Learning Where to Position Parts in 3D

Marco Pedersoli[1,2] Tinne Tuytelaars[1]

[1]ESAT-PSI / iMinds, KU Leuven

[2]LEAR team, Inria Grenoble Rhone-Alpes, Lab. Jean Kuntz-mann, CNRS, Univ. Grenoble Alpes, France.

Introduction

Task: Learn a 3D representation of an object category for object detection and pose/viewpoint estimation.

Motivation: Previous work uses 3D CAD models or other sources of 3D information to estimate the shape of the object. We want to learn a 3D model representation without any explicit 3D info, just from the object location and viewpoint.

Contribution: We show that in the context of deformable models it is possible to augment the latent variables to estimate also the 3D location of the object parts. This produces a coarse estimation of the 3D of the object that improves recognition results in terms of both object detection and pose estimation.

Related work

[1] explicitly associates each part to a 3D landmark from 3D CAD samples.
[2] models the 3D shape as a composition of planar surfaces that are learned from 3D CAD models.
[3] synthesizes a view of an object in the HOG space based on 3D CAD models.
[4] uses annotated 3D landmarks to build a model to estimate the 3D viewpoint of cars.
[5] assumes the object to fit a cuboid representation, while for us that assumption is used only for initialization.

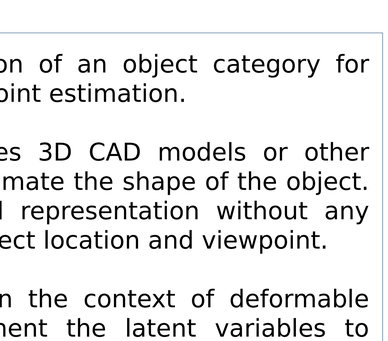
Scoring Function

Object Center: $o = (o_x, o_y, o_z)$

Object Rotation: $\theta = (\theta_x, \theta_y, \theta_z)$

Part Location: $l = (l_x, l_y, l_z)$

Part Orientation: $n = (n_x, n_y, n_z)$



Score of a part relative to camera C:

$$< w, \phi(I, l^C, n^C) >,$$

Considering that:

$$l^C = R_\theta(l^O + o), \qquad n^C = R_\theta n^O$$

the score of a part relative to the object center O is:

$$scr^p(w, l^O, n^O, o, \theta, I) =< w, \phi(I, R_\theta(l^O + o), R_\theta n^O) >,$$

Score of the entire object with respect to O:

$$scr(W, L, N, o, \theta, I) = \sum_i scr^p(w_i, l_i, n_i, o, \theta)$$

with: $L = \{l_0, l_1, ..., l_P\}$ $N = \{n_0, n_1, ..., n_P\}$ $W = \{w_0, w_1, ..., w_P\}$

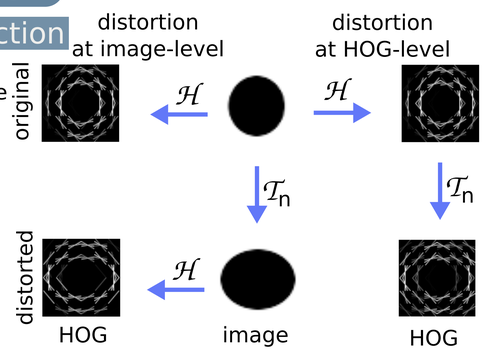
Feature Extraction

Orthographic Projection

Object parts of the same object have the same size which is defined by the distance of the object center to the camera.

HOG features from Image I at location (x,y)
$\mathcal{H}_s(I, x, y)$

Distortion assuming orthogonal projection
$\mathcal{T}_{n^C}(I(x, y)) = I(\eta_x x, \eta_y y)$



Camera Distortion applied at pixel level:
$\phi(I, l^C, n^C) = \mathcal{H}_{l_z^C}(\mathcal{T}_{n^C}(I), l_x^C, l_y^C),$

Camera distortion applied at HOG level:
$\mathcal{H}_{l_z^C}(\mathcal{T}_{n^C}(I), l_x^C, l_y^C) \approx \mathcal{H}_{l_z^C}(I, \eta_x l_x^C, \eta_y l_y^C)$

Speed Up

Quantization error does not affect performance, but speed! Precomputing HOG cells and interpolation gain of almost 16X as in [6]!

3D Deformation

Quadratic deformation cost in 3D defined by (d_x,d_y,d_z) for each part i with location l_i and part displacement m_i=(m_x,m_y,m_z).

$$scr(W, L, N, o, \theta, I) = \sum_i \max_{m_i} \left(scr^p(w_i, l_i + m_i, n_i, \theta, I) - m_i^T Q_{\theta, n_i} m_i\right)$$

$Q_{\theta,n}$ is the 3x3 matrix that encodes the learnable deformation parameters (d_x,d_y,d_z) and the transformations induced by the object rotation θ and the part orientation n_i.

Due to the orthographic projection the 3D deformation cost can be reduced to its 2D projection on the camera plane

Generalized Distance Transform

On the camera plane we can still use the fast 2D distance transform

Deformations are now 2D projection of a 3D quadratic form

We extend the standard distance transform [7] to non aligned axes

Learning

Objective

Annotation s=(Image I_s, label y_s, location o_s, pose θ_s)

W = parts appearance and deformation parameters (to learn)

N = parts orientation (given) L = parts location (given)

$$|W|^2 + C \sum_s \max(0, 1 - y_s scr(W, L, N, o_s, \theta_s, I_s)),$$

We minimize it using coordinate descent and negative mining [8].

Estimation of the parts depth t_z

The location of a part is updated along n_z, the normal of the part.

t_z = displacement of part anchor l_s (latent variable)

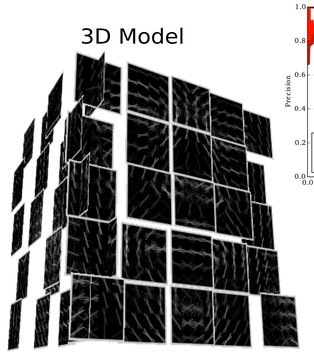
$m_{z,p}$ = displacement of the part with respect to l_s (latent var.)

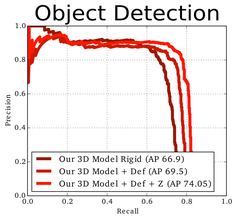
$$l_s^O = l_s^O + t_z n_z \qquad t_z = \arg\min_t \sum_{p \in \mathcal{P}} d_z |m_{z,p} - t|^2$$

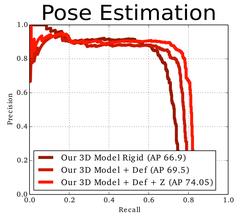
Results

Faces on AFW

3D Model







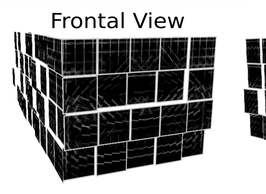
Method	Detection	Pose 15	Pose 30
multiview HOG [9]	75.5	74.6	85.0
3D model from [6]	78.8	71.4	-
TSM [9]	88.0	81.0	89.0
TSM shared [9]	76.2	76.9	87.0
Fisher [10]	88.3	78.6	90.6
DeCaf [10]	88.3	86.5	93.4
Our 3D model	90.18	85.9	92.1

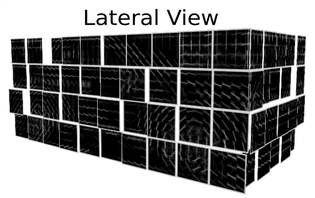
Using 3D deformation and estimating the depth of the parts improve both object detection and pose estimation.

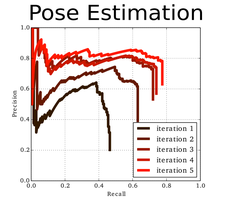
Cars on EPFL

Mean Precision Pose Estim.

Method	MPPE8	MPPE16	MPPE36
MDPM [1]	73.7	66.0	-
3D^{2}PM [1]	77.9	69.1	53.5
Fisher [10]	76.6	72.2	51.8
DeCaf [10]	80.6	67.8	45.9
Our 3D model	81.5	56.4	36.8







Remarks

Estimated a coarse representation of the 3D shape of an object class without any explicit 3D model or annotation

Proposed model faster than previous approaches based on HOG models, but it still requires from 10 to 30 secs on a single CPU

When the 3D class has not only multiple views but also multiple appearances, the 3D estimation of the parts fails

References

[1] B. Pepik, P. Gehler, M. Stark, and B. Schiele. 3dpm - 3d deformable part models. In ECCV 2012.
[2] Y. Xiang and S. Savarese. Estimating the aspect layout of object categories. In ICCV 2012.
[3] Hejrati and D. Ramanan. Analysis by synthesis: 3d object recognition by object reconstruction. In CVPR 2014.
[4] M. Hejrati and D. Ramanan. Analyzing 3d objects in cluttered images. In NIPS 2012.
[5] S. Fidler, S. Dickinson, and R. Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model, In NIPS 2012.
[6] M. Pedersoli and T. Tuytelaars. A scalable 3d hog model for fast object detection and viewpoint estimation. In 3DV 2014.
[7] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical report 2004.
[8] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. PAMI 2010.
[9] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In CVPR 2012.
[10] A. Ghodrati, M. Pedersoli, and T. Tuytelaars. Is 2d information enough for viewpoint estimation? In BMVC 2014.