Hierarchical Probabilistic Models for Object Segmentation

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

Institute for Adaptive and Neural Computation
School of Informatics
The University of Edinburgh

August 8, 2010

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

Classification

Localisation

Segmentation

Chicken and egg problem

$$
\begin{tikzcd}
 & \textbf{Localisation} \arrow[dl, leftrightarrow] \arrow[dr, leftrightarrow] & \\
\textbf{Segmentation} \arrow[rr, leftrightarrow] & & \textbf{Classification}
\end{tikzcd}
$$

Chicken and egg problem

(Panoramio/nicho593)

What is this?

Chicken and egg problem

(Panoramio/nicho593)

Segment this

Outline

1. The task
2. Related research
3. The approach
4. Current progress
5. Discussion

The Segmentation Task

(Pascal VOC, Everingham et al., 2010)

The segmentation task

Object class labelling

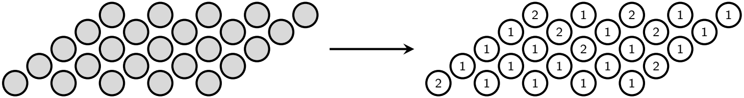
The segmentation task

Foreground/background labelling

The segmentation task

The image **X**

The segmentation **S**

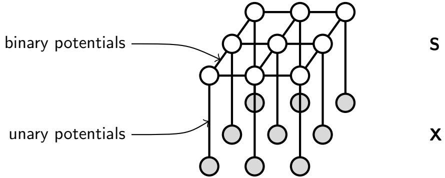


Outline

1. The task
2. Related research
3. The approach
4. Current progress
5. Discussion

Related research

- Continuity-based methods

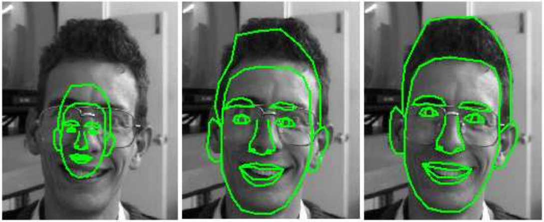


$$p(\mathbf{X},\mathbf{S}) \quad \text{or} \quad p(\mathbf{S}|\mathbf{X}) = \frac{1}{Z}\exp\{-E(\mathbf{X},\mathbf{S})\}$$

- Shape-based methods
  - Global models of shape
  - Parts-based models of shape

Related research

- Continuity-based methods
- Shape-based methods
  - Global models of shape

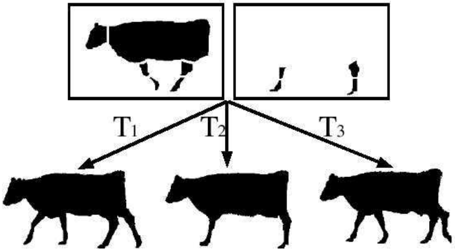


Active Shape and Appearance Models (Cootes et al., 1995)

  - Parts-based models of shape

Related research

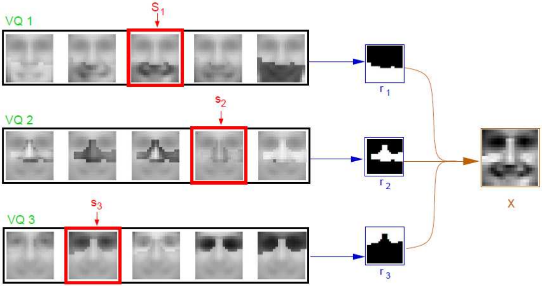
- Continuity-based methods
- Shape-based methods
  - Global models of shape
  - Parts-based models of shape



Layered Pictorial Structures (Kumar et al., 2005)

Related research

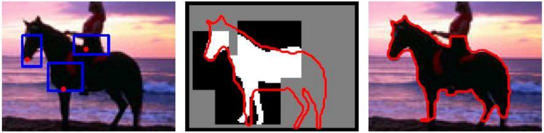
- Continuity-based methods
- Shape-based methods
  - Global models of shape
  - Parts-based models of shape



Multiple Cause Vector Quantization (Ross and Zemel, 2006)

Related research

- Continuity-based methods
- Shape-based methods
  - Global models of shape
  - Parts-based models of shape



Fragment CRF (Levin and Weiss, 2009)

Related research

Summary

Model	Continuity	Shape	Parts	Part shape
LSM (Frey et al., 2003)		✓– FA		
ISM (Leibe et al., 2004)		✓– fragments	✓	$\sim$ – exemplars
GrabCut (Rother et al., 2004)	✓			
OBJCUT (Kumar et al., 2005)	✓	✓– PS	✓	
LOCUS (Winn and Jojic, 2005)	✓	✓– mask		
LHRF (Kapoor and Winn, 2006)	✓	✓– part biases	✓	$\sim$ – CRF
LCRF (Winn and Shotton, 2006)	✓			
SPCRF (Fulkerson et al., 2009)	✓			
FCRF (Levin and Weiss, 2009)	✓	✓– fragments	✓	$\sim$ – exemplars
DPMCRF (Larlus et al., 2009)	✓	✓– DPM		

Related research

Summary

Model	Continuity	Shape	Parts	Part shape
LSM (Frey et al., 2003)		✓– FA		
ISM (Leibe et al., 2004)		✓– fragments	✓	∼ – exemplars
GrabCut (Rother et al., 2004)	✓			
OBJCUT (Kumar et al., 2005)	✓	✓– PS	✓	
LOCUS (Winn and Jojic, 2005)	✓	✓– mask		
LHRF (Kapoor and Winn, 2006)	✓	✓– part biases	✓	∼ – CRF
LCRF (Winn and Shotton, 2006)	✓			
SPCRF (Fulkerson et al., 2009)	✓			
FCRF (Levin and Weiss, 2009)	✓	✓– fragments	✓	∼ – exemplars
DPMCRF (Larlus et al., 2009)	✓	✓– DPM		

Outline

Approach

Shape model type

Three dimensional

Two dimensional

Concerned with tractability

Approach

Part shape variability

Need to model part shape variability

Approach

Aspect variability

Rectangular Circular

Same object, different outlines

Approach

Summary

Model overview

1. Capture the object's shape using a number of **deformable** parts,
2. Combine models of different viewpoints in a mixture,
3. Use this as prior on a random field.

Goal

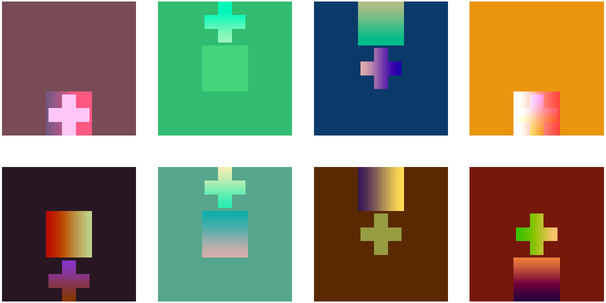
Learning of **dense** object class shape and parts from variable, realistic datasets of images.

- Useful for both **object segmentation** and **object parsing**.
- More expressive power.

Current progress

1. The task
2. Related research
3. The approach
4. Current progress
5. Discussion

Multiple Transformed Masks and Appearances

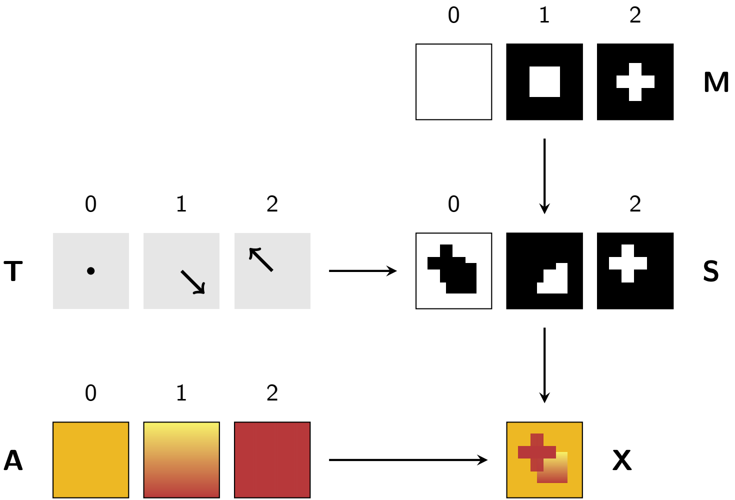


Task

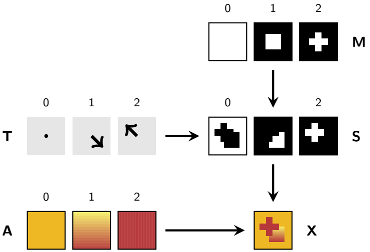
To learn the shapes of the parts and infer their positions and appearances.

Multiple Transformed Masks and Appearances

Schematic diagram



Multiple Transformed Masks and Appearances



$$p(s_{\ell d} = 1 | \mathbf{T}, \boldsymbol{\theta}) = \frac{(\mathbf{T}_\ell \, \mathbf{m}_\ell)_d}{\sum_{k=0}^{L} (\mathbf{T}_k \, \mathbf{m}_k)_d}$$

$$p(\mathbf{x}_d | \mathbf{A}, \mathbf{s}_d) = \prod_{l=0}^{L} \mathcal{N}(\mathbf{x}_d; (\mathbf{W}\mathbf{a}_\ell + \boldsymbol{\mu})_d, \boldsymbol{\Psi}_d)^{s_{\ell d}}$$

Multiple Transformed Masks and Appearances

Learning

$$\mathbf{Z}^i = \{\mathbf{A}^i, \mathbf{S}^i, \mathbf{T}^i\}$$
$$\boldsymbol{\theta} = \{\mathbf{M}\}$$

Use **Expectation Maximisation** algorithm to find a setting of the masks that approximately maximises the likelihood of the parameters given the data $p(\mathbf{D}|\boldsymbol{\theta})$:

1. **Expectation:** Evaluate $p(\mathbf{Z}^i|\mathbf{X}^i, \boldsymbol{\theta}^{\text{old}})$,
2. **Maximisation:** Find $\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ where

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

Multiple Transformed Masks and Appearances

Inference

Goal

Wish to find $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) = p(\mathbf{A},\mathbf{S},\mathbf{T}|\mathbf{X},\boldsymbol{\theta})$.

Approximate

Instead approximate $p(\mathbf{A},\mathbf{S},\mathbf{T}|\mathbf{X},\boldsymbol{\theta})$ by sampling in two steps:

1. Approximate $p(\mathbf{T}|\mathbf{X},\boldsymbol{\theta})$ and draw $K_{\mathbf{T}|\mathbf{X}}$ samples of $\mathbf{T}$,
2. For each sample $\mathbf{T}^{(k)}$, draw from $K_{\mathbf{A},\mathbf{S}|\mathbf{T}}$ samples from $p(\mathbf{S}|\mathbf{A},\mathbf{T},\mathbf{X},\boldsymbol{\theta})$ and $p(\mathbf{A}|\mathbf{S},\mathbf{T},\mathbf{X},\boldsymbol{\theta})$.

$$p(\mathbf{A},\mathbf{S},\mathbf{T}|\mathbf{X},\boldsymbol{\theta}) \simeq \frac{1}{K_{\mathbf{T}|\mathbf{X}}} \sum_{k_1=1}^{K_{\mathbf{T}|\mathbf{X}}} \frac{1}{K_{\mathbf{A},\mathbf{S}|\mathbf{T}}} \sum_{k_2=1}^{K_{\mathbf{A},\mathbf{S}|\mathbf{T}}} \delta(\mathbf{A}^{(k_2)},\mathbf{S}^{(k_2)},\mathbf{T}^{(k_1)})$$

Multiple Transformed Masks and Appearances

Goal

Wish to find $p(\mathbf{Z}|\mathbf{X}, \theta) = p(\mathbf{A}, \mathbf{S}, \mathbf{T}|\mathbf{X}, \theta)$.

Approximate

Instead approximate $p(\mathbf{A}, \mathbf{S}, \mathbf{T}|\mathbf{X}, \theta)$ by sampling in two steps:

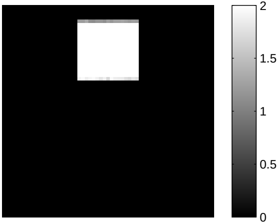
1. Approximate $p(\mathbf{T}|\mathbf{X}, \theta)$ and draw $K_{\mathbf{T}|\mathbf{X}}$ samples of $\mathbf{T}$,
   - Naïve implementation exponential in L, use greedy algorithm (Williams and Titsias, 2004) instead.
2. For each sample $\mathbf{T}^{(k)}$, draw from $K_{\mathbf{A},\mathbf{S}|\mathbf{T}}$ samples from $p(\mathbf{S}|\mathbf{A}, \mathbf{T}, \mathbf{X}, \theta)$ and $p(\mathbf{A}|\mathbf{S}, \mathbf{T}, \mathbf{X}, \theta)$.

$$p(\mathbf{A}, \mathbf{S}, \mathbf{T}|\mathbf{X}, \theta) \simeq \frac{1}{K_{\mathbf{T}|\mathbf{X}}} \sum_{k_1=1}^{K_{\mathbf{T}|\mathbf{X}}} \frac{1}{K_{\mathbf{A},\mathbf{S}|\mathbf{T}}} \sum_{k_2=1}^{K_{\mathbf{A},\mathbf{S}|\mathbf{T}}} \delta(\mathbf{A}^{(k_2)}, \mathbf{S}^{(k_2)}, \mathbf{T}^{(k_1)})$$

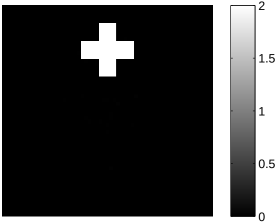
Multiple Transformed Masks and Appearances

Results

- Dataset of 30 images: $n = 30$.
- Transformations discretised into 3 vertical translations: $J = 3$.
- Running time $\sim$3 minutes: 10 EM iterations.



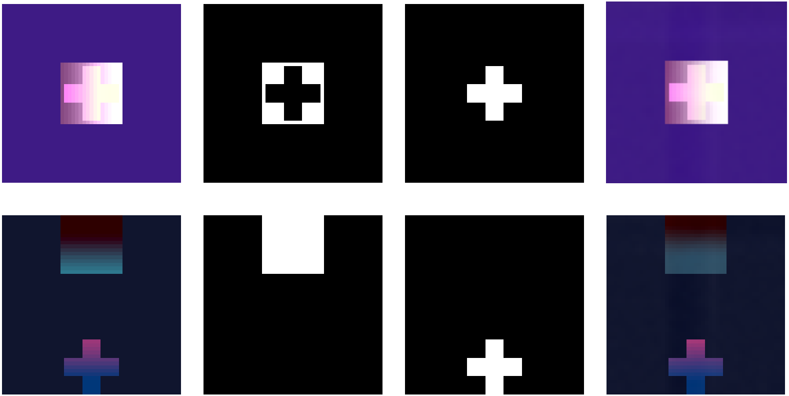
Mask for layer 1, $\mathbf{m}_1$



Mask for layer 2, $\mathbf{m}_2$

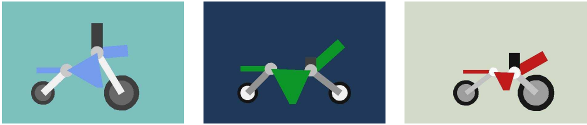
Multiple Transformed Masks and Appearances

Results



Future work

1. Learning inter-part relationships.



2. Incorporating richer part shape models.
3. Determining the number of parts.
4. Incorporating low-level image features.
5. Modelling aspect variability.

Future work

1. Learning inter-part relationships.
2. Incorporating richer part shape models.

3. Determining the number of parts.
4. Incorporating low-level image features.
5. Modelling aspect variability.

Future work

1. Learning inter-part relationships.
2. Incorporating richer part shape models.
3. Determining the number of parts.

4. Incorporating low-level image features.
5. Modelling aspect variability.

Future work

1. Learning inter-part relationships.
2. Incorporating richer part shape models.
3. Determining the number of parts.
4. Incorporating low-level image features.

5. Modelling aspect variability.

Future work

1. Learning inter-part relationships.
2. Incorporating richer part shape models.
3. Determining the number of parts.
4. Incorporating low-level image features.
5. Modelling aspect variability.

Questions

Bibliography I

Cootes, T., Taylor, C., Cooper, D. H., and Graham, J. (1995). Active shape models—their training and application. *Computer Vision and Image Understanding*, 61:38–59.

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J., and Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88:303–338.

Frey, B. J., 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.

Fulkerson, B., Vedaldi, A., and Soatto, S. (2009). Class Segmentation and Object Localization with Superpixel Neighborhoods. In *International Conference on Computer Vision 2009*, pages 670–677.

Bibliography II

Kapoor, A. and Winn, J. (2006). Located Hidden Random Fields: Learning Discriminative Parts for Object Detection. In *European Conference on Computer Vision 2006*, pages 302–315.

Kumar, P., Torr, P., and Zisserman, A. (2005). OBJ CUT. In *IEEE Conference on Computer Vision and Pattern Recognition 2005*, pages 18–25.

Larlus, D., Verbeek, J., and Jurie, F. (2009). Category level object segmentation by combining bag-of-words models with Dirichlet processes and random fields. *International Journal of Computer Vision*, 88:238–253.

Leibe, B., Leonardis, A., and Schiele, B. (2004). Combined Object Categorization and Segmentation With An Implicit Shape Model. In *ECCV Workshop on Statistical Learning in Computer Vision*.

Bibliography III

Levin, A. and Weiss, Y. (2009). Learning to Combine Bottom-Up and Top-Down Segmentation. *International Journal of Computer Vision*, 81:105–118.

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

Rother, C., Kolmogorov, V., and Blake, A. (2004). "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH)*, 23:309–314.

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

Bibliography IV

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

Winn, J. and Shotton, J. (2006). The Layout Consistent Random Field for Recognizing and Segmenting Partially Occluded Objects. In *IEEE Conference on Computer Vision and Pattern Recognition 2006*, pages 37–44.

Multiple Transformed Masks and Appearances

The model

Observed variables

Dataset $\mathbf{D} = \{\mathbf{X}^i\}, i = 1...n$ of images $\mathbf{X}$, each consisting of D pixels $\mathbf{x}_d$, each in a C-dimensional feature space: $\mathbf{x}_d = (x_{dc}), \mathbf{x}_{dc} \in [0, 1]$.

Query variables

For $\mathbf{X}^i$, a segmentation $\mathbf{S}^i$ consisting of D labelings $\mathbf{s}_d$. $\mathbf{s}_d$ is a 1-of-$(L+1)$ encoded variable, where L is the fixed number of 'parts' that combine to generate the images: $\mathbf{s}_d = (s_{\ell d}), s_{\ell d} \in \{0, 1\}, \sum_\ell s_{\ell d} = 1$.

Output

Pixel $\mathbf{x}_d$ background if $s_{0d} = 1$, foreground otherwise.

Multiple Transformed Masks and Appearances

The model

Parameters

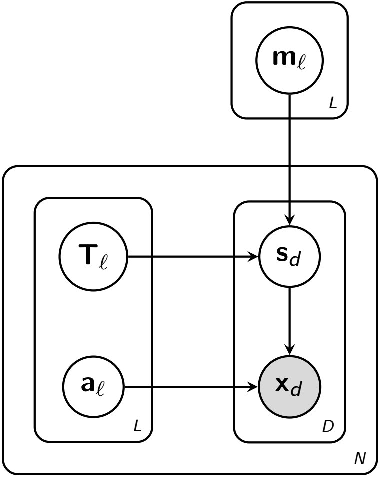
Mask variables $\mathbf{m}_\ell$. Each is a collection of positive real numbers, densely representing the model's preference for part ℓ's shape. Background layer's mask constrained to a vector of ones, i.e. $\mathbf{m}_0 = \mathbf{1}$.

Latent variables

- Transformation variables $\mathbf{T}_\ell$. Each is a permutation matrix, here constrained to 2D translations.
- Appearance variables $\mathbf{a}_\ell$. Can be thought of as low-dimensional latent representations of the parts' appearances.

Multiple Transformed Masks and Appearances

The graphical model



Multiple Transformed Masks and Appearances

Summary of the model

$$\mathbf{Z}^i = \{\mathbf{A}^i, \mathbf{S}^i, \mathbf{T}^i\}$$
$$\boldsymbol{\theta} = \{\mathbf{M}\}$$

$$p(\mathbf{X}^1, ..., \mathbf{X}^n, \mathbf{Z}^1, ..., \mathbf{Z}^n | \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{X}^i, \mathbf{Z}^i | \boldsymbol{\theta})$$

$$\begin{aligned} p(\mathbf{X}, \mathbf{A}, \mathbf{S}, \mathbf{T} | \mathbf{M}) &= p(\mathbf{A})\, p(\mathbf{T}) p(\mathbf{X} | \mathbf{A}, \mathbf{S})\, p(\mathbf{S} | \mathbf{T}, \mathbf{M}) \\ &= p(\mathbf{A})\, p(\mathbf{T}) \prod_{d=1}^{D} p(\mathbf{x}_d | \mathbf{A}, \mathbf{s}_d)\, p(\mathbf{s}_d | \mathbf{T}, \mathbf{M}) \end{aligned}$$

Multiple Transformed Masks and Appearances

Learning

Goal

Approximate $p(\mathbf{T}|\mathbf{X}, \theta)$ and draw $K_{\mathbf{T}|\mathbf{X}}$ samples of $\mathbf{T}$.

Problem

- Discretise each layer's transformation space into J values.
- Inference involves a total of $O(J^L)$ computations.

Solutions

- Variational techniques (Frey et al., 2003).
- Greedy approach (Williams and Titsias, 2004).

Multiple Transformed Masks and Appearances

Learning

Goal

Wish to find $\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$.

Approximate

- Compute $\frac{\partial Q}{\partial m_{\ell d}}$ (involved but can be done efficiently).
- Use Scaled Conjugate Gradients optimisation to maximise Q.
- Results in a **Generalised EM** algorithm.