Human Detection in Image/Video

C. Carincotte
Multitel asbl, Belgium

Ongoing PhD A. Descamps

Human Activity and Vision Summer School
INRIA, Sophia-Antipolis, France
October 1st, 2012
Outlines

- Multitel asbl
  - Missions
  - Scientific and technical activities
  - Computer vision S&T activities

- Human detection
  - Introduction
  - Human detection for images
    - Overview and history of state of the art approaches
    - Standard datasets and evaluation procedure
    - Comparison of methods
  - Human detection for video surveillance
    - Look for video surveillance context
    - Integration of background substraction
    - Comparison with some publicly available approaches
  - Conclusion & Perspectives
Missions

I. R&D in the field of ICT technologies to enterprises
   • Industrial contracts: technology watch/transfer, feasibility study, prototyping, etc.
   • Partnerships in co-funded projects: regional, national, and European projects (ITEA, FP, etc.).

II. Creation and support to ”spin-off” companies

III. Presence on international scene
    • Participation in European initiatives: FP6, FP7, EUREKA
    • Innovation Partner of major industrial groups
    • Participation in conferences and fairs in Europe

IV. Training, seminars
    • Computer networks-Programming - Operating Systems
    • Photonics
    • Signal processing
    • RedHat training (first certification center in Belgium)
# Scientific and technical activities

## Applied photonics
Design and prototyping:
- of fiber lasers
- of passive components and optical sensors

## Signal & speech processing
Development of multimodal human-computer interfaces
Tracking objects and people in real time

## Railway certification
European Reference Laboratory ERTMS
Skills in R & D in the design of new tools (hard / soft)
Service validation and verification in the field of railway signaling

## Networking
Wireless networks Wi-Fi (WLAN)
IP telephony and VoIP (voice over IP)
**IMAGE PROCESSING DEPARTMENT:**

- Intelligent video surveillance applications
- Multimédia content analysis
- Machine vision

**Scientific and technical activities for**

- Partnership in national or European **collaborative projects**
- Development software solutions & prototypes (**industrial projects**)
Human detection / People counting

![Image of a busy corridor with people]

![Graph showing people flow over time]

![Another graph showing count over time]

Computer vision S&T activities
Computer vision S&T activities

Activity clustering / anomaly detection

- Counter flow
- Falling people (people gathering)
- Heckling
- Lost person
- Person distributing leaflets
- Cleaning staff emptying a garbage
- Persons phone calling
- etc.
Activity recognition / object classification
Computer vision S&T activities

Object classification / Defect detection
Content-Based Image Retrieval
Automated video inspection / Optical Character Recognition
Automated video inspection / Pattern recognition
Computer vision S&T activities

Smartphone applications

Open positions in
- Automated video inspection (stereovision), Pattern recognition and OCR
- Panoramic image reconstruction (video surveillance & medical projects)
Introduction

Human detection in images and video
→ very active topic in computer vision

• Many applications: surveillance, robotic, automotive safety, etc.

• Very challenging task:
  – Human is very variable in appearance (clothes, pose, …)
  – Real world problems: low resolution cameras, occlusions management, background issues, mono-view context, moving cameras, …
Introduction

How to detect humans in images?
Introduction

How to detect humans in images?

*Sliding window approach*

- Scan the image for every possible position and scale of the object
How to detect humans in images?

*Sliding window approach*

- Scan the image for every possible position and scale of the object
- For each subwindow, classify as human or non human
How to detect humans in images?

*Sliding window approach*

- Scan the image for every possible position and scale of the object
- For each subwindow, classify as human or non-human
- Pyramid of images for multiscale detection
Sliding window reduces the detection problem to a binary classification problem

Drawbacks:

- Each object instance usually generate multiple detections
- Partial occlusions, cropped persons
- Assumptions of scale invariance and fixed ratio
- Granularity of search space (finite grid)
- Huge number (typically $10^4$ or even $10^5$) of tested subwindows imposes strong constraints on classifier:
  - Very low false positive rate
  - Fast computation
How to classify human vs non-human?

- Feature extraction: extract discriminative features from raw image (human expertise)
- Classifier: classify between human and non-human (learned from training data)

Training of classifier requires a big training dataset (thousands of samples)
Introduction

How to classify human vs non-human?

- Collect positive (from annotation) and negative (random) datasets
- Extract features and train supervised binary classifier
Introduction

Why does it fail?
• Non human class very various : need lots of data
• Sliding window needs very low false positive rate
• Increase number of random negative samples : impracticable

Solution : bootstrapping
1. Collect initial dataset
2. Train on current dataset
3. Apply detector on training negative images
4. Add false detections to negative dataset
5. Return to (2)

Focus training on hard negative
Introduction
Many approaches available in state of the art!

Haar features / Adaboost (Viola, 2001 - Lienhart, 2002)
HOG / SVM (Dalal, 2005)
  Extensions
  – HOG / Adaboost (Zhu, 2006)
  – LBP (Mu, 2008), semantic LBP, etc.
LBP-HOG / Adaboost (Wang, 2009)
Shapelets (Sabzmeydani, 2007)
Covariance matrix (Tuzel, 2007)
Convolutionnal neural network (Szarvas, 2005)
Partial least squares analysis (Schwartz, 2009)
Integral Channel Features (Dollar, 2009)
Fastest pedestrian detector in the west (Dollar, 2010)
Discriminatively trained part based model (Felzenszwalb, 2010)
Multi-resolution model (Park, 2010)
Integration of motion in Viola-Jones (Jones, 2008), HOG (Dalal, 2006)
Integration of background information (Yao, 2008, Descamps, 2011)

…

and many more in whole state of the art…
Introduction

Overview

Human detection for images
- Overview and history of state of the art approaches
- Standard datasets and evaluation procedure
- Comparison of methods

Human detection for videosurveillance
- Look for video surveillance context: (low resolution, static camera etc.)
- Integration of background substraction
- Comparison with some publicly available approaches

Conclusion & Perspectives
Viola-Jones [Viola2001, Lienhart2002]

- **Haar Feature and extended ones**
  - Integral image for computation
  - Designed to respond to different local shapes (vertical, horizontal edge, etc)

- **Cascade of classifiers**

- **Main idea**: Large pool of simple features, let adaboost select/combine them
- Detection very fast, but training slow (weeks)
Human detection in images

Viola-Jones

Performs very well for faces, …
Human detection in images

Viola-Jones

Performs very well for faces, … but poorly with humans
Extensions of Viola-Jones for motion [Viola2005, Jones2008]

- Use difference and shifted difference images to capture motion
- Apply haar filters to both appearance and difference images
- The detector can model human motion over two or more frames and suppress static false detection
- Limited to static cameras

\[
\Delta = \text{abs}(I_t - I_{t+1}) \\
U = \text{abs}(I_t - I_{t+1} \uparrow) \\
L = \text{abs}(I_t - I_{t+1} \leftarrow) \\
R = \text{abs}(I_t - I_{t+1} \rightarrow) \\
D = \text{abs}(I_t - I_{t+1} \downarrow)
\]
Human detection in videos

Extension of Viola-Jones for motion

- Highly reduce false detection rate compared to Viola-Jones
- Allow to detect humans in low resolution (15x20px) and real-time
- Motion model is rough, and limited to walking humans
**HOG** [Dalal2005]

- **Histogram of Oriented Gradient Feature**
- Divide image in cells (e.g. 8x8 pixels squares)
- For each cell, compute weighted histogram of gradient over 8 orientation bins (angles in range 0-180 degrees)
- Normalize histogram over larger blocks
- Classification: linear SVM (non linear much slower / not much better)
Human detection in images

HOG
Largely outperforms previous human detectors
Human detection in images

HOG

Why does it work so well?

Carefully designed feature

- Describe complex shape, edges of object efficiently
- Robust to small deformations
- Good illumination/contrast invariance
- Inspired by popular SIFT

- Original HOG detector is slow : several seconds per image
- Integration in an adaboost cascade
  → Real-time detector (on low res. images) with similar performance

7 years later, HOG features are still used in state of the art approaches for object detection
Human detection in videos

Extension of HOG for motion

*Internal Motion Histogram*

- Compute dense optical flow
- Use local differences of flow ($I^x, I^y$) for orientation vote
- Capture relative movement between different parts of the images
- Complementary information with static HOG
Human detection in videos

Extension of HOG for motion

- IMH improves performance of static HOG for moving people, and don’t decrease for static ones
- Can be used with moving cameras
- Need good optical flow
LBP features [Zhu2006,Mu2008]

- Similarly to HOG, respond to edges in image, but sensitive to curvature of the edge
- Automatically discard noisy (non uniform) regions

Slightly better performance than HOG

Variants: Semantic LBP, HOG-LBP
Covariance matrix [Tuzel2007]

- Extract low level feature maps (gradient/gradient orientation info)

\[ \begin{bmatrix} x & y & |I_x| & |I_y| & \sqrt{I_x^2 + I_y^2} & |I_{xx}| & |I_{yy}| & \arctan \left( \frac{|I_x|}{|I_y|} \right) \end{bmatrix}^T \]

- Compute covariance matrix of these feature over subwindows
- Feature: d*d covariance matrix for each subwindow
Human Detection in images

Covariance matrix

- Flexible: easy to add more low level features, but $O(d^2)$
- Low level features mostly related to gradient/gradient orientation
- Encode complex information about low-level features: variance, spatial distribution, correlations
- Robust to illumination change
- Can be computed by integral image, using $d(d+1)/2$ images

- Not usable directly in standard classifier:
  Modified version of logitboost classify in Riemannian manifold
Covariance matrix
Human detection using partial least squares analysis [Schwartz2009]

Usage of complementary features improve performance:

- HOG
- Color frequency (highest gradient color channel)
- Co-occurrence matrix features (texture descriptor)

→ Results in a very high dimension of feature vector: 170820

SVM training intractable on a so high dimensionnal space
→ Project on lower dimensional space
→ PLS + quadratic classifier
Human Detection in images

Human detection using partial least squares analysis
Integral Channel Features [Dollar2009]

Generalize the features computed by integral image:
1. Compute a set of channels images and their integral images
2. Features are (combination of) integral sum of channel pixels in rectangular subwindows

- Used channels:
  - LUV
  - Gradient magnitude
  - Gradient histogram (HOG)
Integral Channel Features

- Generic definition: any image transformation can be used as channel
- Very simple and fast to compute
- Most existing features can be integrated (Haar, HOG, LBP, …)
- Allow to test easily features in the same classification framework
- State of the art performance

- Very large set of possible features, but random sampling combined with adaboost classifier is efficient
Fastest pedestrian detector in the west  [Dollar2010]

- Many channels are not scale invariant (e.g. gradient), they must be computed at each scale during sliding window detection
- The effect of rescaling can be approximated at nearby scales

→ Faster detector with similar performance
Discriminatively trained part based model [Felzenszwalb2010]

Score is sum of appearance scores plus deformation score
Human Detection in Images

Discriminatively trained part based model
Multi resolution model [Park2010]
- Usual method: train a model at one scale and rescale the image or the model for other scale
- Assumption: appearance is scale invariant -> false
Human Detection in Images

Multi resolution model

Low-resolution model

High-resolution model

Multiresolution model
**INRIA dataset** [Dalal2005]

- Static images from web, personnal digital images
- Training data:
  - 1218 background images
  - 2416 positive samples from 614 images
- Test data:
  - 288 test images containing 589 annotated persons

- Upright persons, wide variety of situations
- **Fairly high resolution (>100px tall) and good quality of images**
Human Detection Tutorial

INRIA dataset

- 82% Shapelet
- 80% PoseInv
- 72% VJ
- 58% FrtMine
- 46% HOG
- 44% LatSvm-V1
- 43% HikSvm
- 40% PIs
- 39% HogLbp
- 36% MultiFtr
- 31% FeatSynth
- 25% MultiFtr+CSS
- 22% ChnFtrs
- 21% FPDW
- 20% LatSvm-V2

false positives per image
INRIA dataset
Integral channel feature results
INRIA dataset
Part based model results
Human Detection: benchmark

INRIA dataset

- Big improvement since Viola Jones (from 70% miss rate to 20%)
- Almost all methods in leading group use HOG
- Two approaches for the best methods:
  - Combination of different features (HOG, texture, color)
  - Part-based models
- Training data is important (diff. between LatSvm-V1 (Pascal training) and LatSvm-V2 (INRIA training))
- There is still room for improvement
**Caltech dataset** [Dollar2011]

- Videos collected from a vehicle driving in urban environment
- 250000 frames, 350000 bounding box with occlusion annotation
- Mostly walking and standing persons
- Wide range of scales and occlusions

**Advantages**
- Large and challenging dataset
- No selection bias
- Allow usage of temporal features
- Allow experiments over persons scales, occlusion level, …
Human Detection Tutorial

Caltech dataset

![Graph showing miss rate vs. false positives per image for various detection methods.

Key:
- 95% VJ
- 91% Shapelet
- 86% PoseIn
- 80% LatSvm-V1
- 74% FtrMine
- 73% HikSvm
- 68% HOG
- 68% MultiFtr
- 68% HogLbp
- 63% LatSvm-V2
- 62% Pls
- 61% MultiFtr+CSS
- 60% FeatSynth
- 57% FPDAW
- 56% ChnFtrs
- 54% Crosstalk
- 51% MultiFtr+Motion
- 48% MultiResC]
Caltech dataset

Near scale (>80px)

Medium scale

Far scale (<30px)
Human Detection: benchmark

Caltech dataset

No occlusion

Partial occlusion (<35%)

Heavy occlusion (<80%)
Caltech dataset

- Globally very challenging dataset
- Best detectors use combination of features (HOG, color, texture, motion)
- Part based method is efficient only in near scale
- Motion features improve performance, but only in near scale
- No method get good result in far/medium scales nor with occlusion
- Scale is important: multiscale detector has best results
- But influence of training data?

"latsvm-v1" is trained with PASCAL, multiFtr with TUD-Brussel, multiresc with Caltech, others with INRIA
Human detection for videosurveillance

It seems a « much easier » task:

- Humans have specific motion patterns, background is mostly static
- Trajectory consistency over many frames may help

But:

- Resolution is usually low
- Finding good features and models for human motion is not easy
- What to do with static persons?

→ Frame by frame detection is still an usual approach

For videosurveillance static cameras, integration of background subtraction can help improve performance
Human detection for videosurveillance

Background subtraction for covariance features [Yao2008]

- Background subtraction: segment foreground objects
- Only for static cameras
- Combined with appearance, may help to detect moving persons and remove false alarms
Human detection for videosurveillance

Fast human detection from videos using covariance features [Yao2008]

Modified covariance low-level features:

\[
\begin{bmatrix}
  x & |I_x| & |I_y| & \sqrt{I_x^2 + I_y^2} & \arctan \frac{|I_y|}{|I_x|} & G & \sqrt{G_x^2 + G_y^2}
\end{bmatrix}
\]

Use foreground probability (G) and edges (\(\sqrt{G_x^2 + G_y^2}\))
Human detection for videosurveillance

Fast human detection from videos using covariance features
Spatio-temporal integral channel features [Descamps2011]

- Features channel relying on foreground mask analysis
Spatio-temporal integral channel features

Spatio-temporal integral feature

First selected features
Spatio-temporal integral channel features

Spatio-temporal integral feature
Videosurveillance datasets
Public datasets: CAVIAR, PETS, ILIDS
Human Detection : benchmark

Videosurveillance datasets
VANAHEIM dataset
Human Detection: benchmark

Videosurveillance datasets

- Training:
  - 9200 positive samples from CAVIAR, PETS2009, AVSS and VANAHEIM
  - Negative sample from various videosurveillance context
- Evaluation:
  - VANAHEIM and CAVIAR data (19 cameras, 3900 person annotation)
  - Indoor context
  - Various point of view, scales and occupancy level
  - Evaluate on unoccluded persons only
Videosurveillance datasets

![Graph showing miss rate vs false detection per frame for different datasets with percentages.]

- ChnFtr + Background (48.5%)
- Cov + Background (54.4%)
- ChnFtr (58.9%)
- HOG (63.9%)
- Vj (66.6%)
Videosurveillance datasets

Near scale (>90px)  Medium scale  Far scale (<45px)
Human Detection: benchmark

Videosurveillance datasets
Integral channel feature with background result
Human Detection : benchmark
Videosurveillance datasets

- Background substraction improve performance, mostly in medium scale
- Good results for single persons in high resolution

Problems with:
- Groups of persons (occlusion, background inefficient)
- Low resolution persons
- Background movements, illumination variations (mostly outdoor)

What should we look for?
- Robust motion features, especially in low resolution
- Good occlusion reasoning models in high resolution
Great improvements in human detection during the last 10 years:

- Classification frameworks: SVM, adaboost cascade, part based models
- Features: HOG, LBP, motion features, …

Detection of single persons in high resolution works fairly well, but still far from human performance

**Main challenges**
Low resolution persons
Occlusions, groups of persons
Research directions

• Explicitly modelize occlusion
• Multiresolution models
• Use context
• Find better motion features, especially for low resolution
• Temporal integration
• More data
• Go beyond sliding window
Bibliography