A 58.6mW Real-Time Programmable Object Detector with Multi-Scale Multi-Object Support Using Deformable Parts Model on 1920x1080 Video at 30fps

Amr Suleiman, Zhengdong Zhang, and Vivienne Sze

Massachusetts Institute of Technology

Why Object Detection?

UAV/Drones

Surveillance/Recognition

ADAS

Self-driving cars

Object Detection System Requirements

High Image Resolution

Real-time & Low Latency

Low Power

Outline

- Detection with Deformable Parts Models (DPM)
- Chip Architecture
- Main Contributions
- Chip Specifications and Comparisons
- Summary

General Object Detection Methodology

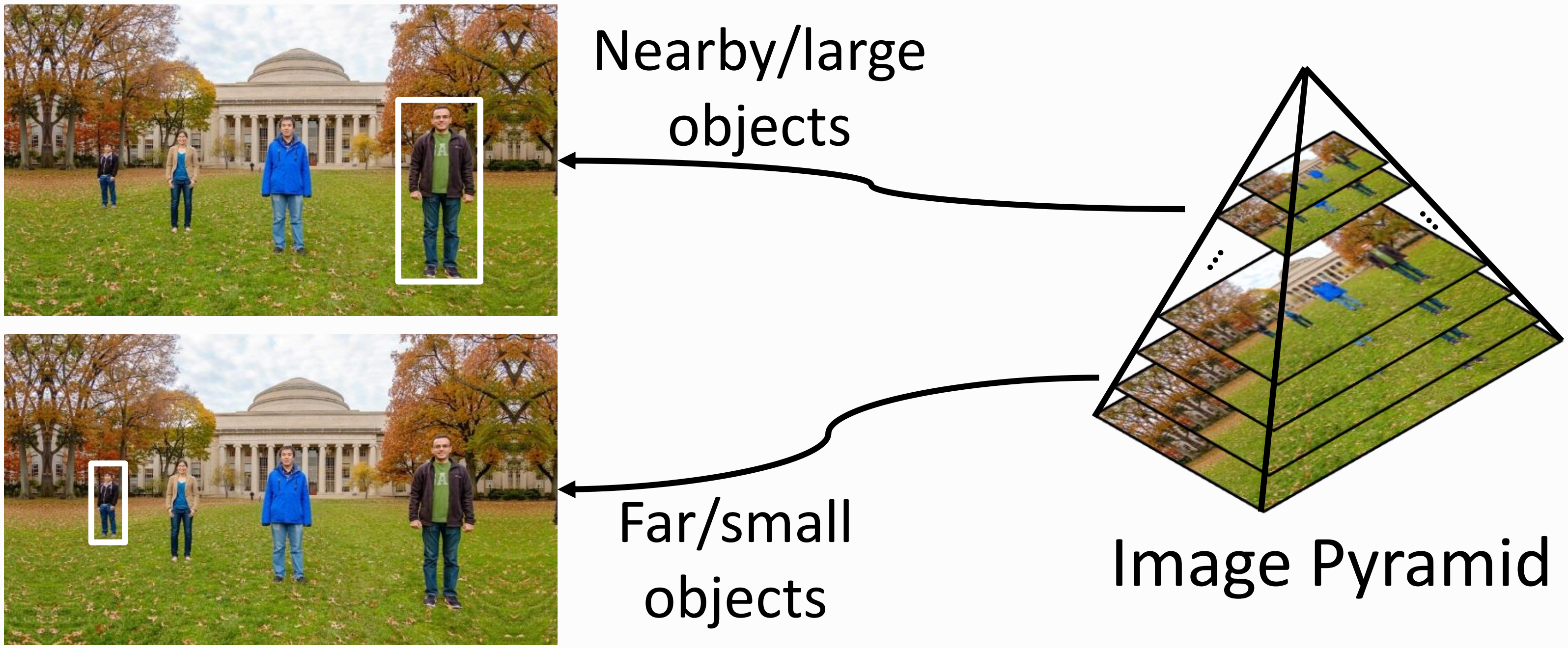
Localization
(Where?)

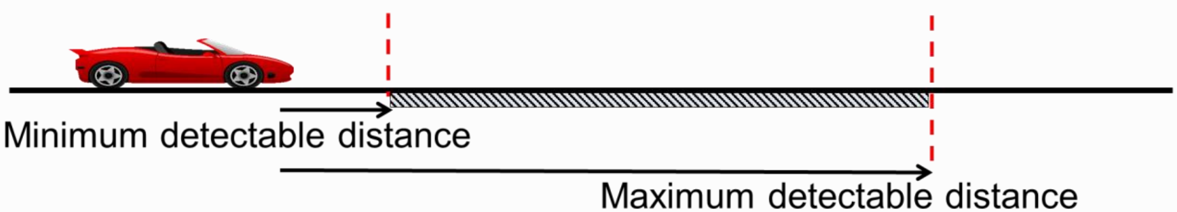
Classification
(True or False?)

Localization: 3D Search

Localization (Where?)

Classification (True or False?)



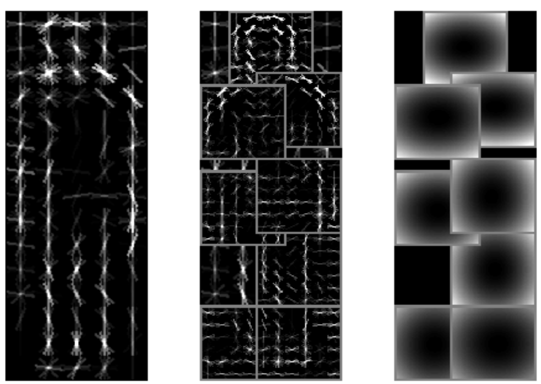


Classification with DPM Templates

Localization
(Where?)

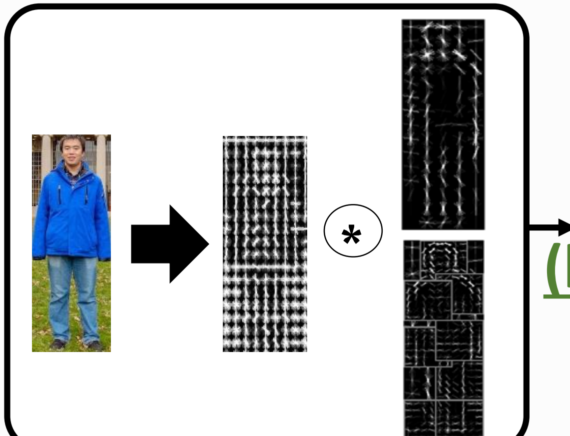
Classification
(True or False?)

Trained **DPM** templates
on HOG feature images

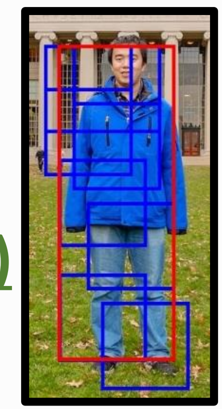


Root Parts Deform

Dot product

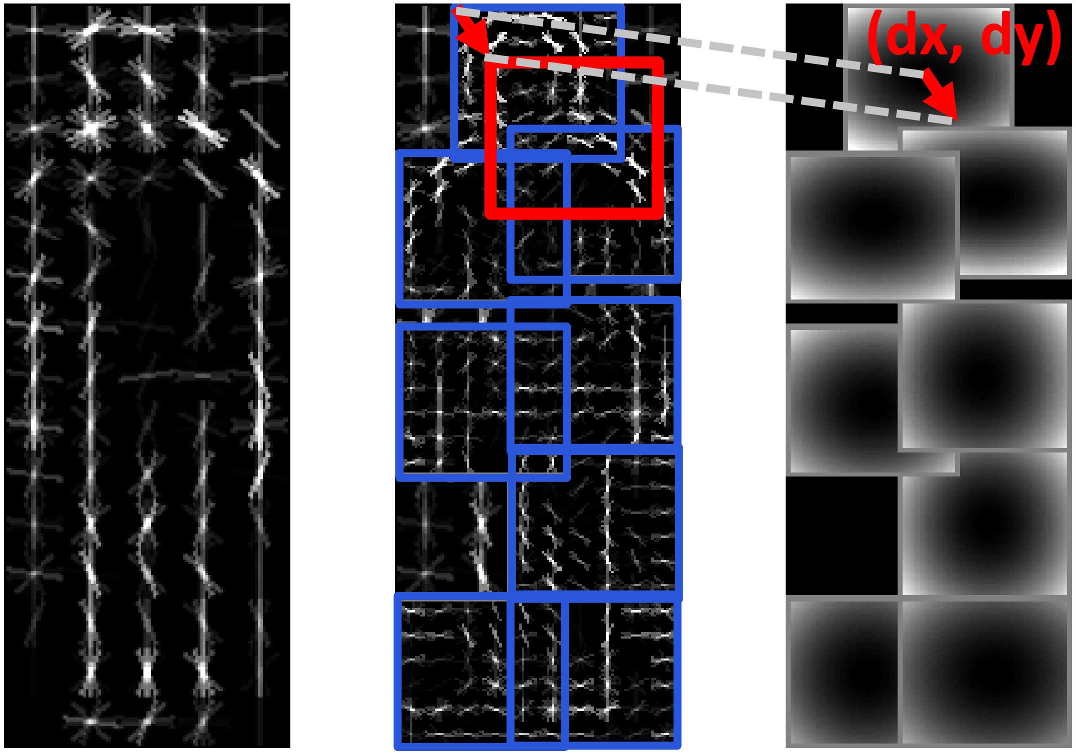


Person
(high score)



<u>HOG</u>: Histogram of Oriented Gradients

How Does DPM Work?



Root | Parts | Deformation Costs

Anchor Position

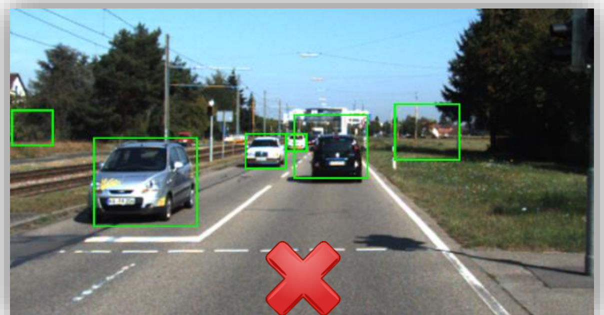
Search Candidate

$\text{DeformCost}_i\,(dx,dy)$
$= a_1 dx^2 + a_2 dx + a_3 dy^2 + a_4 dy$

small penalty — large penalty

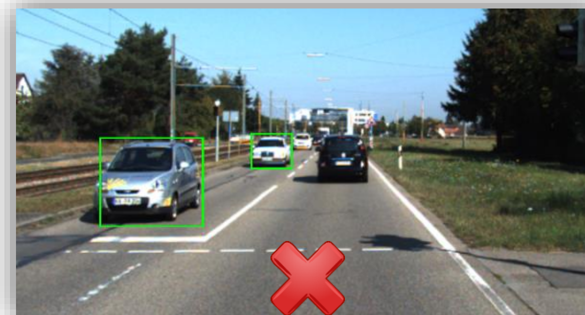
$$\text{DPM Score} = \text{RootScore} + \sum_{i=1}^{8} \max_{dx,dy}\left(\text{PartScore}_i(dx, dy) - \text{DeformCost}_i(dx, dy)\right)$$

Detection Accuracy



High recall
Low precision

High recall
High precision

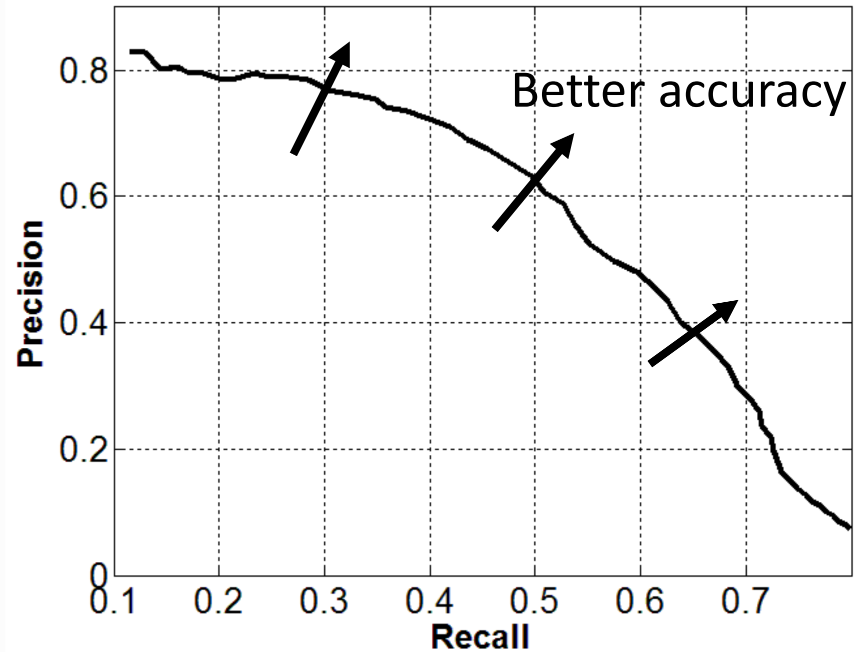


Low recall
High precision

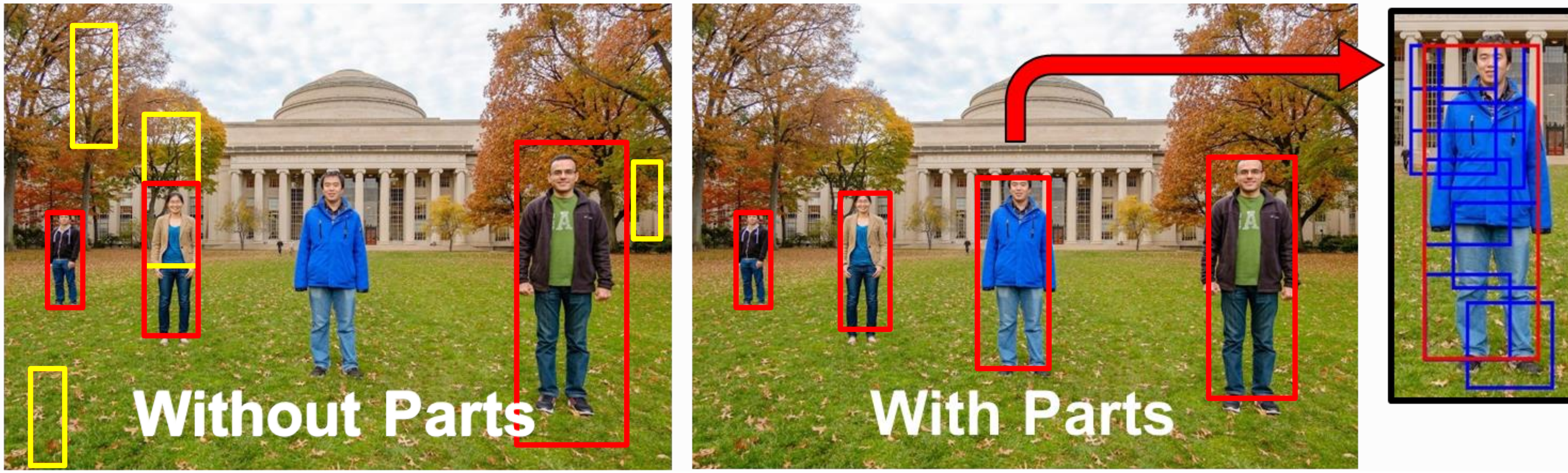
Average precision (AP) =

Area under curve

(0 ≤ AP ≤ 1)



Deformable Parts are More Accurate



Detecting parts enhances the accuracy by 2x

Measured on INRIA person dataset*

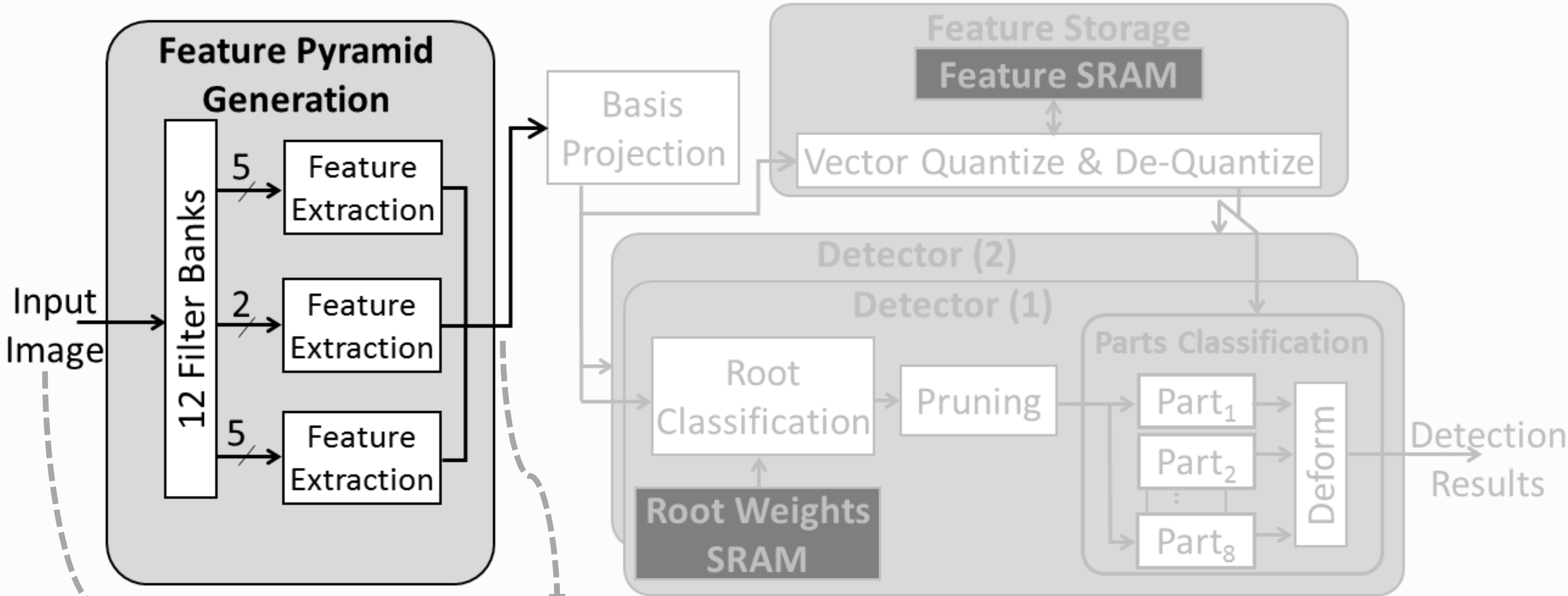
Challenge: DPM has **35x** more computation compared to without parts (rigid body) detection

*[http://pascal.inrialpes.fr/data/human/]

Outline

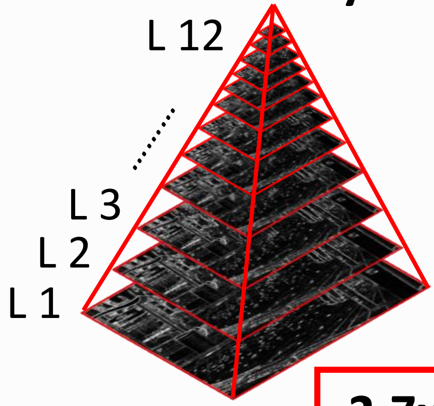
- Detection with Deformable Parts Models (DPM)
- Chip Architecture
- Main Contributions
- Chip Specifications and Comparisons
- Summary

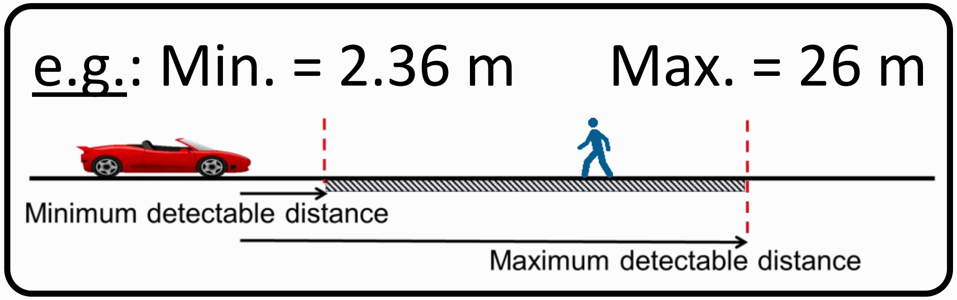
12-level Feature Pyramid



1920x1080 image

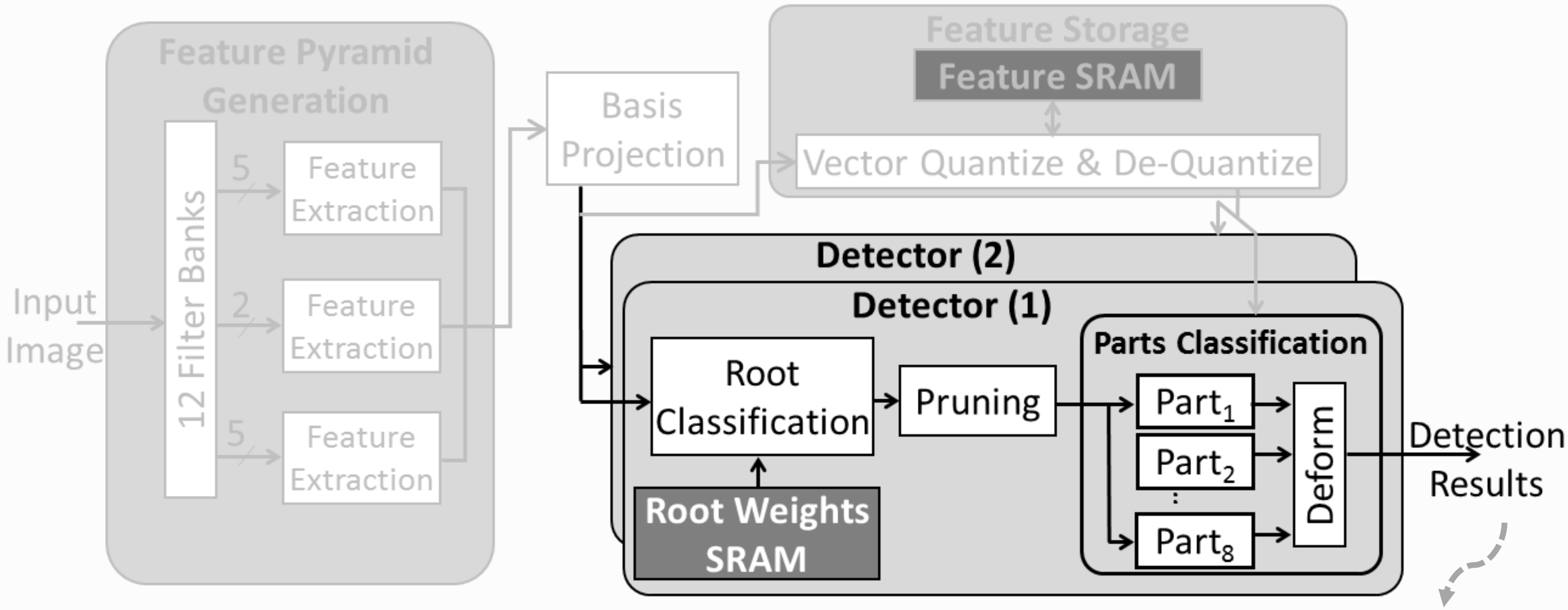
HOG Features Pyramid





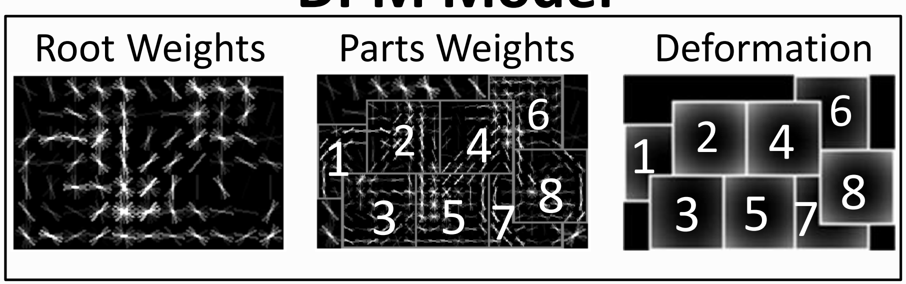
2.7x data expansion for HD (1920x1080) images

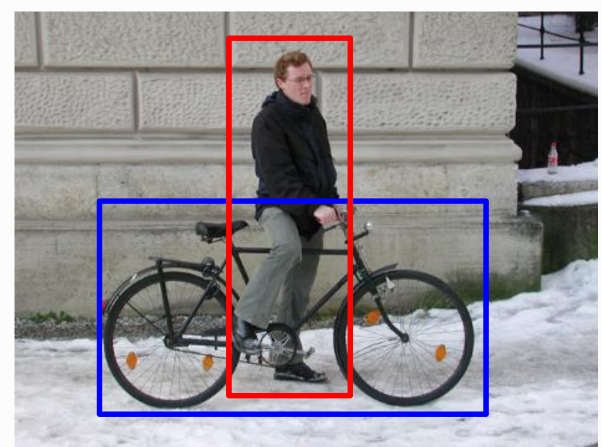
2 Programmable Detectors



Location, size, score

DPM Model



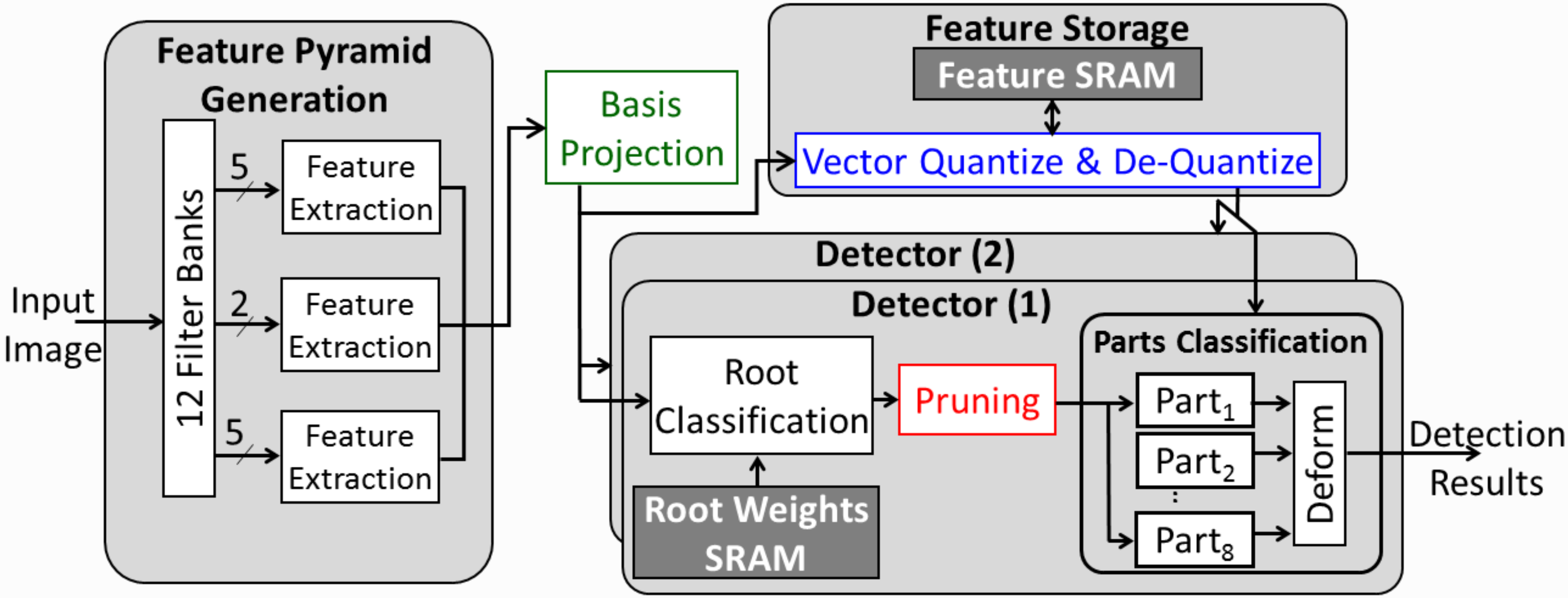


Programmable DPM model with a maximum template size of **128x128 pixels**

Outline

- Detection with Deformable Parts Models (DPM)
- Chip Architecture
- Main Contributions
- Chip Specifications and Comparisons
- Summary

Optimizations for Energy Efficiency



<u>Goal:</u> Reducing the parts classification overhead

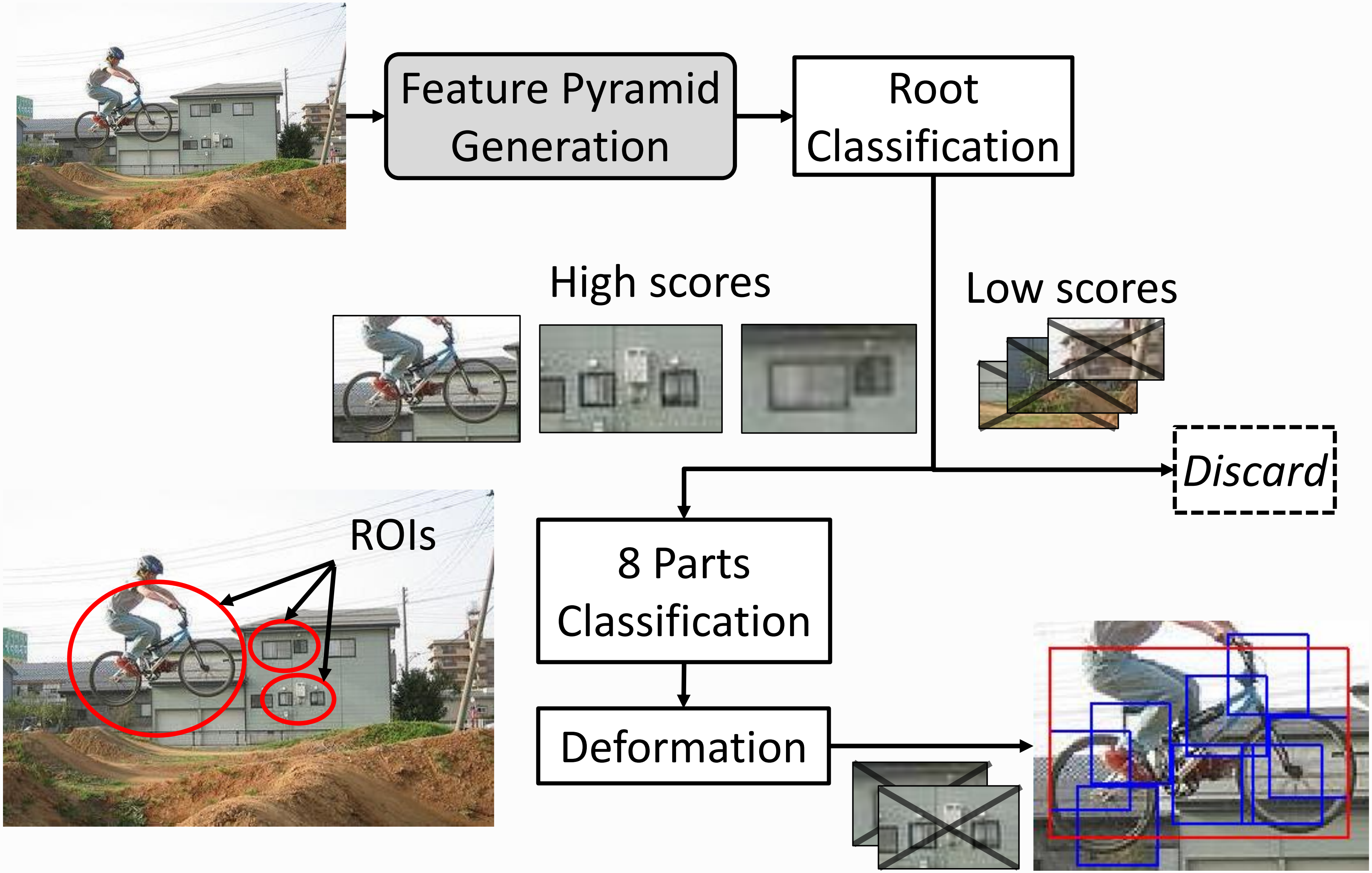
<u>Methods:</u>

1) Reduce the number of classifications
   (**Pruning** & **Vector Quantization**)
2) Reduce the cost of each classification
   (**Basis Projection**)

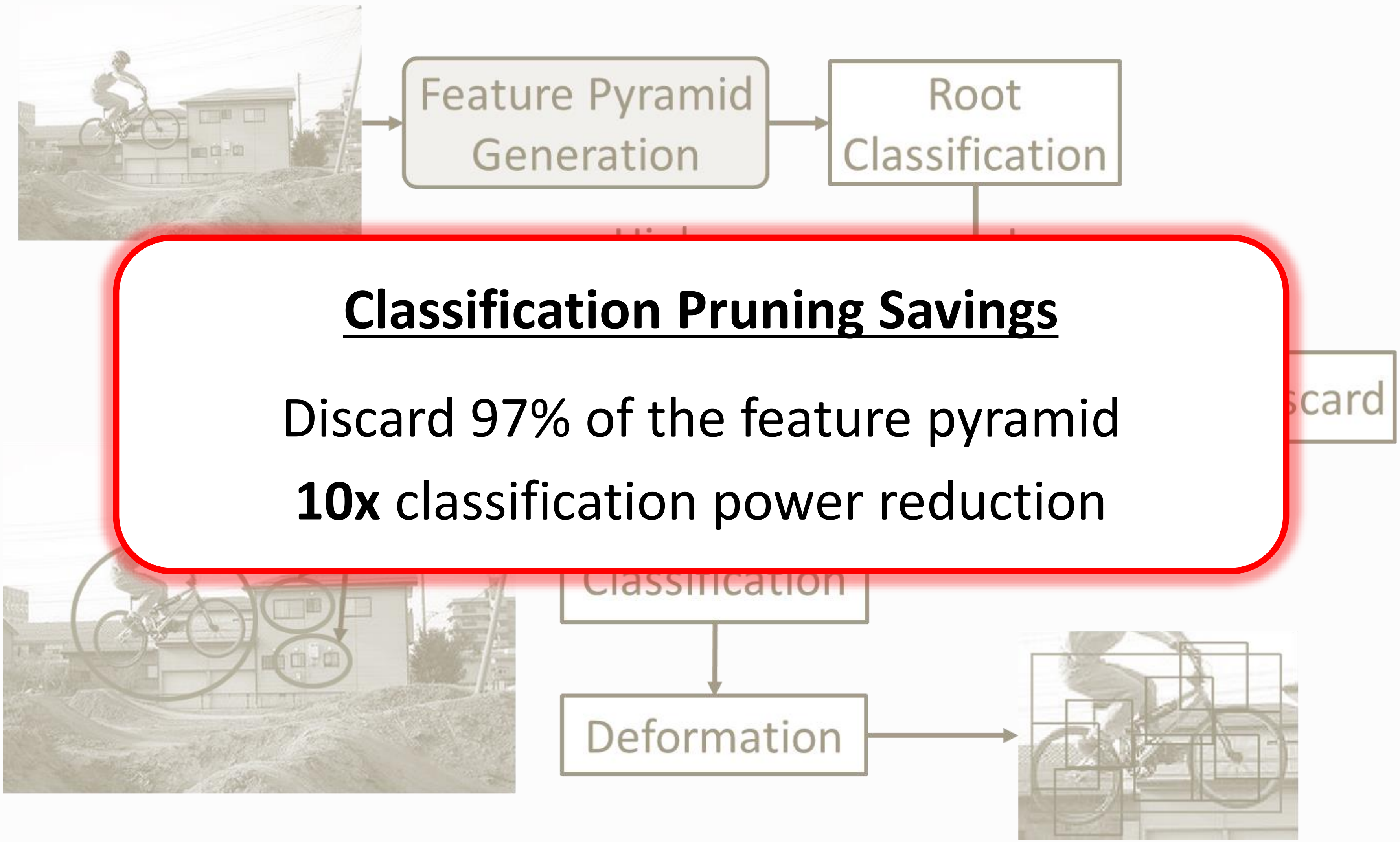
Method 1

Reduce the number of classifications

Parts Classification in Region of Interests



Parts Classification in Region of Interests



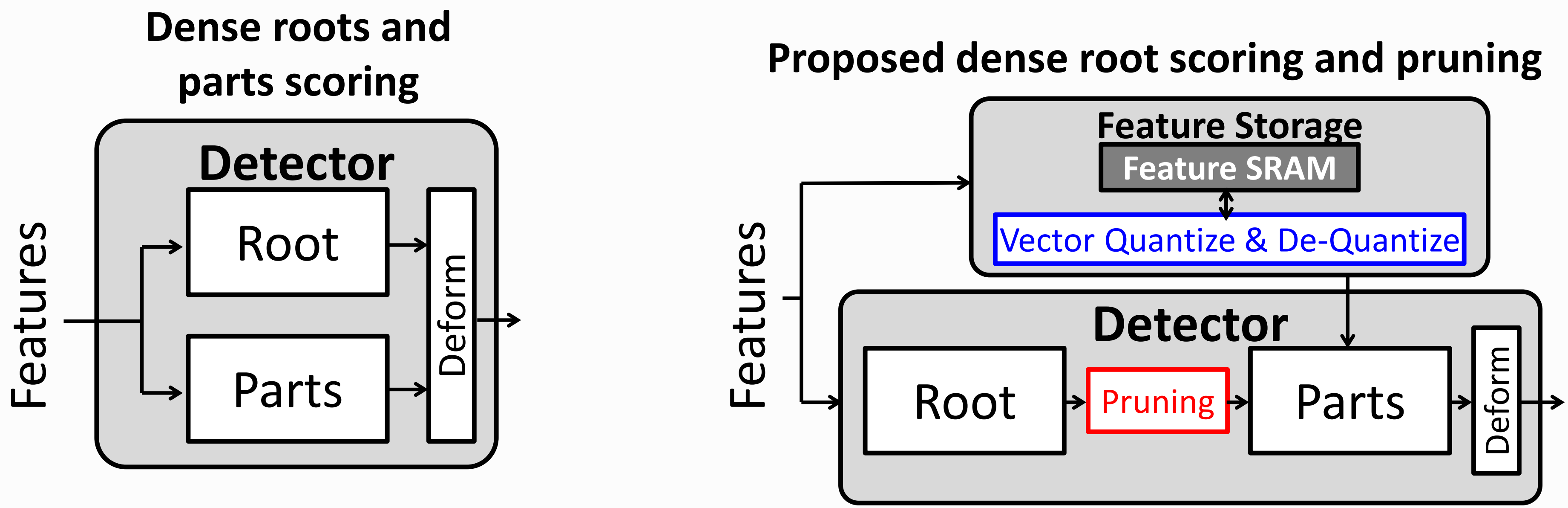
Classification Pruning Savings

Discard 97% of the feature pyramid

10x classification power reduction

Feature Storage for Parts Classification

- Store features for reuse by parts to avoid re-computation



Vector Quantization

16x reduction in memory size (520 KB vs. 32 KB)

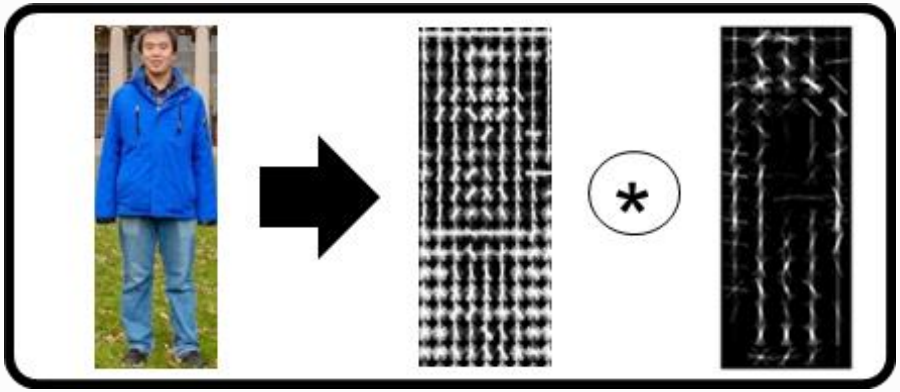
2x reduction in overall chip area

Method 2

Reduce the cost of each classification

Multiplication by Zero Can be Skipped

Classification = Dot product



Dot product → 3 K multiplications
HD image → 88 M multiplications
HD pyramid → 235 M multiplications

With more zero weights:

- Fewer multiplications
- Smaller weights memory size and BW

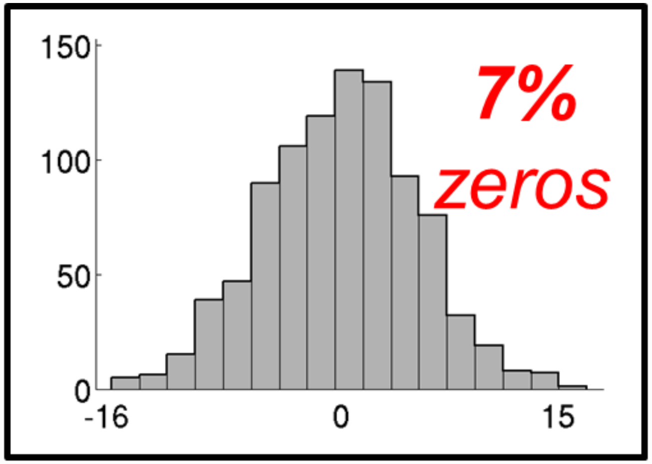
Project the Classification to a Sparse Space

$$\langle \boldsymbol{H}, \boldsymbol{W} \rangle = \left\langle H, \sum_d \alpha_d S_d \right\rangle = \sum_d \alpha_d \langle H, S_d \rangle = \sum_d \alpha_d P_d = \langle \boldsymbol{P}, \boldsymbol{\alpha} \rangle$$

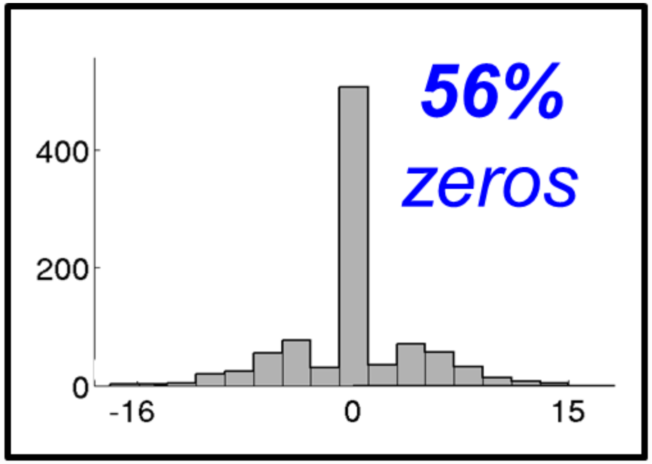
H: HOG features **W:** Template weights

S: Basis vectors

P: Projected features **α:** Projected weights



Histogram of Templates Weights (W)



Histogram of Projected Weights (α)

Project the Classification to a Sparse Space

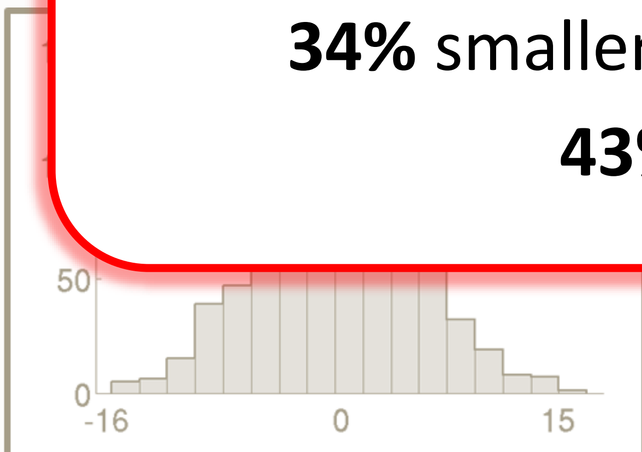
$$\langle \mathbf{H}, \mathbf{W} \rangle = \left\langle H, \sum_d \alpha_d S_d \right\rangle = \sum_d \alpha_d \langle H, S_d \rangle = \sum_d \alpha_d P_d = \langle \mathbf{P}, \boldsymbol{\alpha} \rangle$$

<u>Feature Basis Projection Savings</u>

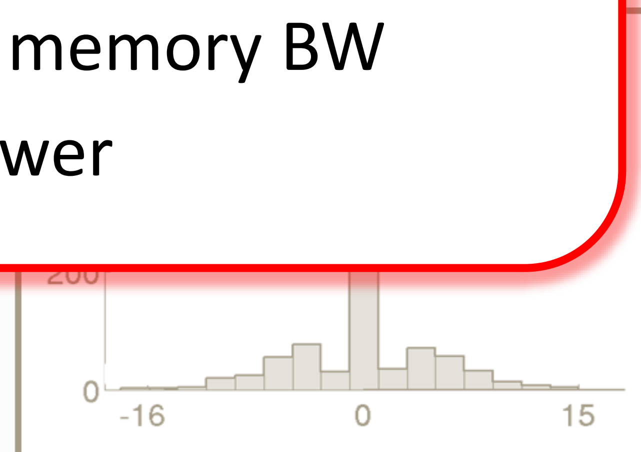
56% fewer multiplications

34% smaller weights memory BW

43% less power

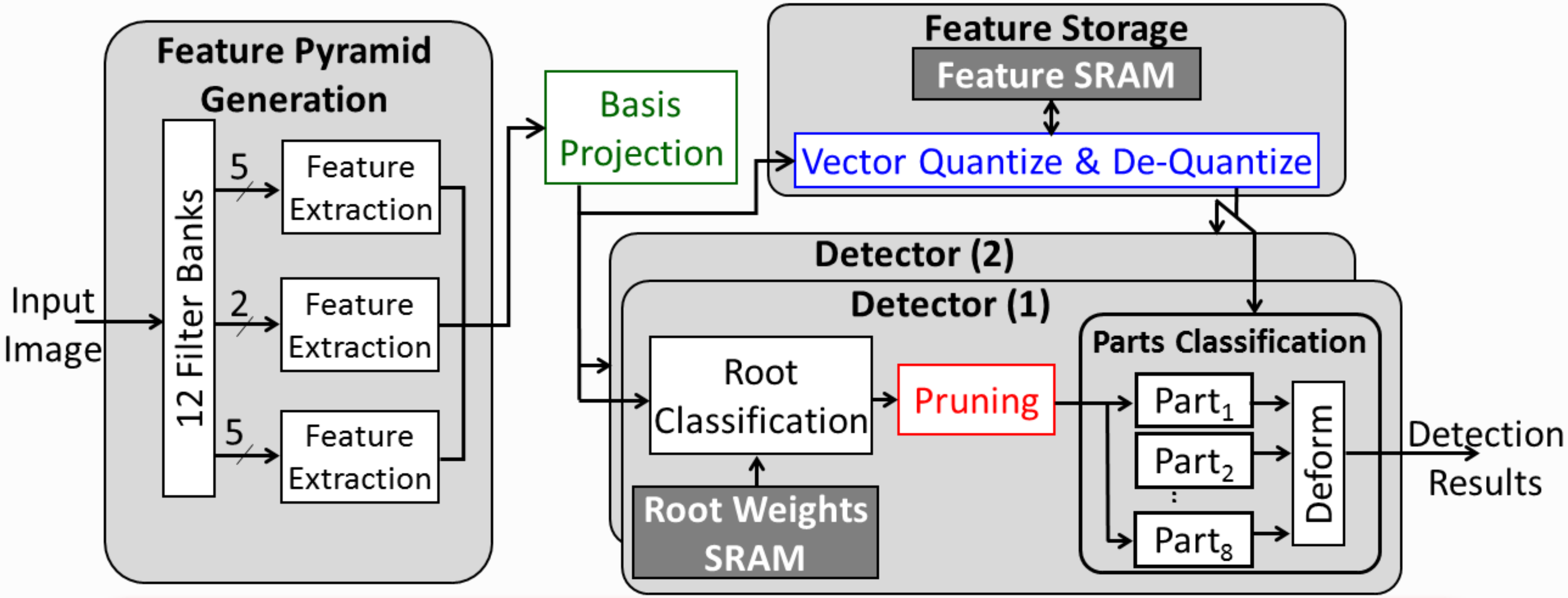


Histogram of Templates Weights (W)



Histogram of Projected Weights (α)

Overall Optimizations Savings



Total 2-Detector System Savings

5x less power

3.6x smaller memory

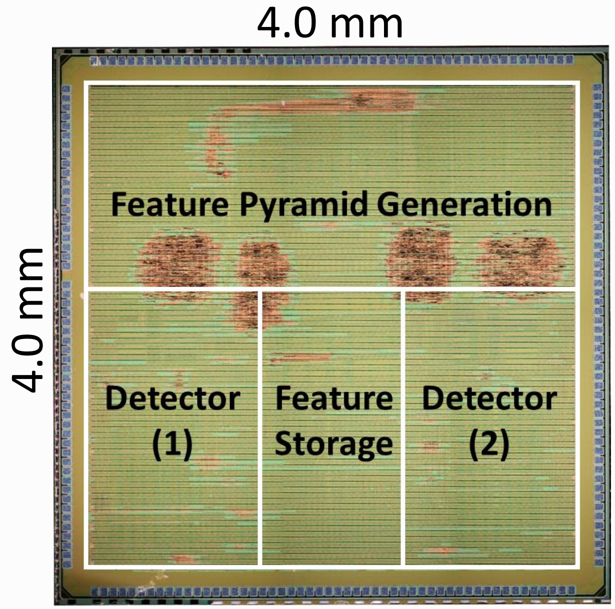
Detection Accuracy (mAP)* : Δ = -4.8%

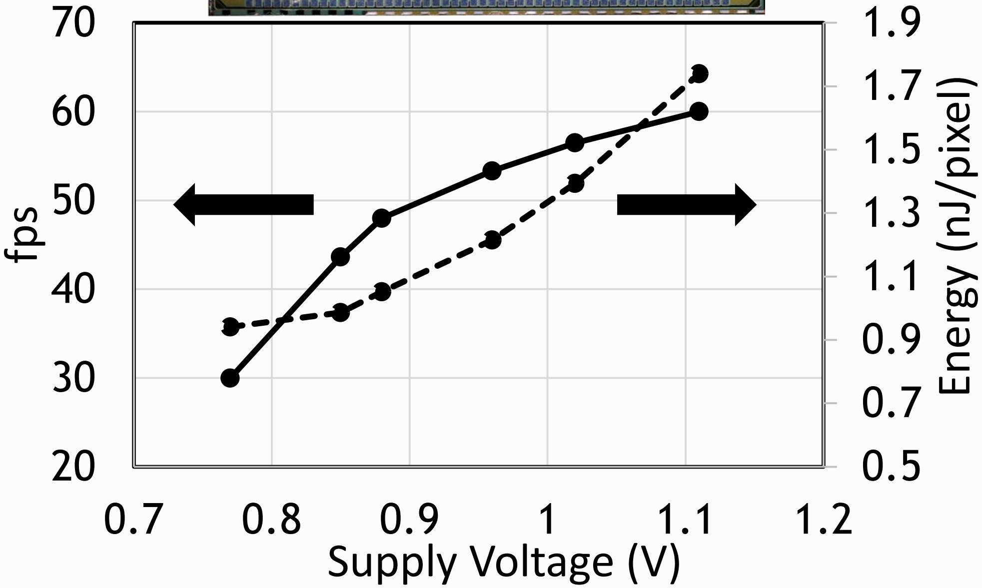
*mAP: mean Average Precision, on PASCAL VOC2007 with 20 classes

Outline

- Detection with Deformable Parts Models (DPM)
- Chip Architecture
- Main Contributions
- Chip Specifications and Comparisons
- Summary

Chip Die Photo and Specifications

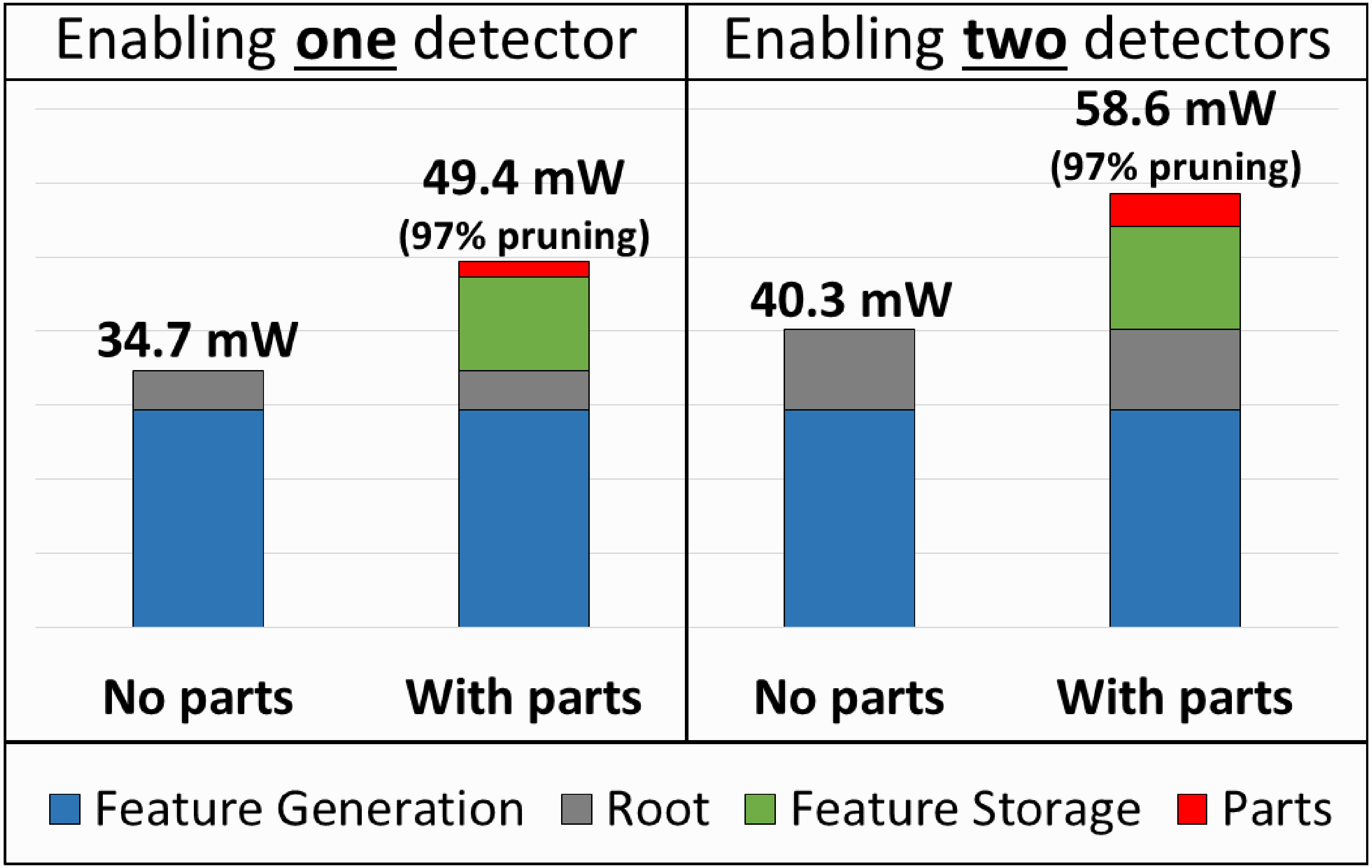




Technology	65nm CMOS
Chip size	4.0 x 4.0 mm^2
Logic gates	3283 kgates
SRAM	280.1 KB
Supply	0.77 – 1.11 V
Frequency	62.5 – 125 MHz
Frame rate	30 – 60 fps
Resolution	1920x1080
Power	58.6 – 216.5 mW
Energy/pixel	0.94 – 1.74 nJ

Two detectors, 97% pruning.

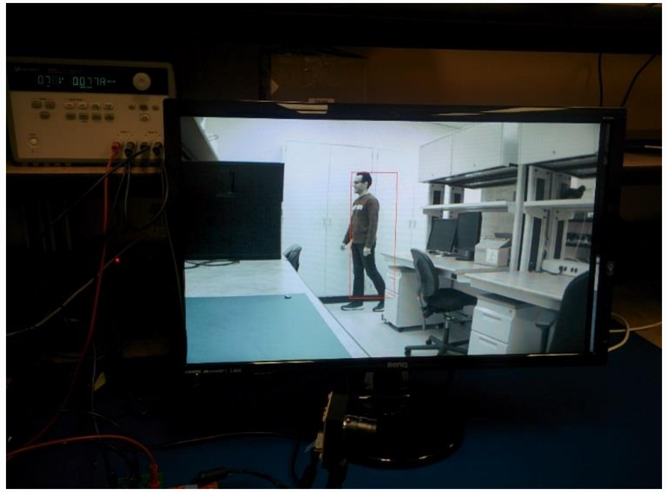
Energy Scalability

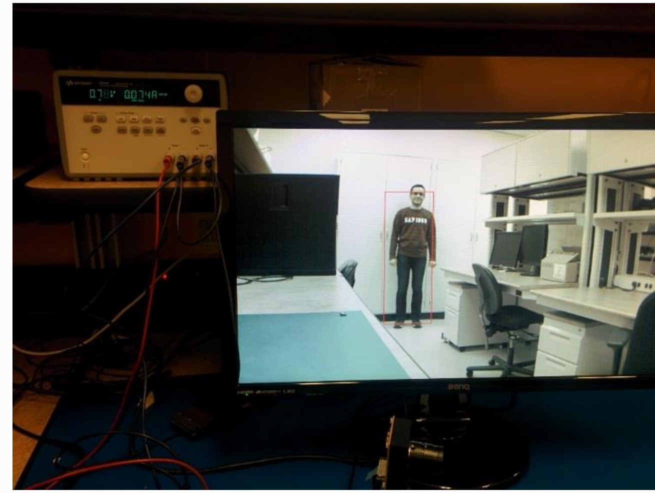


- 1-detector power : **15%** classification & **25%** feature storage
- Adding an extra detector increases power by only **19%**

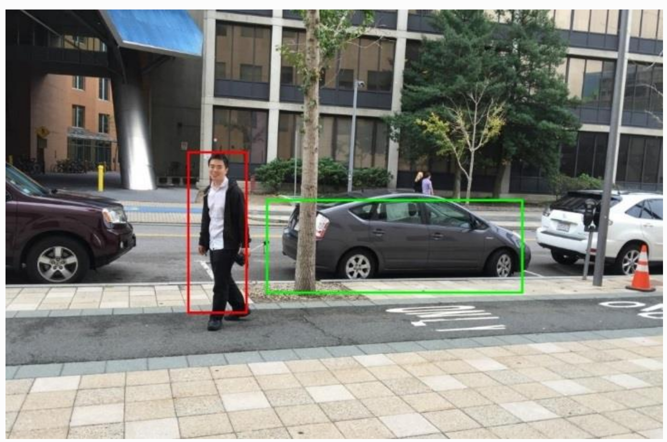
Detection Examples with DPM Chip

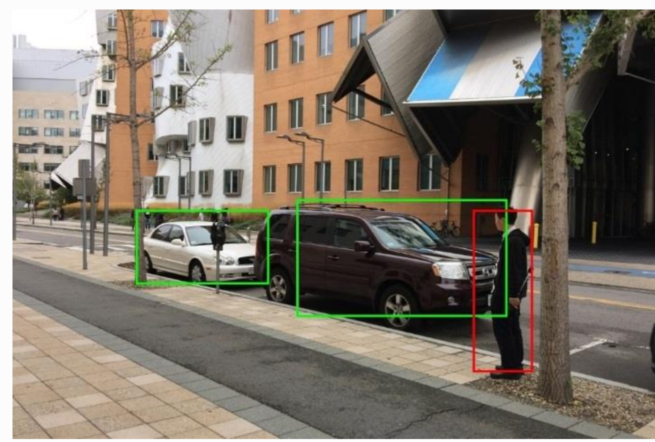
- Live video feed
- 1920x1080
- 30fps
- Detecting pedestrians





- Fixed frames
- 1920x1080
- Detecting cars & pedestrians





Comparison with ASIC Object Detectors

	JSPS 2014	This work
Process	65 nm	65 nm
Chip Size (mm²)	4.2×2.1	4.0x4.0
Voltage	0.7V	0.77V
Resolution	1920x1080	1920x1080
#Object Classes	2	2
Frame rate	30	30
Multi-scale	No	12 levels
Deformable Parts	No	8 parts
Accuracy (AP)	0.166	0.80
Power (mW)	84	58.6
Energy (nJ/pixel)	1.35	0.94

4.7x more accurate*

***INRIA person dataset**

30% less energy

Summary

- A 58.6mW object detection accelerator that processes 1920x1080 videos at 30 fps
  - Uses **deformable parts** for 2x increase in accuracy
  - **Two programmable object detectors** supporting **12 scales**
- Pruning, vector quantization and feature basis projection reduce the DPM classification cost
  - Reduce power by **5x** and memory size by **3.6x**
- This accelerator enables object detection to be as energy-efficient as video compression at < 1nJ/pixel

Acknowledgement

DARPA, Texas Instruments and TSMC University Shuttle