tr>
<td>Pulling</td>
<td>64.82</td>
<td>71.90</td>
</tr>
<tr>
<td>Pusing</td>
<td>84.05</td>
<td>74.52</td>
</tr>
<tr>
<td>Shirt</td>
<td>85.06</td>
<td>79.51</td>
</tr>
<tr>
<td>Short Skirt</td>
<td>85.88</td>
<td>67.43</td>
</tr>
<tr>
<td>Short Sleeve</td>
<td>86.23</td>
<td>64.14</td>
</tr>
<tr>
<td>Skirt</td>
<td>88.52</td>
<td>68.52</td>
</tr>
<tr>
<td>Sport Shoes</td>
<td>80.41</td>
<td>69.28</td>
</tr>
<tr>
<td>SS Bag</td>
<td>68.33</td>
<td>69.88</td>
</tr>
<tr>
<td>Suit Up</td>
<td>62.35</td>
<td>78.52</td>
</tr>
<tr>
<td>Sweater</td>
<td>81.49</td>
<td>90.87</td>
</tr>
<tr>
<td>Talking</td>
<td>76.16</td>
<td>80.43</td>
</tr>
<tr>
<td>Tight</td>
<td>79.44</td>
<td>76.75</td>
</tr>
<tr>
<td>Tight Trousers</td>
<td>82.36</td>
<td>86.88</td>
</tr>
<tr>
<td>T Shirt</td>
<td>72.73</td>
<td>67.61</td>
</tr>
<tr>
<td>Vest</td>
<td>76.97</td>
<td>89.09</td>
</tr>
<tr>
<td>Mean</td>
<td>76.96</td>
<td>78.72</td>
</tr>
</tbody>
</table>

More than 75% of precision and recall for 50 people attributes on RAP

Acceptable results for occluded shapes and good for low resolution shapes
If People Detection still not solved..
..without detection

TRACKING HUMANS IN THE WILD BY JUNCTIONS WITH CMP

(thanks to Fabio Lanzi, and Simone Calderara)
State-of-the-art: Recurrent Nets for object tracking

For long-term-tracking
- YOLO network for detection (fine tuned on PascalVOC)
- NVIDIA GTX1080 GPU 45fps (python TensorFlow)
- 70fps with precomputed YOLO features

Recurrence is provided by an LSTM

<table>
<thead>
<tr>
<th>Tracker</th>
<th>AUC</th>
<th>precision</th>
<th>speed (fps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLT [27]</td>
<td>0.384</td>
<td>0.490</td>
<td>8</td>
</tr>
<tr>
<td>STRUCK [8]</td>
<td>0.496</td>
<td>0.664</td>
<td>10</td>
</tr>
<tr>
<td>DRLT (ours)</td>
<td>0.543</td>
<td>0.635</td>
<td>45</td>
</tr>
<tr>
<td>DRLT-LSTM (ours)</td>
<td>0.543</td>
<td>0.635</td>
<td>270</td>
</tr>
</tbody>
</table>

Very fast. Still very low accuracy.

From [D. Zhang, H. Maei, X. Wang, Y.-F. Wang, Samsung, UCSB ArXiv 2017]
Recurrence with CPM

- **CPM Convolutional Pose Machines**: a sequence of Convolutional nets that repeatedly produce 2D belief maps for the location of interesting parts (human junctions).

- Belief map is a non-parametric encoding of the spatial uncertainty of location.

- CPM learns implicit relationships between parts.

- It is not **recurrent but a multi-stage** network, trained with backpropagation.

* [S-E Wei, V.Ramakrishna, T.Kanade, Y.Sheickh «Convolutional Pose Machines» CVPR 2016]
Without detection: Temporal CPM3

Imagelab: tracking multiple body parts with **T-CPM** (Temporal Convolutional Pose Machines).

An iterative network (CPM) for predicting:

- the position of joints (H)
- their mutual association in space (P)
- their association in time (T)
Three Branches: Heatmaps, PAFs and TAFs

- **Heatmap** models the part locations as gaussian peaks in the map; 1 for each joint (“nose”, “neck”, “left-shoulder”).

- **PAFs: (Part Affinity Field)** to assemble the detected joints. The score of a candidate limb is proportional to the alignment with the PAF associated with that type of limb.

- **TAFs: (Temporal Affinity Field)** to link the corresponding joints of the same person in consecutive frames (for an unknown number of people).
Visual Example
How to provide initial annotation?

- Photorealistic
- Plausible dynamics
- Lifelike entity AI

- Access to native GTA functions
- Customizable
- Extract all the information available to the game engine

ScriptHook Library
The Deep architecture and the software is propriety of ImageLab UNIMORE.
We thanks Jump project funded with EU ER-FESR-2015-2020 program
For Tracking, Action, Behavior Recognition

T-CPMs do not use recurrence but works on sequences of frames and refines with iteration with long convolutional layers
- Problems of vanishing gradient
- Long-Short Term Memory architectures can give solution for time iterations, but not for long time sequences
If target Detection is not required..

SALIENCY DETECTION WITH LSTMS
SAM ARCHITECTURE

(thanks to Marcella Cornia, Giuseppe Serra and Lorenzo Baraldi)
SALIENCY DETECTION @Imagelab SAM

mit saliency benchmark

MIT300 (Itti, Torralba et al) more than 70 competitors since 2014
SALICON (Jiang et al 2015), 10000 images;


Saliency Attentive Model (SAM): ML-NET+ LSTMs
Number of images: 20,000
- 10,000 training images
- 5,000 validation images
- 5,000 test images

GPU: NVIDIA K80 on Supercomputer GALILEO CINECA

Training Time: ~15 hours

Winner of the competition LSUN Challenge CVPR 2017
Groundtruth

Actions in the Eye (Hollywood2) dataset

SAM
Saliency in task-driven video

Bottom-up saliency, detected by ML-NET, trained on SALICON on DR(EYE)VE dataset

http://imagelab.ing.unimore.it/dreyeve

Saliency not driven by a task...

*as a passenger sees*

Saliency trained by driving

*as a driver sees*
SIFT-BASED REGISTRATION FRAME BY FRAME

Collected with SMI ETG 2w, Frontal camera 720p/30fps + Eye pupils cameras at 60fps
GARMIN VirvX, 1080p/25fps + GPS.
Some conclusion (if any)

✓ Computer vision now is a Deep Learning based discipline
✓ Computer vision systems cannot be built without GPUs (both in training and at run-time)
✓ Conv-Nets are fundamental bricks to new architectures
✓ Autoencoders: for image generation
✓ (Conditional) Generative Adversarial Networks: for low-resolution occluded attribute recognition
✓ Multi-layers convolutional networks for emulating recurrency as T-CPM3 for tracking
✓ Recurrent and Long Short Term Memories for short time analysis: saliency and video captioning
✓ ...

Computer Vision

Deep Architectures

GPUs
Thank you!

rita.cucchiara@unimore.it
http://imagelab.ing.unimore.it