Factored Shapes and Appearances for Parts-based Object Understanding

S. M. Ali Eslami
Christopher K. I. Williams

THE UNIVERSITY *of* EDINBURGH
informatics
ianc | Institute for Adaptive and Neural Computation

British Machine Vision Conference
September 2, 2011


The longest-running musical ever.
SKECHERS
DANIEL CRAIG
THE ONE THAT YOU WANT
WICKED
PALACE
LEGALLY BLONDE
CHICAGO
McDonald's
WHERE MAXIMUM PERFORMANCE LIVES
maxell
Broadway
A PROUD SPONSOR OF
X GAMES
FRIDAY'S
tkts
North Fork Bank

Classification

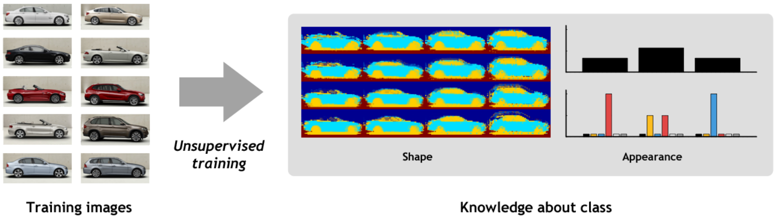
Localisation

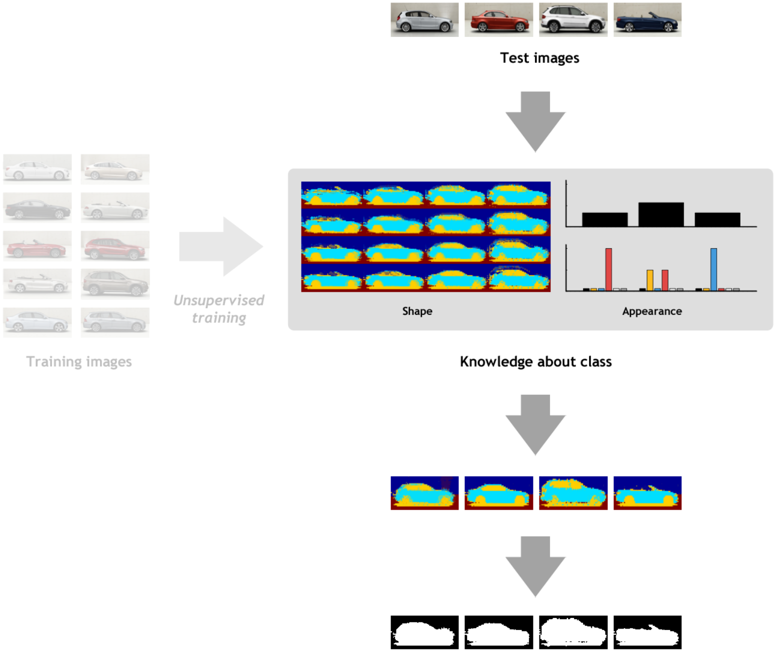
Segmentation

This talk's focus

(Panoramio/nicho593)

Segment this


Unsupervised training
Shape
Appearance
Training images
Knowledge about class

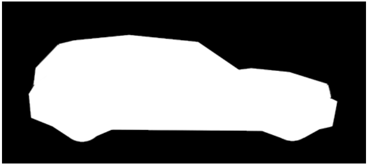

Test images
Training images
Unsupervised training
Shape
Appearance
Knowledge about class

Outline

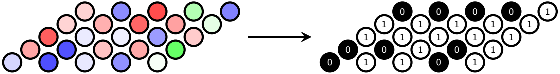
1. The segmentation task
2. The FSA model
3. Experimental results
4. Discussion

The segmentation task

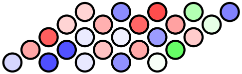
The image $\mathbf{X}$



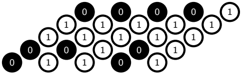
The segmentation $\mathbf{S}$



The segmentation task



The image $\mathbf{X}$

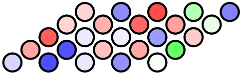


The segmentation $\mathbf{S}$

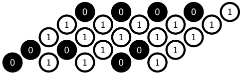
The generative approach

- Construct a joint model of $\mathbf{X}$ and $\mathbf{S}$ parameterised by $\boldsymbol{\theta}$: $p(\mathbf{X}, \mathbf{S}|\boldsymbol{\theta})$
- Learn $\boldsymbol{\theta}$ given dataset $\mathbf{D}^{\text{train}}$: $\arg\max_{\boldsymbol{\theta}} p(\mathbf{D}^{\text{train}}|\boldsymbol{\theta})$
- Return probable segmentation $\mathbf{S}^{\text{test}}$ given $\mathbf{X}^{\text{test}}$ and $\boldsymbol{\theta}$: $p(\mathbf{S}^{\text{test}}|\mathbf{X}^{\text{test}}, \boldsymbol{\theta})$

The segmentation task



The image $\mathbf{X}$



The segmentation $\mathbf{S}$

The generative approach

- Construct a joint model of $\mathbf{X}$ and $\mathbf{S}$ parameterised by $\boldsymbol{\theta}$: $p(\mathbf{X}, \mathbf{S}|\boldsymbol{\theta})$
- Learn $\boldsymbol{\theta}$ given dataset $\mathbf{D}^{\text{train}}$: $\arg\max_{\boldsymbol{\theta}} p(\mathbf{D}^{\text{train}}|\boldsymbol{\theta})$
- Return probable segmentation $\mathbf{S}^{\text{test}}$ given $\mathbf{X}^{\text{test}}$ and $\boldsymbol{\theta}$: $p(\mathbf{S}^{\text{test}}|\mathbf{X}^{\text{test}}, \boldsymbol{\theta})$

Some benefits of this approach

- Flexible with regards to data:
  - Unsupervised training,
  - Semi-supervised training.
- Can inspect quality of model by sampling from it.

Factored Shapes and Appearances

Goal

Construct a joint model of **X** and **S** parameterised by $\boldsymbol{\theta}$: $p(\mathbf{X}, \mathbf{S} | \boldsymbol{\theta})$.

Factor appearances

- Reason about object **shape** independently of its **appearance**.

Factored Shapes and Appearances

Goal

Construct a joint model of $\mathbf{X}$ and $\mathbf{S}$ parameterised by $\boldsymbol{\theta}$: $p(\mathbf{X}, \mathbf{S}|\boldsymbol{\theta})$.

Factor appearances

- Reason about object **shape** independently of its **appearance**.

Factor shapes

- Represent objects as collections of **parts**,
- Systematic **combination** of parts generates objects' complete shapes.

Factored Shapes and Appearances

Goal

Construct a joint model of $\mathbf{X}$ and $\mathbf{S}$ parameterised by $\boldsymbol{\theta}$: $p(\mathbf{X}, \mathbf{S}|\boldsymbol{\theta})$.

Factor appearances

- Reason about object **shape** independently of its **appearance**.

Factor shapes

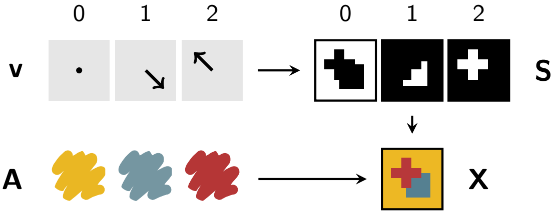
- Represent objects as collections of **parts**,
- Systematic **combination** of parts generates objects' complete shapes.

Learn everything

- Explicitly model **variation** of appearances and shapes.

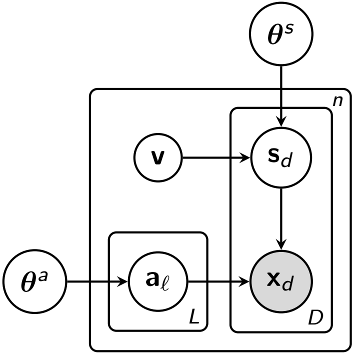
Factored Shapes and Appearances

Schematic diagram



Factored Shapes and Appearances

Graphical model



n number of images

L parts

D pixels in each image

Parameters

$\boldsymbol{\theta}^s$ – shape statistics

$\boldsymbol{\theta}^a$ – appearance statistics

Latent variables

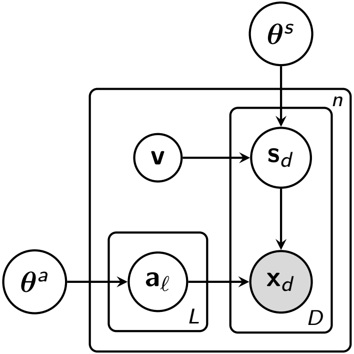
$\mathbf{a}_\ell$ – per part appearance

$\mathbf{v}$ – global shape type

$\mathbf{s}$ – segmentation

Factored Shapes and Appearances

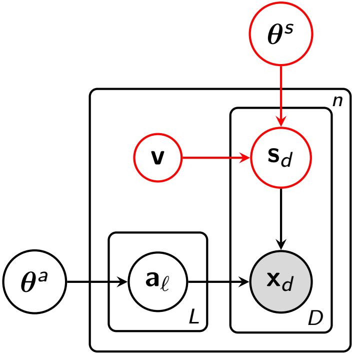
Shape model



$$p(\mathbf{X}, \mathbf{A}, \mathbf{S}, \mathbf{v} | \theta) = p(\mathbf{v})\, p(\mathbf{A}|\theta^a) \prod_{d=1}^{D} p(\mathbf{s}_d | \mathbf{v}, \theta^s)\, p(\mathbf{x}_d | \mathbf{A}, \mathbf{s}_d, \theta^a)$$

Factored Shapes and Appearances

Shape model



$$p(\mathbf{X}, \mathbf{A}, \mathbf{S}, \mathbf{v} | \theta) = p(\mathbf{v})\, p(\mathbf{A}|\theta^a) \prod_{d=1}^{D} p(\mathbf{s}_d | \mathbf{v}, \theta^s)\, p(\mathbf{x}_d | \mathbf{A}, \mathbf{s}_d, \theta^a)$$

Factored Shapes and Appearances

Shape model

Continuous parameterisation

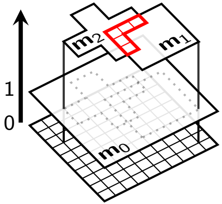
$$p(s_{\ell d} = 1|\mathbf{v}, \boldsymbol{\theta}) = \frac{\exp\{m_{\ell d}\}}{\sum_{k=0}^{L} \exp\{m_{k d}\}}$$

Efficient

- Finds probable assignment of pixels to parts without having to enumerate all part depth orderings.
- Resolve ambiguities by exploiting knowledge about appearances.

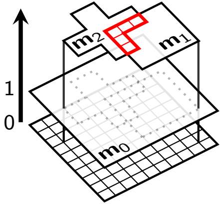
Factored Shapes and Appearances

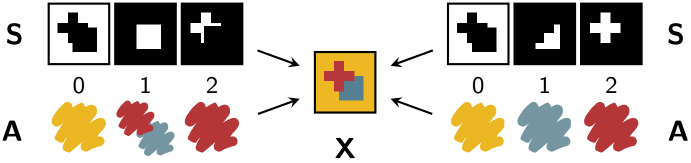
Handling occlusion



Factored Shapes and Appearances

Handling occlusion





Factored Shapes and Appearances

Learning shape variability

Goal

Instead of learning just a template for each part, learn a *distribution* over such templates.

Linear latent variable model

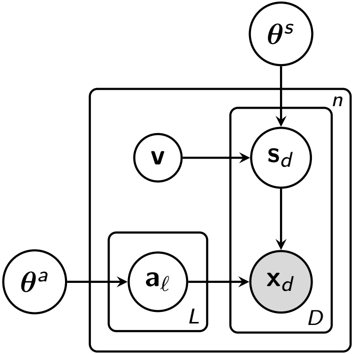
Part ℓ's mask $\mathbf{m}_\ell$ is governed by a Factor Analysis-like distribution:

$$
\begin{aligned}
p(\mathbf{v}) &= \mathcal{N}(\mathbf{0}, \mathbf{I}_{H \times H}) \\
\mathbf{m}_\ell &= \mathbf{F}_\ell \mathbf{v} + \mathbf{c}_\ell,
\end{aligned}
$$

where $\mathbf{v}_\ell$ is a low-dimensional latent variable, $\mathbf{F}_\ell$ is the factor loading matrix and $\mathbf{c}_\ell$ is the mean mask. Shape parameters $\boldsymbol{\theta}^s = \{\{\mathbf{F}_\ell\}, \{\mathbf{c}_\ell\}\}$.

Factored Shapes and Appearances

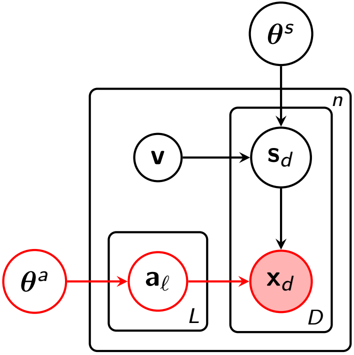
Appearance model



$$p(\mathbf{X}, \mathbf{A}, \mathbf{S}, \mathbf{v} | \theta) = p(\mathbf{v})\, p(\mathbf{A}|\theta^a) \prod_{d=1}^{D} p(\mathbf{s}_d | \mathbf{v}, \theta^s)\, p(\mathbf{x}_d | \mathbf{A}, \mathbf{s}_d, \theta^a)$$

Factored Shapes and Appearances

Appearance model



$$p(\mathbf{X}, \mathbf{A}, \mathbf{S}, \mathbf{v} | \theta) = p(\mathbf{v})\, p(\mathbf{A} | \boldsymbol{\theta}^a) \prod_{d=1}^{D} p(\mathbf{s}_d | \mathbf{v}, \boldsymbol{\theta}^s)\, p(\mathbf{x}_d | \mathbf{A}, \mathbf{s}_d, \boldsymbol{\theta}^a)$$

Factored Shapes and Appearances

Appearance model

Goal

Learn a model of each part's RGB values that is as informative as possible about its extent in the image.

Position-agnostic appearance model

- Learn about distribution of colours *across* images,
- Learn about distribution of colours *within* images.

Factored Shapes and Appearances

Appearance model

Goal

Learn a model of each part's RGB values that is as informative as possible about its extent in the image.

Position-agnostic appearance model

- Learn about distribution of colours *across* images,
- Learn about distribution of colours *within* images.

Sampling process

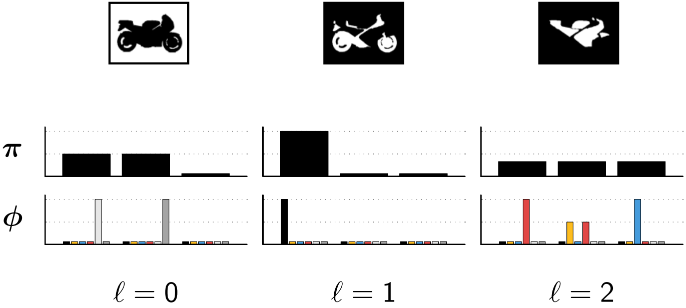
For each part:

1. Sample an appearance 'class' for the current part,
2. Sample the part's pixels from the current class' feature histogram.

Factored Shapes and Appearances

Appearance model

Training data



Factored Shapes and Appearances

Learning

Use **EM** to find a setting of the shape and appearance parameters that approximately maximises their likelihood given the data $p(\mathbf{D}^{\text{train}}|\boldsymbol{\theta})$:

1. **Expectation:** Block Gibbs and elliptical slice sampling (Murray et al., 2010) to approximate $p(\mathbf{Z}^i|\mathbf{X}^i, \boldsymbol{\theta}^{\text{old}})$,
2. **Maximisation:** Gradient descent optimisation to find $\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ where

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{i=1}^{n} \sum_{\mathbf{Z}^i} p(\mathbf{Z}^i|\mathbf{X}^i, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}^i, \mathbf{Z}^i|\boldsymbol{\theta}).$$

Related work

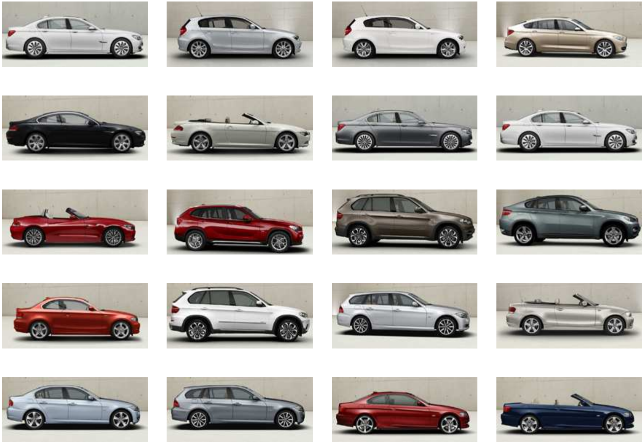
	FACTORED PARTS	FACTORED SHAPE AND APPEARANCE	SHAPE VARIABILITY	APPEARANCE VARIABILITY
LSM Frey et al.	✓ (layers)	-	✓ (FA)	✓ (FA)
Sprites Williams and Titsias	✓ (layers)	-	-	-
LOCUS Winn and Jojic	-	✓	✓ (deformation)	✓ (colours)
MCVQ Ross and Zemel	-	✓	-	✓ (templates)
SCA Jojic et al.	-	✓	✓ (convex)	✓ (histograms)
FSA	✓ (softmax)	✓	✓ (FA)	✓ (histograms)

Outline

1. The segmentation task
2. The FSA model
3. Experimental results
4. Discussion

Learning a model of cars

Training images



Learning a model of cars

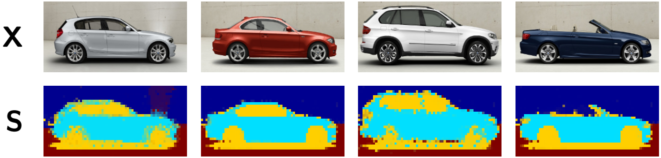
Model details

- Number of parts $L = 3$,
- Number of latent shape dimensions $H = 2$,
- Number of appearance classes $K = 5$.

Learning a model of cars

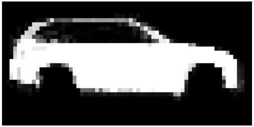
Model details

- Number of parts $L = 3$,
- Number of latent shape dimensions $H = 2$,
- Number of appearance classes $K = 5$.

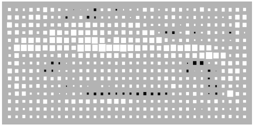


Learning a model of cars

Shape model weights

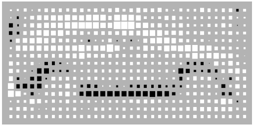


$\ell = 2$



$\mathbf{F}_2$ column 1

Convertible $\longleftrightarrow$ Coupé

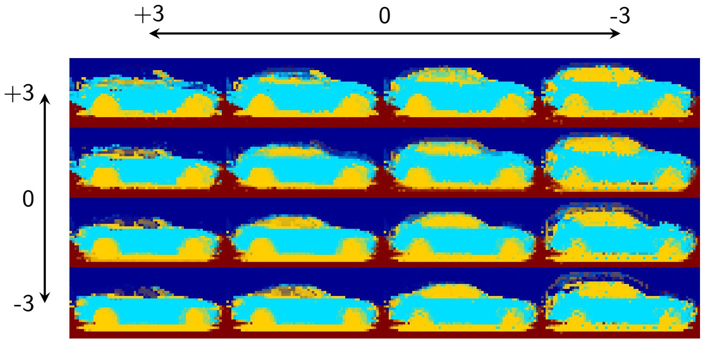


$\mathbf{F}_2$ column 2

Low $\longleftrightarrow$ High

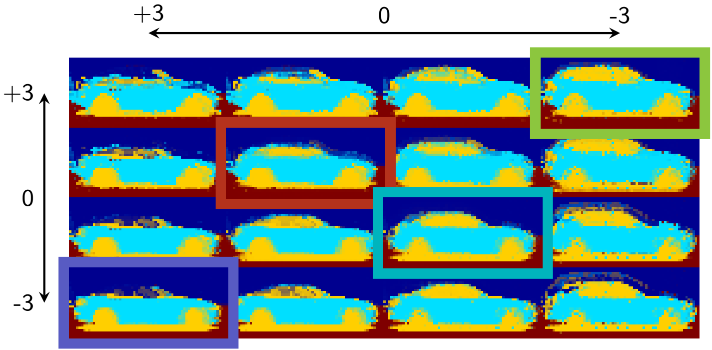
Learning a model of cars

Latent shape space



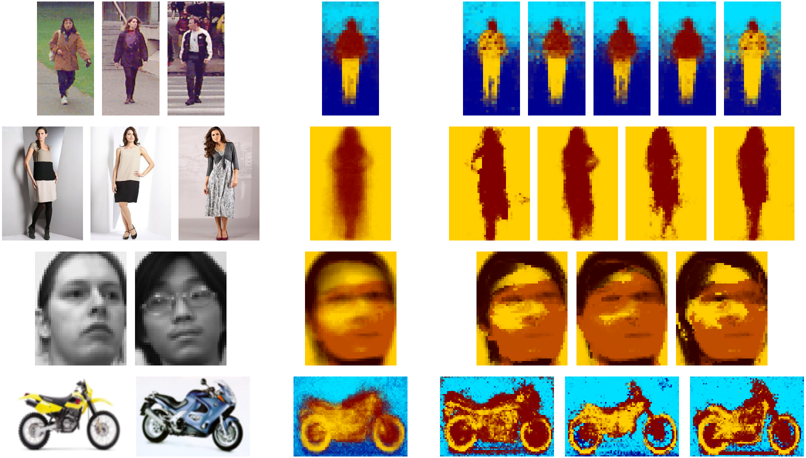
Learning a model of cars

Latent shape space



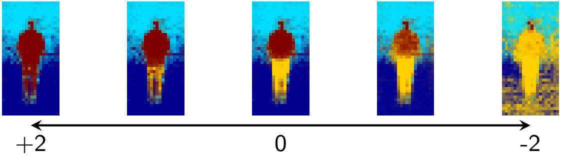
Saloon – Hatchback – Convertible – SUV

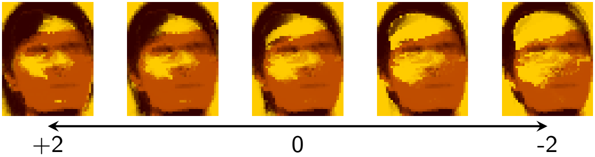
Other datasets



Training data Mean model FSA samples

Other datasets





Segmentation benchmarks

Datasets

- **Weizmann horses**: 127 train – 200 test.
- **Caltech4**
  - Cars: 63 train – 60 test,
  - Faces: 335 train – 100 test,
  - Motorbikes: 698 train – 100 test,
  - Airplanes: 700 train – 100 test.

Two variants

- **Unsupervised FSA**: Train given only RGB images.
- **Supervised FSA**: Train using RGB images *and* their binary masks.

Segmentation benchmarks

	Weizmann		Caltech4		
	Horses	Cars	Faces	Motorbikes	Airplanes
GrabCut Rother et al.	83.9%	45.1%	83.7%	82.4%	84.5%
Borenstein et al.	**93.6%**	-	-	-	-
LOCUS Winn et al.	93.1%	91.4%	-	-	-
Arora et al.	-	**95.1%**	92.4%	83.1%	**93.1%**
ClassCut Alexe et al.	86.2%	93.1%	89.0%	90.3%	89.8%
Unsupervised FSA	87.3%	82.9%	88.3%	85.7%	88.7%
Supervised FSA	88.0%	93.6%	**93.3%**	**92.1%**	90.9%

Competitive – despite lack of CRF-style pixelwise dependency terms.

Summary

FSA is a probabilistic, generative model of images that

- Reasons about object **shape** independently of its **appearance**,
- Represent objects as collections of **parts**,
- Explicitly models **variation** of both appearances and shapes.

Object segmentation with FSA is competitive.

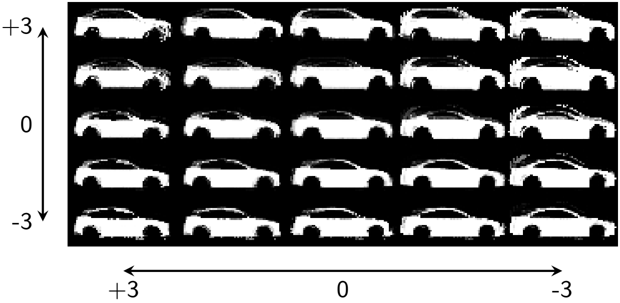
The same FSA model can potentially also be used to

- **Classify** objects into sub-categories (using latent $\mathbf{v}$ variables),
- **Localise** objects (using a sliding window or branch and bound),
- **Parse** objects into meaningful parts.

Questions

Learning a **supervised** model of cars

Latent shape space



Bibliography I

Alexe, B., Deselaers, T., and Ferrari, V. (2010). ClassCut for unsupervised class segmentation. In *Proceedings of the 11th European conference on Computer vision: Part V*, pages 380–393.

Arora, H., Loeff, N., Forsyth, D., and Ahuja, N. (2007). Unsupervised Segmentation of Objects using Efficient Learning. *IEEE Conference on Computer Vision and Pattern Recognition 2007*, pages 1–7.

Borenstein, E., Sharon, E., and Ullman, S. (2004). Combining Top-Down and Bottom-Up Segmentation. In *CVPR Workshop on Perceptual Organization in Computer Vision*.

Frey, B., Jojic, N., and Kannan, A. (2003). Learning appearance and transparency manifolds of occluded objects in layers. In *IEEE Conference on Computer Vision and Pattern Recognition 2003*, pages 45–52.

Jojic, N., Perina, A., Cristani, M., Murino, V., and Frey, B. (2009). Stel component analysis: Modeling spatial correlations in image class structure. In *IEEE Conference on Computer Vision and Pattern Recognition 2009*, pages 2044–2051.

Murray, I., Adams, R. P., and MacKay, D. J. (2010). Elliptical slice sampling. *Journal of Machine Learning Research*, 9:541–548.

Bibliography II

Ross, D. and Zemel, R. (2006). Learning Parts-Based Representations of Data. *Journal of Machine Learning Research*, 7:2369–2397.

Williams, C. K. and Titsias, M. (2004). Greedy learning of multiple objects in images using robust statistics and factorial learning. *Neural Computation*, 16(5):1039–1062.

Winn, J. and Jojic, N. (2005). LOCUS: Learning object classes with unsupervised segmentation. In *International Conference on Computer Vision 2005*, pages 756–763.