Putting Objects in Perspective
Derek Hoiem, Alexei A. Efros, Martial Hebert

presented by
Phongsathorn Eakamongul

Department of Computer Science
Asian Institute of Technology

2009, July 2

Outline

1. Abstract

2. Introduction

3. Projective geometry

4. Modeling the scene

5. Training + Evaluation

6. Viewpoint by Example

7. Summary

Abstract

Statistic Framework for placing local object detection by modeling interdependence of

- camera viewpoint (position/orientation)
- objects identities
- surface orientations (3D scene geometry)

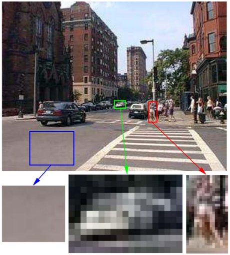
using single image.

Outline

1. Abstract
2. Introduction
3. Projective geometry
4. Modeling the scene
5. Training + Evaluation
6. Viewpoint by Example
7. Summary

General Object Detection cannot be solved locally

We cannot recognize car, person, and the road in isolation



Pixel of car can be easily be interpreted as person's shoulder, mouse, stack of books, balcony, etc.

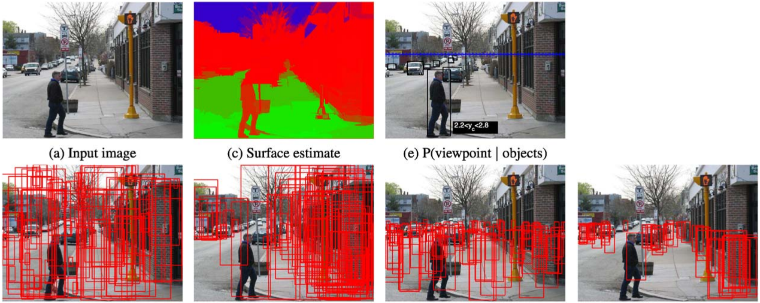
Psychophysical Evidence

context plays crucial role in scene understanding (i.e. Biederman 1981; Torralba 2005).
(1) car is sitting on the road
(2) right size relative to other object in the scene.

Related Work

- Heuristic Image understanding systems
  - ACRONYM (Brooks 1979), VISIONS (Hanson and Riseman 1978), outdoor scene understanding system (Ohta and Kanade 1985), intrinsic images (Barrow and Tenenbaum 1978), etc.
- Small image windows (at all locations and scales) to find specific objects
  - works wonderfully with face detection (Schneiderman 2004; Viola and Jones 2004) – since inside the face is much more important than boundary.
  - unreliable for cars, pedestrians, especially at smaller scale.
- Modeling within 2D image
  - modeling direct relationships between objects and other objects (Kumar and Herbert 2003; Murphy et al. 2003), regions (He et al. 2004; Kumar and Hebert 2005; Tu et al. 2005), scene geometries (Murphy et al. 2003; Sudderth et al. 2005)
- Modeling 3D image
  - estimate rough 3D scene geometry from single image (Hoiem et al. 2005), Viewpoint and mean scene depth (Torralba and Sinha 2001), geometric consistency of object hypotheses in simple scenes using hard algebric constriants (Forsyth et al. 1994)
- Relationship between camera parameters and objects
  - well calibrated camera (Jeong et al. 2001, etc.), stationary surveillance camera (Krahnstoever and Mendonca 2005), or both (Greienhagen et al. 2000)

Overview



(a) Input image (c) Surface estimate (e) P(viewpoint | objects)

(b) P(person) = uniform (d) P(person | geometry) (f) P(person | viewpoint) (g) P(person|viewpoint,geometry)

100 boxes sampled
(b) local object detector (Murphy et al. 2003; Dalal and Triggs 2005) at any location/scale
(c) estimate rough surface geometry (Hoiem et al. 2005)
red, green, blue channels indicate vertical, ground, sky
(e) estimate camera viewpoint from scale of objects in image

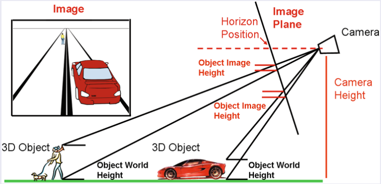
This is like typical human eye-tracking, where subjects are asked to search for a person in an image.

Outline

1. Abstract
2. Introduction
3. Projective geometry
4. Modeling the scene
5. Training + Evaluation
6. Viewpoint by Example
7. Summary

Projective geometry

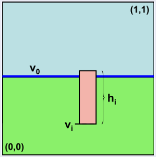
object’s height in image can be determined from it’s height in the world



y_c : camera height

–

y_i : object height



v_0 : horizon position

u_i, v_i : lower-left corordinate

h_i : object image height

–

v_c : center of image

v_t, v_b : top and bottom of object

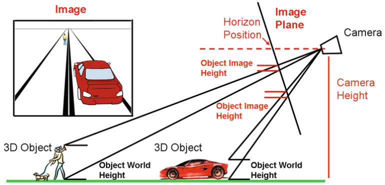
Assume

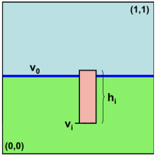
- all objects of interest rest on the ground plane. (cannot find people on the roof)
- objects perpendicular to the ground

We can estimate object world height

$$y_i \approx \frac{h_i y_c}{v_0 - vi}$$

Camera Tilt





camera tilt

$$\theta_x = 2arctan(\frac{v_c - v_0}{2f})$$

Proof

Assume

- camera roll = 0 or image is rectified
- camera intrinsic parameters are typical (skew = 0, unit aspect ratio, typical f)

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \frac{1}{z} \begin{bmatrix} f & 0 & u_c \\ 0 & f & v_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos\theta_x & -sin\theta_x & y_c \\ 0 & sin\theta_x & cos\theta_x & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}$$

From this,

$$y = \frac{z(fsin\theta_x - (v_c - v)cos\theta_x - fy_c)}{(v_c - v)sin\theta_x + fcos\theta_x} \tag{1}$$

Since y=0 at v_b,

$$z = \frac{fy_c}{fsin\theta_x - (v_c - v_b)cos\theta_x} \tag{2}$$

put (2) in (1)

$y = \frac{fy_c(fsin\theta_x) - (v_c - v_t)cos\theta_x / fsin\theta_x - (v_c - v_b)cos\theta_x - fy_c}{(v_c - v_t)sin\theta_x + fcos\theta_x}$

Consider condition that camera is on the ground,

- Since θ_x is small; $cos\theta_x \approx 1$, $sin\theta_x \approx \theta_x$, $\theta_x \approx \frac{v_c - v_0}{f}$

  $y \approx y_c \frac{v_t - v_b}{v_0 - v_b} / (1 + (v_c - v_0)(v_c - v_t)/f^2)$
- Since tilt tend to be small; $v_c - v_0 \approx 0$, and $f > 1$

  $y \approx y_c \frac{v_t - v_b}{v_0 - v_b}$

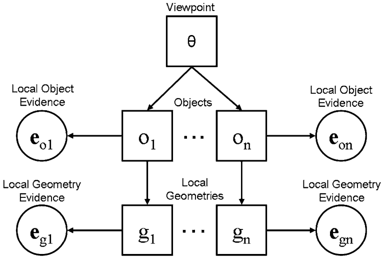
so $y_i \approx \frac{h_i y_c}{v_0 - vi}$

Outline

1. Abstract
2. Introduction
3. Projective geometry
4. Modeling the scene
5. Training + Evaluation
6. Viewpoint by Example
7. Summary

Probabilistic model to describe a scene

Assume θ, o, g are conditionally independent.



θ : viewpoint

o : object identity

g : surface geometry

e_o : object evidence

e_g : geometry evidence

From model

$$P(\theta, o, g, e_g, e_o) = P(\theta) \prod_i P(o_i|\theta) P(e_{oi}|o_i) P(g_i|o_i) P(e_{gi}|g_i)$$

The likelihood of the scene conditioned on image evidence using Bayes rule

$$P(\theta, o, g|e_g, e_o) = P(\theta) \prod_i P(o_i|\theta) \frac{P(o_i|e_{oi})}{P(o_i)} \frac{P(g_i|e_{gi})}{P(g_i)}$$

Classifier can be incorporated using probabilistic output $P(o_i|e_{oi})$

Inference

- Pearl's belief propagation algorithm (Pearl 1988) is optimal and efficient.
- inference by quantizing continuous variable v_0 and y_c into evenly-spaced bins (50 and 100 bins, respectively)
- Their implementation make use of Bayes Net Toolbox (Murphy 2001)
- After inference, they can query, such as
  - "what is expected height of this object ?"
  - "what are marginal probability of cars ?"
  - "What is the most probable explanation of the scene ?"
- They report results based on marginal probability from sum-product algorithm (this allows ROC curve to be computed)

Camera Viewpoint (θ)

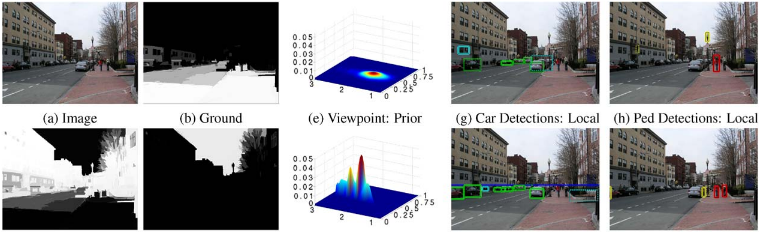
Consider v_0, y_c to be independent a priori.

$$P(\theta) = P(v_0)P(y_c)$$

- v_0 distribution is estimated using Simple Gaussian prior
- y_c distribution is estimated using kernel density estimation .

A priori,
Most likely $y_c = 1.67$, at eye level of typical adult male
Most likely $v_0 = 0.50$

Camera Viewpoint



(a) Image (b) Ground (e) Viewpoint: Prior (g) Car Detections: Local (h) Ped Detections: Local

(c) Vertical (d) Sky (f) Viewpoint: Full (i) Car Detections: Full (j) Ped Detections: Full

(e) Viewpoint prior θ (left axis - y_c, right axis - v_0) – high variance, more informative than uniform distribution that implicitly assumed when scale is considered irrelevant.

(g, h) local object detection

(b, c, d) Geometry estimates g_i

–

Using this model, they improve estimates of

(f) Viewpoint : full

Viewpoint peak likelihood increases from 0.0037 a priori to 0.0503 after inference.

(i, j) local detection improve when viewpoint and surface geometry are considered.

At the same false positive (cars: cyan, peds: yellow) rate, the true detection (cars: green, peds: red) rate doubles.

Note : detections with lower confidence exist for both local and full model but not shown here.

Object Identities (o_i)

o_i consists of

- t_i : $type \in \{object, background\}$ (i.e. pedestrian) : identity
- $bbox_i$: $boundingbox = \{u_i, v_i, w_i, h_i\}$
  - w_i : width
  - h_i : height

Prob of object occurring at a particular location/scale

$$P(t_i = obj, bbox_i | I_i) = \frac{P(I_i | t_i = obj, bbox_i) P(o_i)}{P(I_i | t_i = obj, bbox_i) P(o_i) + P(I_i | t_i \neq obj, bbox_i)(1 - P(o_i))}$$

$$= \frac{1}{1 + exp[-c_i - log \frac{P(o_i)}{1 - P(o_i)}]}$$

$P(o_i)$: prior

I_i : image information at ith bounding box

class-conditional log-likelihood ratio : $c_i = log \frac{P(I_i | t_i = obj, bbox_i)}{P(I_i | t_i \neq obj, bbox_i)}$

Object Identities (o_i)

- Typically, researcher perform non-maxima suppression , assuming that high detection responses at neighboring positions could be due to an object at either of those positions (but not both).
- Applied non-maxima suppression + form a point distribution out of the non-maxima (rather than discarding them)
- An object candidate is formed out of a group of closely overlapping bounding boxes.

Object Identities (o_i)

The candidate likelihood

$P(t_i = obj|e_o)$ = the likelihood of the highest confidence bounding box.
$P(bbox_i|t_i = obj, e_o) \propto P(t_i = obj, bbox|I)$

Object's height depends on its position when given viewpoint.

$P(o_i|\theta) \propto p(h_i|t_i, v_i, \theta)$
(due to the uniformity of $P(t_i, v_i, w_i|\theta)$)

From $y_i \approx \frac{h_i y_c}{v_0 - vi}$
if y_i is normal distribution, with parameters $\{\mu_i, \sigma_i\}$
h_i is also normal distribution with parameters $\{\frac{\mu_i(v_0 - v_i)}{y_c}, \frac{\sigma_i(v_0 - vi)}{y_c}\}$

Surface Geometry (g_i)

(Hoiem et al 2005) produces "confidence maps" of geometric surface for 3 classes

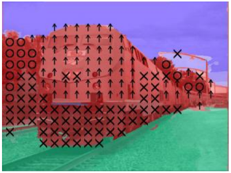


Figure: Geometric Label

- sky
- vertical, contains 5 subclasses
  - planar, facing
    - left (⟵)
    - center (↑)
    - right (⟶)
  - non-planar
    - solid (O)
    - porous (X)
- ground

Define g_i = three values, corresponding to

- whether "object surface" is visible in detection window
- whether the ground is visible just below the detection window

i.e.

- consider "car" as "planar or non-planar solid"
- consider "pedestrian" as "non-planar solid"

Surface Geometry (g_i)

- Compute $P(g_i|o_i)$ and $P(g_i)$ by counting occurrences of value of g_i for both people and cars in training set
  - If o_i is background, consider $P(g_i|o_i) \approx P(g_i)$
- Estimate $P(g_i|e_{gi})$ based on confidence maps.
- They found that "average geometric confidence in a window" is a "well-calibrated probability for the geometric value".

Outline

1 Abstract

2 Introduction

3 Projective geometry

4 Modeling the scene

5 Training + Evaluation

6 Viewpoint by Example

7 Summary

y_c distribution is estimated using kernel density estimation .

- Manually labeled cars (i.e. vans, trucks), pedestrians in 60 outdoor images from LabelMe dataset (Russell et al. 2005).
- Calculate maximum likelihood estimate of camera height (y_c) based on labeled horizon and height distributions of cars and people in world.
- y_c is estimated using "kernel density estimation" (ksdensity in Matlab).

- use local detector (Murphy et al. 2003)
    - same local patch template features + 6 color features (encode average L*a*b*color of detection window, and difference between detection window and surrounding area.)
    - verified using PASCAL challange training/validation set : did not search for car shorter than 10% of image height (the official entries could not detect small cars). Result : average precision of 0.423 (the best scores reported by top 3 groups: 0.613, 0.489, 0.353)

Training

Object Identities (o_i)

- logistic regression version of Adaboost classifier (Collins et al 2002) to boost 8-node decision tree classifiers
  - For car, they trained 2 views (front/back: 32x24 px, side: 40x16 px)
  - For pedestrains, they trained 1 view (16x40 px)

  using full PASCAL dataset (PASCAL 2005)

- Estimate distribution of
  - cars' height (y_i) : use data from Consumer Reports (www.consumerreports.org)
    Result : μ = 1.59 m, σ = 0.21 m
  - pedestrains' height (y_i) : use data from National Center for Health Statistics (www.cdc.edu/nchs)
    Result : μ = 1.7 m, σ = 0.085 m

Training

Surface Geometry (g_i)

- Compute $P(g_i|o_i)$ by counting occurrences of value of g_i for both people and cars in 60 training set from LabelMe
- Set $P(g_i)$ to be uniform , because the learned values of $P(g_i)$ over-relying on geometry.

  The over-reliance may be due to

  - labeled image (general outdoor) being drawn from a different distribution than test set (streets of Boston)
  - lack of modeled direct dependence between surface geometries.

Evaluation

- Test Set : 422 random outdoor images from LabelMe dataset (Russell et al. 2005)
  - no car or pedestrians : 60 images
  - only pedestrians : 40 images
  - only cars : 94 images
  - both cars and pedestrians : 225 images

  Total 923 cars, 720 pedestrians
- Detect cars' height as small as 14 px
  Detect Pedestrians' height as small as 36 px
- To get detection confidences for each window, they reverse the process in (in Object identities slide).
  Then determine bounding boxes of objects in standard way (thresholding the confidences and performing non-maxima suppression)

Viewpoint and Surface geometry aids Object detection

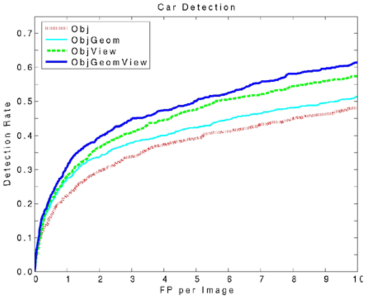
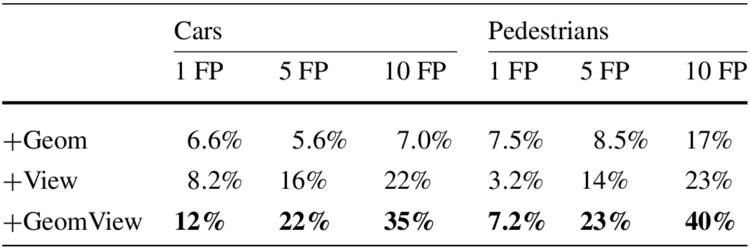


Figure: Object Detection Result : ROC curves

About 20% of cars and pedestrians missed by the baseline are detected for the same FP

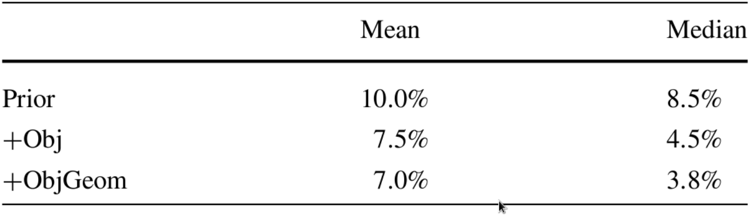


	Cars			Pedestrians		
	1 FP	5 FP	10 FP	1 FP	5 FP	10 FP
+Geom	6.6%	5.6%	7.0%	7.5%	8.5%	17%
+View	8.2%	16%	22%	3.2%	14%	23%
+GeomView	**12%**	**22%**	**35%**	**7.2%**	**23%**	**40%**

Figure: FP : false positive rates, and % reductions in false negatives

Object and geometry improve horizon estimation

Manually labeled horizon of 100 images that contained Car+Ped.



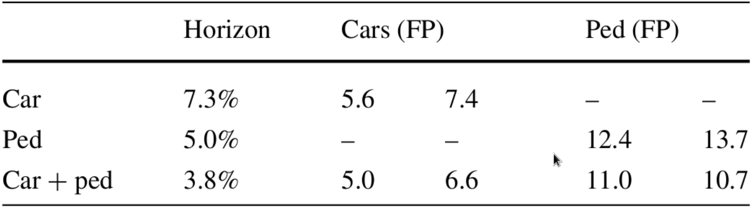
	Mean	Median
Prior	10.0%	8.5%
+Obj	7.5%	4.5%
+ObjGeom	7.0%	3.8%

Figure: Horizon estimation results

Mean/median absolute error (as percentage of image height) are shown for horizon estimates.

Horizon estimation, and object detection are more accurate when "more object models" of the scene are known.

detect only car, only pedestrians, and both



	Horizon	Cars (FP)		Ped (FP)	
Car	7.3%	5.6	7.4	–	–
Ped	5.0%	–	–	12.4	13.7
Car + ped	3.8%	5.0	6.6	11.0	10.7

Figure: (Horizon) : median absolute error, (First Number) : FP/image (at 50% detection rate) computed over all images, (Second Number) : FP/image (at 50% detection rate) computed over subset of images that contain both cars and people

Any local object detector can be plugging into their framework

- After converting SVM outputs to probabilities using method of (Platt 2000).
- Training 80x24 car detector; 32x96 big pedestrian detector, 16x48 small pedestrian detector

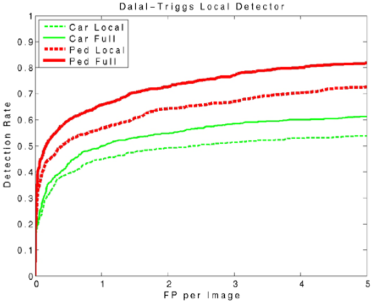
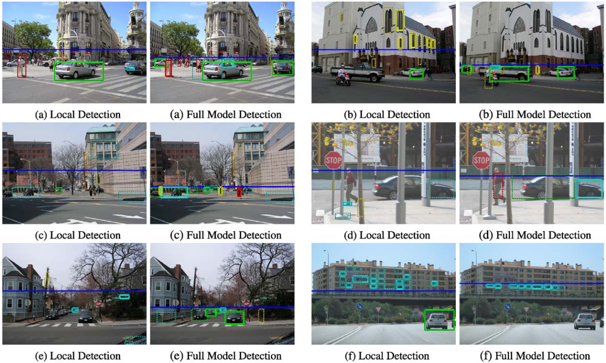


Figure: Detal-Triggs local detector (Detal and Triggs 2005)

Detection Example

baseline local detector (Murphy et al. 2003), and full model detector



blue line = horizon estimate (0.5 initially)
green box = true car, cyan box = false car
red box = true ped, yellow = false ped
solid line – high confidence detection (0.5 FP/Image)
dotted line – low confidence detection (2 FP/Image)

- In most case, horizon line is correctly recovered, obj detection improves considerably.
- Box that make no sense (i.e. wrong scale (d), above horizon (b), in the middle of the ground (e)) is removed.
- Predestrians are often hallucinated (c, e) in places where they could be (but are not).
- (f) : a bad geometry estimate and repeated false detections along the building windows causes the horizon estimate to become worse and the car to be missed

Outline

1 Abstract

2 Introduction

3 Projective geometry

4 Modeling the scene

5 Training + Evaluation

6 Viewpoint by Example

7 Summary

improve estimating v_0

using detections of known objects in scene

If image does not include easily detectable known object ?

- "Edges and perspective geometry" using a lots of parallel lines (Manhattan Worlds) (Kosecka and Zhang 2022), (Coughlan and Yuille 2003)
  - fail for less structured environments.
- using "spatial-frequency" statistics (the "gist") of image to recover a general sense of space of the scene. (Oliva and Torralba 2006)
  - They estimate rough "scene envelop" of properties such as "openess, smoothness, depth".
- "gist" statistics could also provide a rough estimate of "horizon position" (v_0). (Torralba and Sinha 2001)

Use gist descriptor to estimate v_0 instead of Simple Gaussian Prior

To obtain more precise estimate, we require

1. large amount of training set (image/viewpoint pairs)
2. a way of accurately labeling this data, automatically

Automatically labeling data

to get v_0 and new distribution of y_i

using iterative EM-like algorithm (Lalongde te al. 2007)

- initially guess of μ and σ of people's height (y_i)
- iterates
  1. estimate "viewpoint" (v_0) for images that contain objects of known "size distributions" (y_i)
  2. estimate "size distributions" (y_i) of objects that are contained in images with known "viewpoints" (v_0).

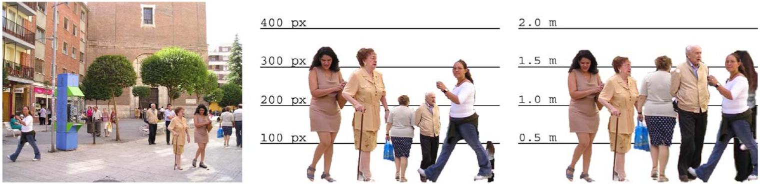


Figure: automatic object height estimation result, (left) picture from LabelMe (center) original pixel size (right) automatically estimated 3D height (y_i)

Automatically labeling data

to get v_0 and new distribution of y_i

- Training : infer viewpoint of over 5,000 images and height of 13,000 object instances in 50 object classes.
- based on inferred object heights (y_i), re-estimate y_i distribution of cars ($\mu = 1.51m$, $\sigma = 0.191m$), people ($\mu = 1.70m$, $\sigma = 0.103m$)
- consider camera viewpoint v_i estimates to be reliable for 2,660 images (excluding images in test set) that contains at least 2 objects with known height distributions.
- calculate gist statistics (8x8 blocks at 3 scales with 8,4,4 orientations, giving 1280 variables per gist vector).

gist descriptor

- Evaluate gist nearest neighbour classifiers with various distance matrics
  - nearest neighbor regression method (Navot et al. 2006)
  - GRANN (Generalized Regression neural networks) – provide lowest error after tuning $\alpha = 0.22$ on training set

The horizon estimation using GRANN ($\tilde{v_0}$)

$\tilde{v_0}(x) = \frac{\sum_i v_{0i} w_i}{\sum_i w_i}$; $w_i = exp[\frac{-||x - x_i||^2}{2\alpha^2}]$

x : gist statistics

v_{0i} : horizontal position for ith training sample

$p(v_0|\tilde{v_0}) = \frac{1}{2s_0} exp[\frac{-|v_0 - \tilde{v_0}|}{s_v}]$ (laplace probability density)

s_v : scale parameter = expected error

- Result : correlation (coefficient 0.27) between error and Euclidean distance to the nearest neighbour in gist statistics.
- Provide better estimate, by consider $\tilde{d}$ (nearest neighbour distance)
  fit $s_v = 0.022 + 0.060\tilde{d}$, by maximum likelihood estimation over training set

Evaluation

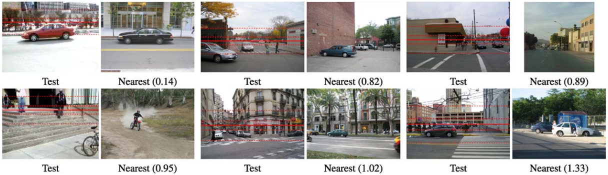
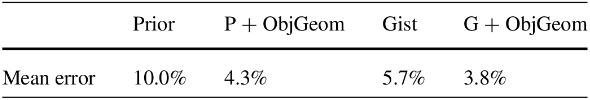


Figure: horizontal estimation using gist nearest neighbour and distance $\tilde{d}$



	Prior	P + ObjGeom	Gist	G + ObjGeom
Mean error	10.0%	4.3%	5.7%	3.8%

Figure: mean error of horizon position estimation (% of image height)

- Gaussian Prior
- Gaussian Prior + surface geometry, object (using Datal-Triggs detectors)
- Gist estimation
- Gist estimation + surface geometry, object (using Datal-Triggs detectors)

Evaluation

improve detection rates (This is measured at 1 FP/image using Dalal-Triggs object detectors)

- cars detection 50 % to 52 %
- pedestrians 66 % to 68 %

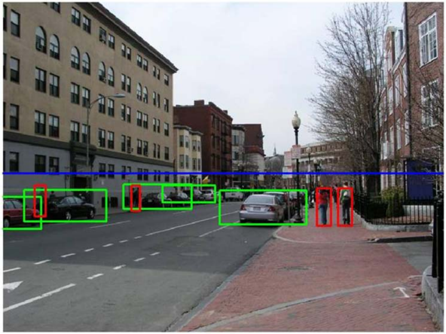
Improve estimating y_c

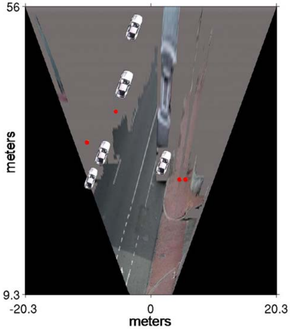
- cannot improve significantly over prior estimate
- may because small change in y_c has little impact on the global statistics.
- Therefore, re-estimate y_c using larger training set

Outline

1. Abstract
2. Introduction
3. Projective geometry
4. Modeling the scene
5. Training + Evaluation
6. Viewpoint by Example
7. Summary

Summary





- Project estimated ground surface into overhead view, using estimated camera viewpoint.
- Assume cars mostly face down the same line.

What they have done : "bag of features/blackbox" classification method

1. subtle relationship (i.e. object sizes relate through the viewpoint) can be easily represented.
2. additions and extensions to the model are easy.

could be extended by modeling other scene properties (i.e. scene category)

Summary

- geometry assumptions
- not consider as conditionally independent, if two object patches are very similar. i.e. repeated windows in (f) causes object detection responses to be correlated. solution : correlation should be modeled. i.e. if two object patches are very similar, consider their evident jointly.
- the surface trained in this experiment is outdorr image only.