Factored Shapes and Appearances for Parts-based Object Understanding

S. M. Ali Eslami
s.m.eslami@sms.ed.ac.uk
Christopher K. I. Williams
ckiw@inf.ed.ac.uk

School of Informatics,
University of Edinburgh,
Edinburgh,
United Kingdom

One of the long-standing open problems in machine vision has been the task of foreground-background segmentation. There is broad agreement that this task is coupled to that of object recognition. In this paper we focus on one side of this relationship; given the ground truth value of the object's identity in an image region specified by a bounding box, how accurately can we segment that image?

When the object's pixel intensities are near constant in the dataset (*e.g.* in videos), statistics of its appearance have been used to guide segmentation [2]. However for many datasets of interest the object's appearance is too variable to be modelled effectively by these methods. Recently, a number of models have been proposed that carry out segmentation by incorporating prior knowledge about the object's shape instead [1, 3, 4, 5]. In such cases, techniques mainly differ in how accurately they represent and learn about the variability in the object's shape.

In this paper we present a novel image representation that is *parts-based* and learns from datasets that exhibit variability in *both shape and appearance*. The model's latent representations can be interpreted as 'parsings' of images.

In the Factored Shapes and Appearances (FSA) model we consider datasets of images of an object class. We assume that the images are constructed through some combination of a fixed number of parts. Given a dataset of such images $\mathbf{X}$, we wish to infer a segmentation $\mathbf{S}$ for each image. Each segmentation consists of a labelling $\mathbf{s}_d$ for pixels, where L is the fixed number of parts that combine to generate the foreground and $\mathbf{s}_d$ is a 1-of-$(L+1)$ encoded variable. Accurate inference of $\mathbf{S}$ is driven by FSA's models for 1) part shapes and 2) part appearances (see Fig. 1).

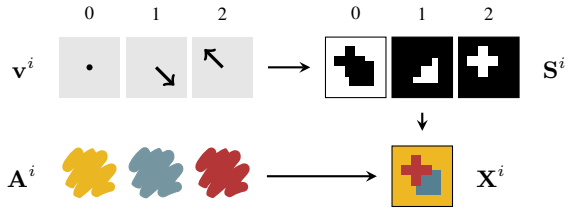

Figure 1: Schematic of the FSA model for a single image $\mathbf{X}^i$ (L=3). Pixel intensities $\mathbf{X}$ are modelled via appearance random variables ($\mathbf{A}$). The model's belief about each part's shape is captured by a latent variable ($\mathbf{v}$). Segmentation variables ($\mathbf{S}$) assign each image pixel to a part.

Let $\mathbf{m}_l$ be a collection of real numbers of the same size as the image, densely representing the model's preference for part l's shape at each location. These 'masks' combine via a softmax activation function to generate the segmentation $\mathbf{S}$: $p(\mathbf{s}_{ld} = 1) = \exp\{\mathbf{m}_{ld}\} / \sum_{k=0}^{L} \exp\{\mathbf{m}_{kd}\}$. In order to be able to allow for part shape variability, the model is designed to capture a *distribution* over $\mathbf{m}_l, l = 1...L$ (the background's mask $\mathbf{m}_0$ is fixed to equal $\mathbf{1}$). Specifically, the probability distribution over $\mathbf{m}_l$ is defined by a Factor Analysis-like model:

$$\mathbf{m}_l = \mathbf{F}_l \mathbf{v} + \mathbf{c}_l, \qquad p(\mathbf{v}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_{H \times H}).$$

Part appearances are modelled as mixtures of Gaussians in colour space. Our formulation captures pixel statistics of parts within a single image, as well statistics of appearance across images in the dataset. A more detailed description of the shape and appearance models can be found in the paper.

Inference in the model is performed by iteratively sampling $\mathbf{S}$, $\mathbf{v}$ and appearance variables $\mathbf{A}$ (samples of $\mathbf{v}$ are obtained efficiently using an elliptical slice sampler). We use the EM algorithm to find estimates of the maximum likelihood parameters.

One illustrative dataset we experiment on consists of 20 images of cars. In addition to appearance variability, the cars exhibit significant shape variability across the dataset. The segmentations inferred by an unsupervised FSA model on this dataset can be seen in Fig. 2(a).

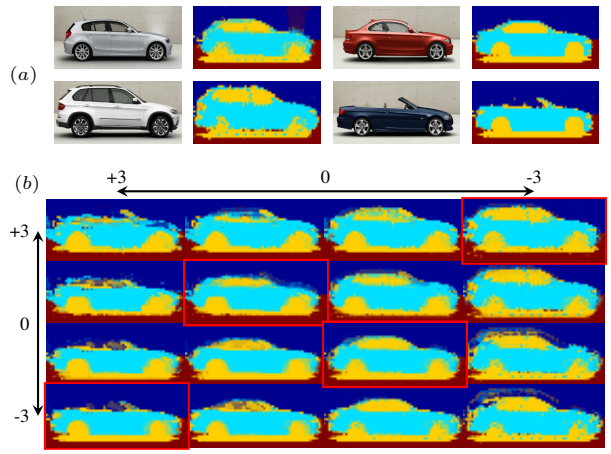

Figure 2: (a) A subset of the training images with their inferred segmentations. Distinct colours indicate assignments of pixels to different parts. (b) A plot of the joint segmentation for a grid of $\mathbf{v}$ values in 2D latent space. Prototypical shapes of 4 different car *types* shown in red.

We can inspect how the latent $\mathbf{v}$ variable is projected by $\mathbf{F}_l$ and $\mathbf{c}_l$ into masks for the parts. In Fig. 2(b) we plot the car body's mask for a grid of $\mathbf{v}$ values in 2-dimensional latent space. Notice how FSA learns a model of shape that gradually morphs between the parts' possible outlines. In doing so it learns a model of object class shape that is more informative than just a mean. We also observe that the inferred $\mathbf{v}$s can be used as discriminative indicators of the object's *type*. In our experiments, using a leave-one-out SVM classifier on *only* the inferred $\mathbf{v}$s, we can classify the cars into the 5 distinct categories with 100% accuracy.

We additionally evaluate FSA's performance at segmentation (see Table 1). Even though the FSA model does not use CRF-style pixelwise dependency terms, its performance is comparable to that of state-of-the-art methods on common benchmarks.

Table 1: Accuracy defined as the average percentage of correctly labelled pixels. Supervised FSA is trained *with* ground-truth masks and $L = 1$.

	Weizmann	Caltech4			
	Horses	Cars	Faces	Bikes	Planes
GrabCut [1]	83.9%	45.1%	83.7%	82.4%	84.5%
Borenstein *et al.* [3]	93.6%	-	-	-	-
LOCUS [5]	93.1%	91.4%	-	-	-
Arora *et al.* [4]	-	95.1%	92.4%	83.1%	93.1%
ClassCut [1]	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%

[1] Alexe B., Deselaers T., and Ferrari V. ClassCut for unsupervised class segmentation. In *Proc. ECCV*, pages 380–393, 2010.

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

[3] Borenstein E., Sharon E., and Ullman S. Combining Top-Down and Bottom-Up Segmentation. In *CVPR*, 2004.

[4] Arora H., Loeff N., Forsyth D., and Ahuja N. Unsupervised Segmentation of Objects using Efficient Learning. *CVPR*, pages 1–7, 2007.

[5] Winn J. and Jojic N. LOCUS: Learning object classes with unsupervised segmentation. In *ICCV*, pages 756–763, 2005.