People Tracking in RGB-D Data With On-line Boosted Target Models

Arash Akbarinia, Corina Barbalata, Raquel Gil, Eduardo Tusa

November 13, 2012

Motivation

People detection and tracking are fundamental components for many:

- Robots,

Motivation

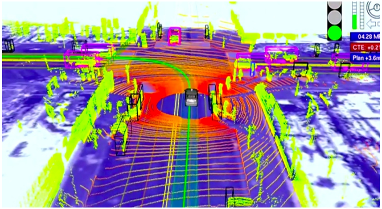
People detection and tracking are fundamental components for many:

- Robots,
- Interactive systems,

Motivation

People detection and tracking are fundamental components for many:

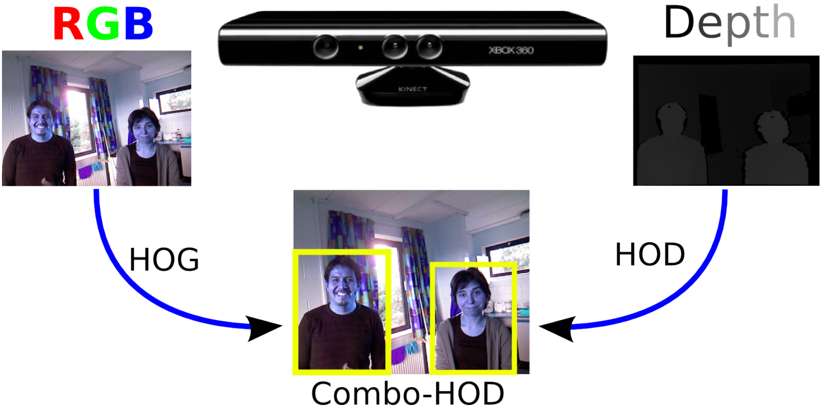
- Robots,
- Interactive systems,
- Intelligent vehicles.



Why RGB-D sensors?

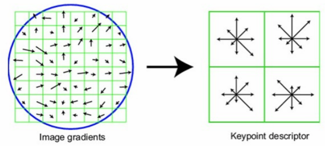
1. RGB-D sensors are cheap nowadays, e.g. *Kinect*.
2. Detection of objects in unfavourable light-conditions, e.g. a bar.
3. Detection of objects made of very low-return materials, e.g. our glasses.
4. Detection confidence can be combine from colour and depth images.

Description



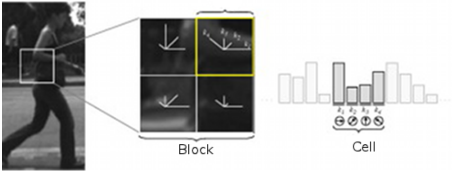
HOG: Histograms of Oriented Gradients

- The method considers a fixed-size detection window which is densely subdivided into an uniform grid of cell.
- For each cell, the gradient orientations over the pixels are computed and collected in a 1D histogram.
- Group of adjacent cells, blocks, are used to normalise.
- The descriptor is built concatenating all block histograms.



HOG: Histograms of Oriented Gradients

For detecting people the window is scrolled over the image at different scales.

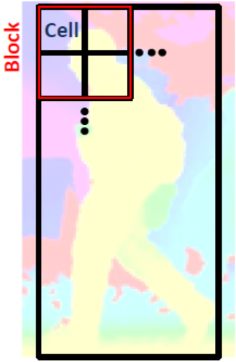


For each position and scale, *HOG* descriptors are computed and classified with the learnt *SVM*.

HOD: Histograms of Oriented Depths

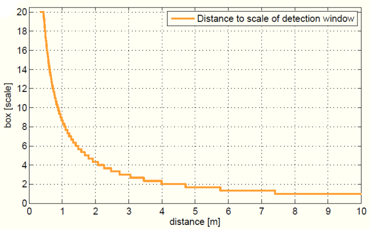
Inspired by *HOG*

- A window is fixed and divided into squared cells. Cells are used for computing 1D histograms of oriented depth changes.
- Blocks are used for normalization of cells histograms.
- Concatenated histograms from all blocks.
- Linear-SVM as classifier.



Depth Informed Scale Space Search

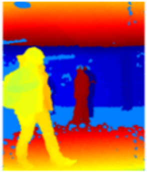
Scale to distance regression



$$s = \frac{F_y H_m}{d} \frac{1}{H_w}$$

- F_y is the vertical focal length of IR camera,
- H_m average height of a person,
- H_w height in m of the window at scale 1.

1. Compute s for every pixel in the image, *scale list*, $\mathcal{S}$.
2. For each $s \in \mathcal{S}$ we get the depth relative to this scale.



Depth Image

Depth relative to each scale (layers)

3. Run the detector in the non-empty areas by using integral image in that layer.

Examples

Scale 3 at 2.8m.

Scale 1.3 at 5.6m.

Combo-HOD

Depth Data

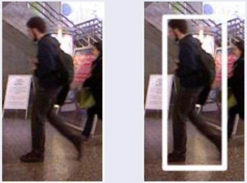
- Robust to illumination conditions, returns 3D data.
- Sensitive to surface properties, suffers from a limited resolution.

Image Data

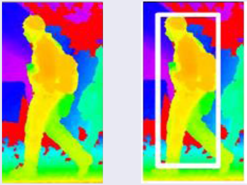
- Rich in color and texture, high angular resolution.
- Suffer from illumination changes.

Density probability

HOG $\longrightarrow$ p_G

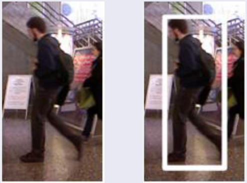


HOD $\longrightarrow$ p_D

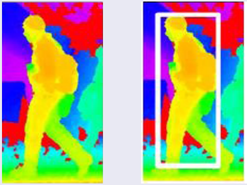


Density probability

HOG $\longrightarrow$ p_G



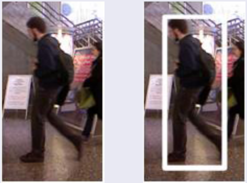
HOD $\longrightarrow$ p_D



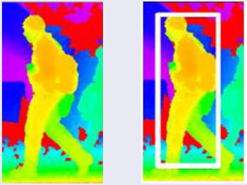
$$\boxed{p = p_D + k(p_G - p_D)}$$

Density probability

HOG $\longrightarrow$ p_G



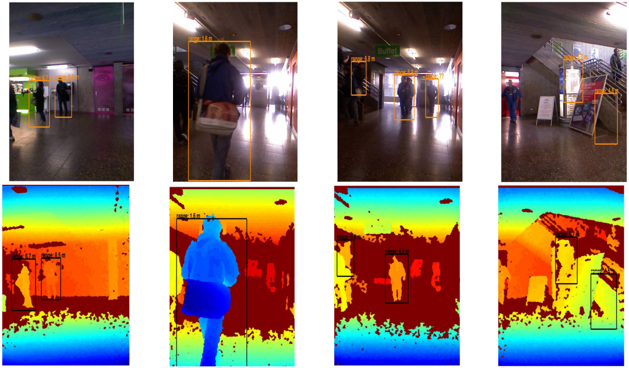
HOD $\longrightarrow$ p_D



$$p = p_D + k(p_G - p_D)$$

- $k = \frac{\sigma_D^2}{\sigma_D^2 + \sigma_G^2}$
- $\sigma_D^2 = \frac{\#\text{false positives } HOD}{\#\text{false positives } HOG}$
- $\sigma_G^2 = 1 - \sigma_D^2$

Experiments

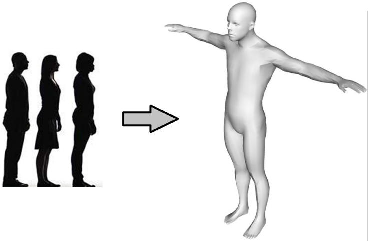


People are detected at several ranges at varying partial occlusions and in different visual and depth clutter.

Experiments

On-Line Boosting

The detector learns a generic person model from a priori labeled data.



On-line boosting is used to learn target appearance models in RGB-D data.

Boosting

AdaBoost [Freund and Shapire, 95]

Given training samples x with labels y:
$(x_1, y_1), ..., (x_m, y_m); x_i \in \chi, y_i \in \{-1, +1\}$

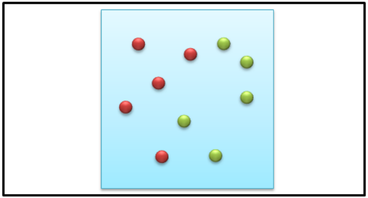


Figure: Training samples

Boosting

AdaBoost [Freund and Shapire, 95]

This method trains weak classifiers $h(x)$ from labeled training samples (x, y).

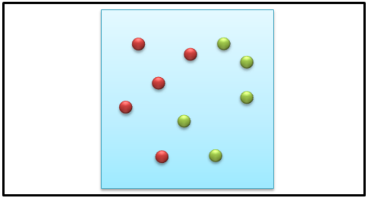


Figure: Training samples

Boosting

AdaBoost [Freund and Shapire, 95]

This method is initialized with uniform weights w_i associated to each x.

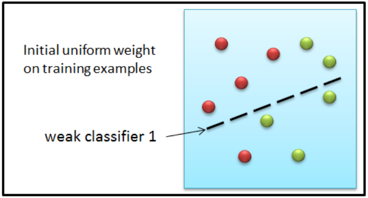


Figure: Weak classifier $h_1(x)$

Boosting

AdaBoost [Freund and Shapire, 95]

Learning is done in rounds where the weights are updated based on the mistakes of the previous weak learner.

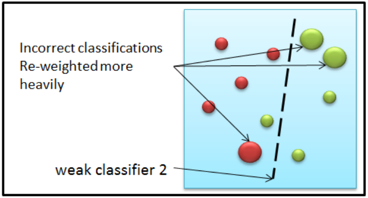


Figure: Weak classifier $h_2(x)$

Boosting

AdaBoost [Freund and Shapire, 95]

Learning is done in rounds where the weights are updated based on the mistakes of the previous weak learner.

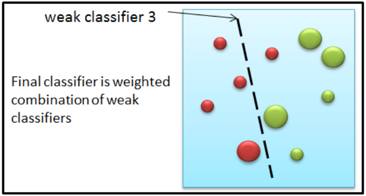


Figure: Weak classifier $h_3(x)$

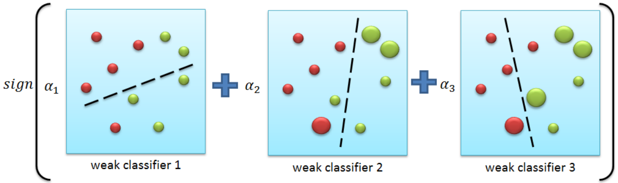
Boosting

AdaBoost [Freund and Shapire, 95]

A strong classifier $H(x)$ is computed as linear combination of a set of weighted hypotheses called weak classifiers $h_i(x)$.

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x))$$

Derivation of a Strong classifier $H(x)$



Boosting

On-line Boosting [Oza and Russell ,01]

- One training sample at a time.
- Strong classifier is initialized at the beginning and is updated by each training sample.
- Selectors are introduced in this approach.

On-line Boosting

Updating Weak Classifiers

The weight of a sample (called importance λ in this context) can be estimated by propagating it through a fixed chain of weak classifiers.

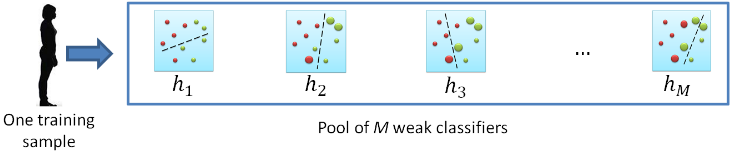


Figure: Updating Weak Classifiers

On-line Boosting

Updating Weak Classifiers

The error of the i-th weak classifier is estimated from the summed weights of the correctly (λ_i^{corr}) and wrongly (λ_i^{wrong}) classified samples

$$e_i = \frac{\lambda_i^{wrong}}{\lambda_i^{wrong} + \lambda_i^{corr}}$$

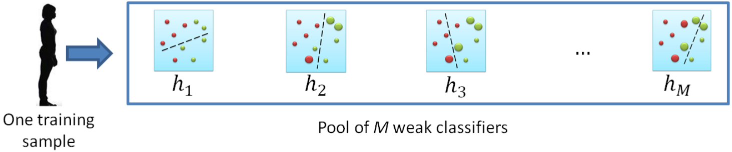
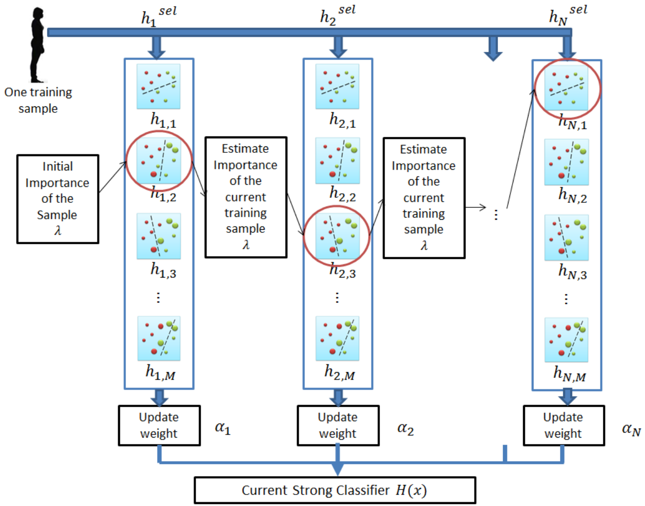


Figure: Updating Weak Classifiers


h_1^{sel}
h_2^{sel}
h_N^{sel}
One training sample
Initial Importance of the Sample λ
$h_{1,1}$
$h_{1,2}$
$h_{1,3}$
$h_{1,M}$
Estimate Importance of the current training sample λ
$h_{2,1}$
$h_{2,2}$
$h_{2,3}$
$h_{2,M}$
Estimate Importance of the current training sample λ
$h_{N,1}$
$h_{N,2}$
$h_{N,3}$
$h_{N,M}$
Update weight
α_1
Update weight
α_2
Update weight
α_N
Current Strong Classifier $H(x)$

On-line Boosting

Feature selection

A selector h^{sel} selects the best weak classifier h_i from a pool of M weak learners with "best" being defined by the lowest error.

Strong classifier

The strong classifier is finally obtained by computing the confidence as a linear combination of the N selectors

$$H(x) = sign(\sum_{n=1}^{N} \alpha_n h_n^{sel}(x))$$

Features [Viola and Jones, 01]

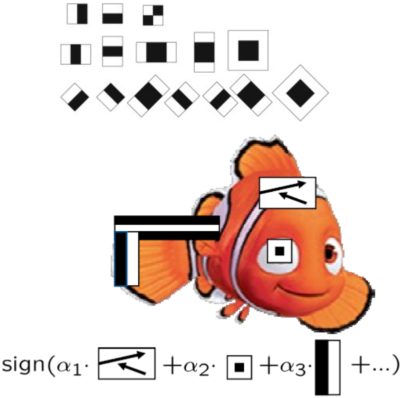


Figure: Haar-like features Illustration

Features [Viola and Jones, 01]

RGB-D Data

- Haar-like features in the intensity image (converted from the RGB values)
- Haar-like features in the depth image.
- Illumination agnostic Lab color features in the RGB image.

.

Features [Viola and Jones, 01]

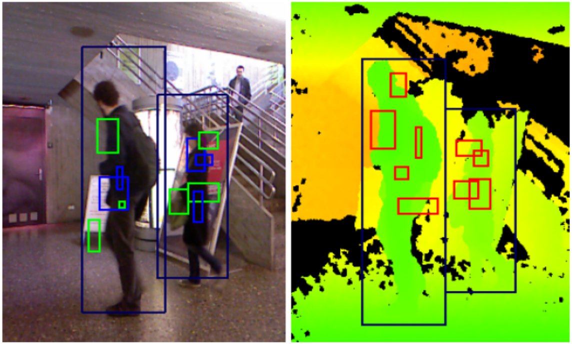


Figure: Bounding boxes of two detected persons in the RGB and depth images. The ten best features of each on-line detector are marked with colored rectangles.

On-line Boosting for Tracking [Grabner and Bischof, 06]

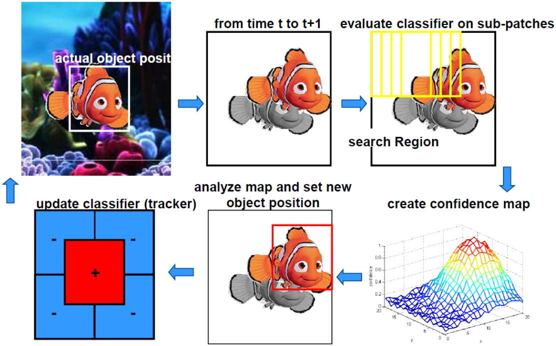
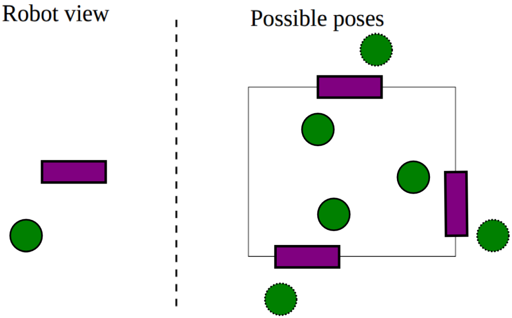


Figure: The main steps of tracking by a classifier.

Description

- A detected door gives several possible poses for the robot.
- Closely related to the problem of multi-target tracking.

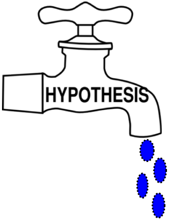


Multiple Hypotheses Tracking (MHT)

Multiple Hypotheses Tracking

Is a type of tracking algorithm that keep at each time step multiple hypotheses about the past and current associations.

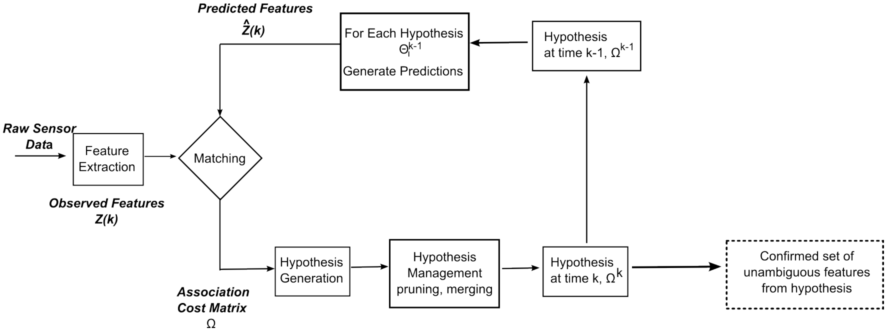
- Belief is represented by multiple hypotheses.
- Each hypotheses is tracked by a Kalman filter.



Multiple Hypotheses Tracking (MHT)

Multiple Hypotheses Tracking

Exponentially growing hypothesis tree pruned by Murty's algorithm.



Measurement likelihood for standard MHT

Measurement likelihood

Quantifies how well an observation matches a predicted measurement based on position and velocity.

$$p(z_i(t) \mid \psi_j^t, \Omega_{p(l)}^{t-1}) = \mathcal{N}(z_i(t); \hat{z}_j(t), S_{ij}(t))$$

Consists of:

- observations interpreted as new tracks and false alarms
- matched observations with innovation covariance matrix

Joint Likelihood Data Association

Joint likelihood

Accounts for both motion state and appearance:

$$p(z_i(t) \mid \psi_j^t, \Omega_{p(l)}^{t-1}, H^{t-1}) = p(z_i(t) \mid \psi_j^t, \Omega_{p(l)}^{t-1}) \cdot p(x_i(t) \mid H^{t-1})$$

Appearance Likelihood:

- added by the on-line classifier H
- expresses how much the observed target's appearance matches the learnt model.

Feeding Data Association Back to On-line Boosting

Tracker produces information for on-line boosting

Feeding Data Association Back to On-line Boosting

Tracker produces information for on-line boosting

- New target - new track is initialized, new classifier created.

Feeding Data Association Back to On-line Boosting

Tracker produces information for on-line boosting

- New target - new track is initialized, new classifier created.
- Existing target associated to observation - classifier is updated.

Feeding Data Association Back to On-line Boosting

Tracker produces information for on-line boosting

- New target - new track is initialized, new classifier created.
- Existing target associated to observation - classifier is updated.
- MHT declares a track as occluded:

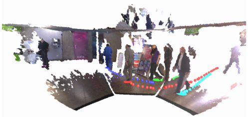
Feeding Data Association Back to On-line Boosting

Tracker produces information for on-line boosting

- New target - new track is initialized, new classifier created.
- Existing target associated to observation - classifier is updated.
- MHT declares a track as occluded:
  - occlusion
  - misdetection

Experiments – Dataset Collection

- Recording data set in the lobby of a large university.



- Priori detector trained with background in different building.
- 3021 people in 1133 frames and 31 tracks labelled.

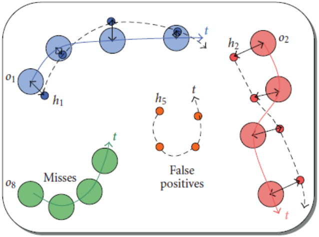
Experiment – Sensory Setup

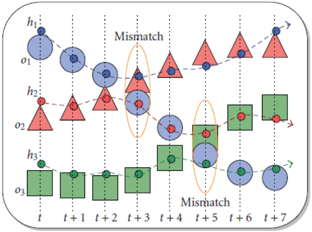
- 3 Kinect, field of view $130^\circ \times 50^\circ$. Mounted at $1.2m$.

- Cameras calibrated.
- All data available in their website.

Experiment – Parameters

- MHT parameters learnt from data of 600 frames.
- Maximal number of hypothesis set to 100.
- CLEAR MOT metrics for assessment [6].





Results – Videos

- Tracking people

Results – Confidence

- Confidence evaluation reaches steady state at 0.8.

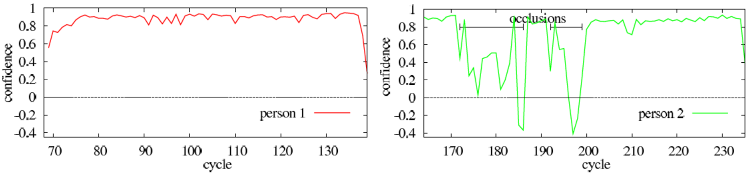


Figure: Evaluation of the confidence of the online detector.

- Occlusions drops the confidence.
- Confidence increases immediately after reappearing.

Results – CLEAR MOT

- Clear improvements.

	FN	FP	ID	MOTA
Baseline	1502	168	42	62%
Online boosting	751	201	32	78%
Improvement	50%	-19%	24%	16%

Table: CLEAR MOT Results

- Delayed deletion of tracks increases FP.
- ID improvement could have been even larger if dataset had higher occlusion.

Summary

- Detect with Combo-HOD.
- Treat tracking as classification, use online boosting.
- Verify tracking with multiple Kalman Filters.

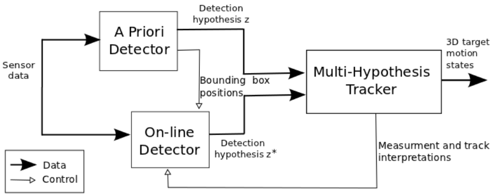


Figure: Decisional framework.

Conclusions

- Novel 3D people detection and tracking in RGB-D.
- Combination of online learning of target appearance and multi-hypothesis tracking.
- Framework to integrate offline detector, online boosting classification and MHT.

References I

- M. Luber, L. Spinello & Kai O. Arras, *People Tracking in RGB-D Data With On-line Boosted Target Models*, University of Freiburg, Germany.
- L. Spinello & Kai O. Arras, *People Detection in RGB-D Data*, University of Freiburg, Germany.
- Y. Freund & R. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting*, in Computational Learning Theory, 1995.
- N. C. Oza & S. Russell,, *Online bagging and boosting*, in Artificial Intelligence and Statistics, 2001, pp. 105, 112.

References II

- H. Grabner & H. Bischof, *On-line boosting and vision*, in Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, New York, USA, 2006.
- Keni Bernardin and Rainer Stiefelhagen, *Evaluating multiple object tracking performance: the CLEAR MOT metrics*. J. Image Video Process. Article 1 (January 2008), 10 pages.
- John J. Leonard, Bradley A. Moran, Ingemar J. Cox & Matthew L. Miller, *Underwater Sonar Data Fusion Using an Efficient Multiple Hypothesis*. IEEE Robotics and Automation, 1995.

THANK YOU