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

Abstract

We present a novel generative framework for learning parts-based representations of object classes. Our model, *Factored Shapes and Appearances* (FSA), employs a highly factored representation to reason about appearance *and* shape variability across datasets of images. We propose Markov Chain Monte Carlo sampling schemes for efficient inference and learning, and evaluate the model on a number of datasets. Here we consider datasets that exhibit large amounts of variability, both in the shapes of objects in the scene, and in their appearances. We show that the FSA model extracts meaningful parts from training data, and that its parameters and representation can be used to perform a range of tasks, including object parsing, segmentation and fine-grained categorisation.

1 Introduction

One of the long-standing open problems in machine vision has been the task of foreground-background segmentation, in which an image is partitioned into two sets of pixels: those that belong to the object of interest in the foreground, and those that do not. There is broad agreement that this task is coupled to that of object recognition. Knowledge of the object's class can lead to more accurate segmentations, and in turn accurate segmentations can be used to obtain higher recognition rates. 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?

There is a rich history of work on probabilistic models that segment by only considering low-level, pairwise pixel statistics *e.g.* [5, 27]. To see why this type of approach is not sufficient on its own, one only has to examine the kinds of images that these models find difficult to segment. Errors can typically be attributed to a lack of high-level, cross-image understanding about the object in question. 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 [7, 8, 12, 28]. However for many datasets of interest the foreground object's appearance is too variable to be modelled effectively by these methods. Recently, a number of models have been proposed that obtain more accurate segmentations by incorporating prior knowledge about the foreground object's shape instead [4, 14, 17, 18, 20, 21, 29]. In such cases, probabilistic techniques mainly differ in how accurately they represent and learn about the variability in the object's shape.

Though advances have recently been made for handwritten digits [15] and human silhouettes [16], the task of learning accurate distributions of holistic shape in general still remains a challenging and open problem. One natural approach to this problem is the parts-based one in which the variability of the object's shape is reasoned about in terms of the relationships of its constituent parts [3]. The main idea behind such approaches is that the parts combine *factorially* in some way to generate the object's shape [4, 11, 21, 23]. The challenge predominantly lies in two areas: 1) how the object's parts are extracted from data and 2) how they combine to generate the whole. Additionally, much of the existing research in this area is focused on learning parts-based representations of datasets that exhibit limited variability in either shape [17, 26] or appearance [19, 21], or are applicable only to small patches [22].

The main contributions of this paper are as follows: 1) We present a novel image representation that is *parts-based*, and learns from datasets that exhibit variability in *both shape and appearance*. Our experiments on a variety of datasets demonstrate the advantages of FSA's explicit modelling of part deformations over related methods. 2) We demonstrate that the model's latent representations can be interpreted as 'parsings' of images, and show that these parsings are accurate enough to be used for tasks like fine-grained categorisation. 3) We apply FSA to the foreground-background segmentation task, and find that even without CRF-style pixelwise dependency terms its performance is comparable to that of the state-of-the-art on a number of benchmark datasets.

In Secs. 2 and 3 we present FSA and propose an efficient inference and learning scheme for the model. In Sec. 4 we explain how FSA generalises and extends previous work in the field. We provide an experimental evaluation of the model in Sec. 5 and conclude with a discussion in Sec. 6.

2 The FSA generative model

In FSA 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 (which can alternatively be thought of as layers). Given a dataset $\mathbf{D} = \{\mathbf{X}^i\}, i = 1...n$ of such images $\mathbf{X}$, each consisting of D pixels $\{\mathbf{x}_d\}$ in some feature space, we wish to infer a segmentation $\mathbf{S}$ for the image. Each segmentation consists of a labelling $\mathbf{s}_d$ for every pixel, 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. In other words, $\mathbf{s}_d = (s_{ld})$, $l = 0...L$, $s_{ld} \in \{0,1\}$ and $\sum_l s_{ld} = 1$. Note that the background is also treated as a 'part' ($l = 0$). Accurate inference of $\mathbf{S}$ is driven by FSA's models for 1) part shapes and 2) part appearances. In the following sections we describe how the two components are defined.

Shape: 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-like activation function to generate the segmentation $\mathbf{S}$. Let

$$\sigma_{ld} = \frac{\exp\{\mathbf{m}_{ld}\}}{\sum_{k=0}^{L} \exp\{\mathbf{m}_{kd}\}}, \tag{1}$$

then the distribution on the labelling of pixel d, $p(\mathbf{s}_{ld} = 1|\boldsymbol{\theta})$, is given by $\epsilon + (1 - L\epsilon - \epsilon) \cdot \sigma_{ld}$. Here ϵ is a parameter that helps prevent over-confident predictions by 'smoothing out' the distribution imposed by the model on segmentations.

In order to be able to allow for part shape variability, the model is designed to capture a *distribution* over $\mathbf{m}_l, l = 1...L$ ($\mathbf{m}_0$ is fixed to equal $\mathbf{1}$). Specifically, the probability distribution over $\mathbf{m}_l$ is defined by a Factor Analysis-like model:

$$\mathbf{m}_l = \mathbf{F}_l\mathbf{v} + \mathbf{c}_l, \qquad p(\mathbf{v}) = \mathcal{N}(\mathbf{0}, \mathbf{I}_{H\times H}). \tag{2}$$

Here $\mathbf{v}$ is an H-dimensional latent variable, $\mathbf{F}_l$ is a $D \times H$ matrix analogous to the *factor loading matrix* in Factor Analysis literature and $\mathbf{c}_l$ is the mean mask. An L^1-norm prior on $\mathbf{F}$ is used to reduce the amount of noise in its values.

We additionally consider an alternative shape variability model in which we use *separate*, $\bar{H}$ dimensional latent variables $\mathbf{v}_l$ for every part ($H = L \times \bar{H}$). This **local** model can be thought of as a special case of the **global** model presented earlier, in which most of the columns of each $\mathbf{F}_l$ are forced to equal 0. The local model is useful in cases where we believe the shapes of any two pairs of parts in the data to be independent (*e.g.* the pose of upper and lower parts of human bodies), and we wish to explicitly build this knowledge into the model.

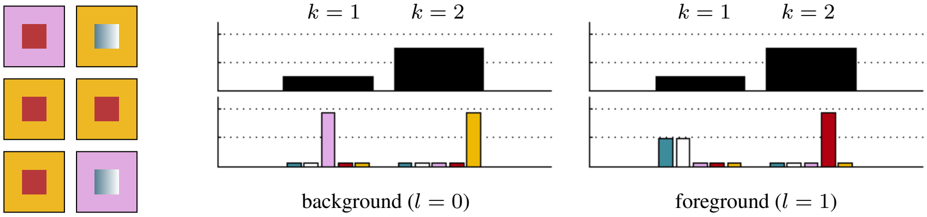

Figure 1: **Appearance modelling:** Given a dataset of images and their segmentations, we construct a model of the parts' appearances. **Left:** The dataset. The foreground and background appear with 2 different styles. **Right:** The corresponding appearance model. The top row depicts $\boldsymbol{\pi}_l$ for the two parts and the bottom row depicts $\boldsymbol{\phi}_l$. In this example, $L = 1$, $K = 2$ and $W = 5$.

Appearance: Pixels corresponding to each part in a given image are assumed to have been generated by W fixed Gaussians in feature space (in this paper we only use Lab colour features). In the pre-training phase, the means $\{\boldsymbol{\mu}_w\}$ and covariances $\{\boldsymbol{\Sigma}_w\}$ of these Gaussians are extracted by training a Gaussian mixture model with W components on every pixel in the dataset, ignoring image and part structure. It is also assumed that each of the L parts have different appearances in different images, and that these appearances can be clustered into K classes. The classes differ in how likely they are to use each of the W Gaussian components when 'colouring in' the part.

The generative process is as follows. For part l in a given image, one of the K classes is chosen (represented by a 1-of-K indicator variable $\mathbf{a}_l$). Given $\mathbf{a}_l$, the probability distribution defined on pixels associated with part l is given by a Gaussian mixture model with means $\{\boldsymbol{\mu}_w\}$ and covariances $\{\boldsymbol{\Sigma}_w\}$ and mixing proportions $\{\phi_{lkw}\}$. Therefore the distribution on the image pixel values is given by

$$p(\mathbf{x}_d|\mathbf{A}, \mathbf{s}_d, \boldsymbol{\theta}) = \prod_{l=0}^{L} p(\mathbf{x}_d|\mathbf{a}_l, \boldsymbol{\theta})^{s_{ld}} = \prod_{l=0}^{L} \left(\prod_{k=1}^{K} \left(\sum_{w=1}^{W} \phi_{lkw} \mathcal{N}(\mathbf{x}_d|\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w) \right)^{a_{lk}} \right)^{s_{ld}} \tag{3}$$

The prior on $\mathbf{A} = \{\mathbf{a}_l\}$ specifies the probability of each appearance class being selected for the parts in any given image:

$$p(\mathbf{A}|\boldsymbol{\theta}) = \prod_{l=0}^{L} p(\mathbf{a}_l|\boldsymbol{\theta}) = \prod_{l=0}^{L}\prod_{k=1}^{K}(\pi_{lk})^{a_{lk}}. \tag{4}$$

See Fig. 1 for an illustration of the appearance model. In our experiments the model typically performs best when $K \simeq 10$, and $W \simeq 30$. We additionally place a hyper-prior on $\boldsymbol{\phi}$ of the form $p(\boldsymbol{\phi}) \propto \exp\{-E(\boldsymbol{\phi})\}$ where

$$E(\boldsymbol{\phi}) = \sum_{l=0}^{L}\left(\lambda_{\text{self}} \cdot \sum_{k=1}^{K} H(\boldsymbol{\phi}_{lk}) - \lambda_{\text{others}} \cdot \sum_{m \neq l} D_{\text{KL}}(\boldsymbol{\phi}_l \parallel \boldsymbol{\phi}_m) + D_{\text{KL}}(\boldsymbol{\phi}_m \parallel \boldsymbol{\phi}_l)\right). \tag{5}$$

Here $H(\boldsymbol{\phi}_{lk})$ is the entropy of the distribution defined by $\boldsymbol{\phi}_{lk}$ (similar to that used in [6]) and $D_{\text{KL}}(\boldsymbol{\phi}_l \parallel \boldsymbol{\phi}_m)$ is the Kullback-Leibler divergence from the distribution defined by $\boldsymbol{\phi}_l$ to that defined by $\boldsymbol{\phi}_m$. This hyper-prior encourages settings of $\boldsymbol{\phi}$ that define distributions on the appearance components that are 1) low-entropy, and 2) dissimilar from each other, and can be very effective in accelerating convergence of the parameters during training. Suitable values of λ_{self} and λ_{others} are found through trial and error.

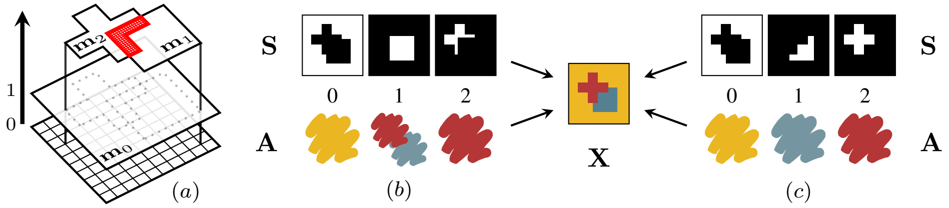

Figure 2: **Lazy occlusion reasoning:** (a) Given the image $\mathbf{X}$, the masks are deformed to their most likely states. In this example, the model has learned that the cross and square always appear in front of the background and that they are equally likely to be in the foreground. The highlighted pixels (red) are equally likely to belong to either shape at this stage. (b) One setting of $\mathbf{A}$ and $\mathbf{S}$ that can explain $\mathbf{X}$. Note the two-tone appearance for part 1. (c) The most likely setting of $\mathbf{A}$ and $\mathbf{S}$. Out of all such competing segmentations, the most likely $\mathbf{S}$ is the one for which the corresponding choice of appearances is most probable.

Handling occlusion: Instead of modelling part occlusion using an explicit random variable, FSA captures knowledge about part-ordering *implicitly* in the shape parameters. By increasing the magnitude of m_{ld} for a particular l, the model can capture the increased likelihood of part l occluding other parts at pixel d. In cases where the multiple parts are equally likely to occlude each other, the appearance model is used to resolve this ambiguity in the posterior. See Fig. 2 for an illustration of this effect.

Combined model: To summarise, the latent variables $\mathbf{Z}^i$ for image $\mathbf{X}^i$ are $\mathbf{A}^i$, $\mathbf{S}^i$ and $\mathbf{v}^i$, the model's active parameters $\boldsymbol{\theta}$ include shape parameters $\boldsymbol{\theta}^s = \{\{\mathbf{F}_l\}, \{\mathbf{c}_l\}\}$ and appearance parameters $\boldsymbol{\theta}^a = \{\{\pi_{lk}\}, \{\phi_{lkw}\}\}$, and

$$p(\mathbf{X}^i, \mathbf{A}^i, \mathbf{S}^i, \mathbf{v}^i|\boldsymbol{\theta}) = p(\mathbf{v}^i)\, p(\mathbf{A}^i|\boldsymbol{\theta}^a) \prod_{d=1}^{D} p(\mathbf{s}_d|\mathbf{v}^i, \boldsymbol{\theta}^s)\, p(\mathbf{x}_d^i|\mathbf{A}, \mathbf{s}_d^i, \boldsymbol{\theta}^a). \tag{6}$$

See Fig. 3 for an illustration of the complete FSA graphical model. During learning, we find the values of $\boldsymbol{\theta}$ that maximise the likelihood of the training data $\mathbf{D}$, and segmentation is performed on previously-unseen image by querying the marginal distribution $p(\mathbf{S}|\mathbf{X}^{\text{test}},\boldsymbol{\theta})$.

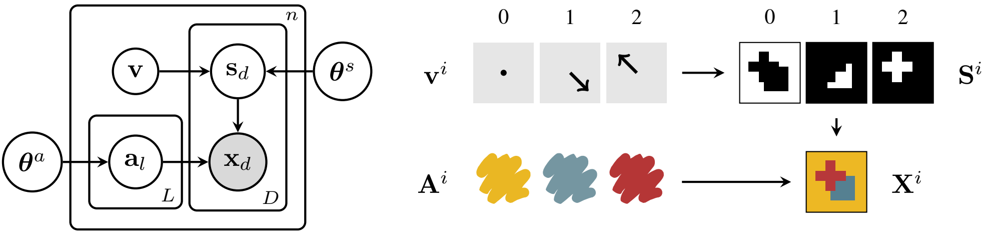

Figure 3: **Left:** Directed graphical representation of the global FSA model. Pixel intensities $\mathbf{x}_d$ are modelled via L appearance random variables ($\mathbf{a}_l$). The model's belief about each part's shape is captured by a latent variable ($\mathbf{v}$). Segmentation random variables ($\mathbf{s}_d$) assign each image pixel to a part. **Right:** Schematic diagram of the model for a single image $\mathbf{X}^i$.

3 Inference and learning

We use the expectation-maximisation (EM) algorithm to find estimates of the maximum likelihood parameters. For the E-step, we wish to find $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}) = p(\mathbf{A},\mathbf{S},\mathbf{v}|\mathbf{X},\boldsymbol{\theta})$. However, the exact evaluation of this distribution is intractable. Instead we approximate $p(\mathbf{A},\mathbf{S},\mathbf{v}|\mathbf{X},\boldsymbol{\theta})$ by drawing samples of $\mathbf{A}$, $\mathbf{S}$ and $\mathbf{v}$ using block-Gibbs Markov Chain Monte Carlo (MCMC).

The appearance variable $\mathbf{A}^i$ is sampled given each image $\mathbf{X}^i$ and its corresponding segmentation $\mathbf{S}^i$. The conditional distribution of appearance class k being chosen for part l (*i.e.* the binary variable a_{lk} being set to 1) is given by:

$$p(a_{lk}=1|\mathbf{S},\mathbf{v},\mathbf{X},\boldsymbol{\theta}) = \frac{\pi_{lk}\prod_{d=1}^{D}\left(\sum_{w=1}^{W}\phi_{lkw}\,\mathcal{N}(\mathbf{x}_d|\boldsymbol{\mu}_w,\boldsymbol{\Sigma}_w)\right)^{s_{ld}}}{\sum_{k=1}^{K}\left[\pi_{lk}\prod_{d=1}^{D}\left(\sum_{w=1}^{W}\phi_{lkw}\,\mathcal{N}(\mathbf{x}_d|\boldsymbol{\mu}_w,\boldsymbol{\Sigma}_w)\right)^{s_{ld}}\right]}. \tag{7}$$

The segmentation variable $\mathbf{S}^i$ is then sampled given $\mathbf{v}^i$ and $\mathbf{A}^i$. It can be shown that the conditional distribution of the segmentation factorises over the pixels in the image. The probability of pixel d being associated with part l is:

$$p(s_{ld}=1|\mathbf{A},\mathbf{v},\mathbf{X},\boldsymbol{\theta}) = \frac{p(s_{ld}=1|\mathbf{v},\boldsymbol{\theta})\,p(\mathbf{x}_d|\mathbf{A},\mathbf{s}_d)}{\sum_{m=1}^{L}p(s_{md}=1|\mathbf{v},\boldsymbol{\theta})\,p(\mathbf{x}_d|\mathbf{A},\mathbf{s}_d)}. \tag{8}$$

Finally, $\mathbf{v}^i$ is sampled given the segmentation $\mathbf{S}^i$. To do this we use an efficient elliptical slice sampling scheme [24]. In each iteration of the top-level block-Gibbs sampler, the sample for $\mathbf{v}^i$ is set to equal the mean of the samples returned by the elliptical slice sampler after a burn-in period.

For the M-step we are looking to find $\arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$, where

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \ln p(\boldsymbol{\theta}) + \sum_{i=1}^{n} \sum_{\mathbf{Z}^i} p(\mathbf{Z}^i | \mathbf{X}^i, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}^i, \mathbf{Z}^i | \boldsymbol{\theta}). \tag{9}$$

To do this, we compute the derivative of Q with respect to $\boldsymbol{\theta}^a$ and $\boldsymbol{\theta}^s$. The gradients are used in a numerical optimisation routine to find the settings of the parameters at which Q is maximised. We use independent scaled conjugate gradients (SCG) routines to update the shape and appearance parameters. Note that special care needs to be taken to ensure that the $\boldsymbol{\pi}$ and $\boldsymbol{\phi}$ variables sum to 1. We re-parametrise the model such that $\pi_{lk} = \exp\{\alpha_{lk}\} / \sum_{c=1}^{K} \exp\{\alpha_{lc}\}$ and $\phi_{lkw} = \exp\{\beta_{lkw}\} / \sum_{v=1}^{W} \exp\{\beta_{lkv}\}$, and optimise Q with respect to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ instead.

4 Related work

Existing parts-based image models can be categorised by the amount of variability they expect to encounter in the data and by how they model this variability. For example, in the Layered Subspace Manifold (LSM) of Frey *et al.* [12] *videos* are partitioned into layers that translate independently of each other. The layers exhibit limited shape and appearance variability from frame to frame, and are modelled using Factor Analysers and a fixed, explicit occlusion ordering. With the Sprites model [28], Williams and Titsias show how such layered models can be efficiently learned one layer at a time, however they do not model shape or appearance variability. By contrast, FSA is designed to work on datasets of *images* that exhibit significant shape and appearance variability from image to image, and does not impose any layer ordering into the model.

With Multiple Cause Vector Quantisation (MCVQ) [26], Ross and Zemel present an alternative part-based representation of images. The model learns a fixed partitioning of the image, and it is assumed that a fixed number of appearance templates generate the pixels within each part. When applied to highly variable data, the model finds it difficult to learn meaningful parts as it can only make very limited variation in the partitionings from image to image. The authors also present Multiple Cause Factor Analysis (MCFA), which uses a Factor Analysis model for part appearances, however this remains too restrictive for most datasets of interest. By contrast, FSA explicitly models the variability of pixel assignments to parts, therefore learning sharp partitions, and it models part appearance variation in a more flexible way.

Table 1: Comparison of a number of different parts-based models.

	FACTORED PARTS	**FACTORED SHAPE AND APPEARANCE**	**SHAPE VARIABILITY**	**APPEARANCE VARIABILITY**
LSM [12]	✓(layers)	-	✓(FA)	✓(FA)
Sprites [28]	✓(layers)	-	-	-
LOCUS [29]	-	✓	✓(deformation)	✓(colours)
MCVQ [26]	-	✓	-	✓(templates)
SCA [18]	-	✓	✓(convex)	✓(histograms)
FSA	✓(softmax)	✓	✓(FA)	✓(histograms)

The closest works to ours are LOCUS [29] and Stel Component Analysis (SCA) [18]. In the basic formulation of LOCUS, the model uses only one 'part' to account for the foreground object, but this restriction can be relaxed with the deformable probabilistic index map (dPIM) [29]. Shape variability between images is accounted for using a deformation field that warps the partitioning to fit each image. Since the formulation imposes only local smoothness constraints on deformations, samples from the model in the absense of an image are unlikely to capture global properties of the object in consideration (*e.g.* pose of a horse).

The SCA model, on the other hand, accounts for shape variability by learning a fixed number of templates for each part. The templates are restricted such that any pixelwise, convex combination of templates results in a valid probabilistic index map (*i.e.* one in which the probabilities of part assignments for each pixel sum to 1). The SCA distribution over segmentations is accurate only in the posterior – in the absence of an image, the defined distribution over segmentations is 'blurry'. Thus samples of partitionings generated by LOCUS and SCA will not have much resemblance to their training images, even though the are both generative models of image partitionings.

In FSA part shapes vary accurately even *in the prior* and segmentations randomly sampled by the model are similar to those found in the training data. Additionally, both LOCUS (with dPIMs) and SCA define global distributions over partitionings that do not factorise over part shape. In FSA parts can be modelled independently of each other allowing further developments to be made by incorporating specialised part models that concentrate on the shape, position and scale of each individually. We summarise these differences in Table 1.

5 Experiments

FSA, as a generative model for images of objects, can be used to accomplish a variety of tasks in computer vision. Here we demonstrate its performance on several datasets. FSA segments all images across the dataset simultaneously to learn a parts-based object model. In addition to the segmentations made by the algorithm, we inspect the parameters learned by the model. We show that these parameters form an intuitive reflection of the algorithm's 'understanding' of the object class.

Cars dataset: The first real dataset we consider[1] contains 20 images of cars that have been downloaded from a manufacturer's website[2]. In addition to appearance variability, the cars exhibit significant shape variability across the dataset (*e.g.* hatchback, SUV, convertible coupé, saloon, estate). The segmentations inferred by an unsupervised FSA model with $L = 3$ and $H = 2$ are shown in Fig. 4(a).

It is informative to inspect how the latent $\mathbf{v}$ variable is projected by $\mathbf{F}_l$ and $\mathbf{c}_l$ into masks for the parts. In Fig. 4(b) we plot columns of one of the $\mathbf{F}$ matrices, and in Fig. 4(c) 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[3]. Also note that the model learns a mask for the roof-less 'convertible-coupé' body type. A deformation field like the one used in LOCUS [29] would find this kind of variability difficult to represent. Finally, we observe that the inferred $\mathbf{v}$s can be used as discriminative indicators

[1] See supplementary material for illustrative results on synthetic data.

[2] `http://bmw.com`

[3] See supplementary material for samples from a *supervised* FSA model on the same dataset.

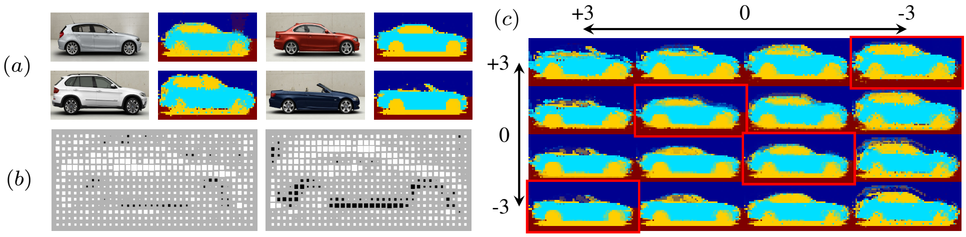

Figure 4: (a) A subset of the training images with their inferred segmentations. Distinct colours indicate assignments of pixels to different parts. (b) Hinton diagrams of the two columns of $\mathbf{F}_2$ corresponding to the car body (cyan). (c) A plot of the joint segmentation for a grid of $\mathbf{v}$ values in 2D latent space. Prototypical shapes of 4 of the different car types have been highlighted in red.

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.

Other datasets: We apply the FSA model to a number of other datasets including 100 MIT pedestrians [25], 200 UMIST faces [13] and 127 Caltech motorbikes [9], as well as 138 images of dresses obtained from a fashion retailer's website[4,5]. The results of these experiments can be seen in Fig. 5. The model does a good job of learning about class shape across the dataset. In our experiments we observed that it uses this information effectively to guide inferences for more difficult images that cannot be segmented based on appearance cues alone. The fact that it has the flexibility to learn about shape deformations increases its chances of transferring shape information in a useful way. For example, having correctly learned about the shape of a human in an unusual pose in an image with strong appearance cues, the model uses this information to correctly segment more difficult images of humans with the same pose. The mean pose in this case would do more harm than good in providing cues for segmentation.

Segmentation accuracy: We additionally evaluate the performance of the FSA model at segmenting the Weizmann horse [4] and Caltech4 [10] datasets, where the ground truths are readily available. The train-test split for the datasets were as follows. Weizmann horses: 127-200; Caltech cars: 63-60; faces: 335-100; motorbikes 698-100 and airplanes: 700-100.

The baseline we consider is the batch GrabCut algorithm described by Alexe *et al.* [1]. GrabCut is initialised by training a foreground colour model on the central 25% of each test image and a background colour model using the remainder of its pixels. In supervised FSA, training is performed given the ground-truth segmentations for each image ($L = 1$).

The results of these experiments can be seen in Table 2. For comparison we also include accuracies reported by Borenstein *et al.* [4] (supervised), Winn and Jojic [29] (unsupervised, colour model) and Alexe *et al.* [1] (unsupervised). FSA uses knowledge about shape to increase the average accuracy over the baseline. The discrepancy with LOCUS and Borenstein *et al.*'s approach on the Weizmann dataset is likely due to the lack of low-level edge features in our implementation of FSA. Supervised FSA outperforms the other models on the face and motorbike datasets, in part due to the way in which it learns to classify pixels belonging

[4] http://marksandspencer.com

[5] Pedestrians: $L = 3$, $H = 2$; faces: $L = 2$, $H = 2$; motorbikes: $L = 3$, $H = 20$; dresses: $L = 1$, $H = 5$.

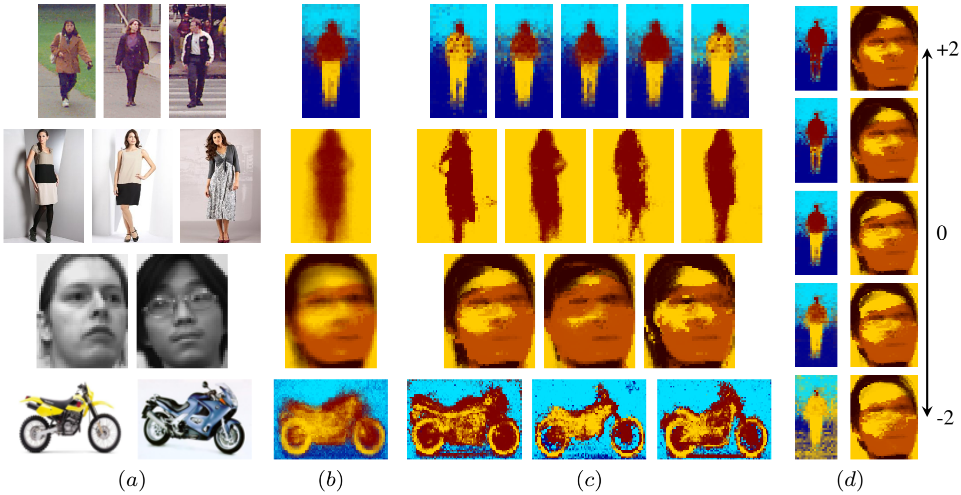

Figure 5: (a) Training images. (b) Partitioning learned by an FSA model with no shape deformation component (equivalent to a PIM). Distinct colours indicate probabilities of assignments of pixels to different parts. (c) A selection of samples from complete FSA models. Notice in row 1) captured variability of clothing styles and leg separation, 2) body poses, 3) face highlights and hair styles, and 4) motorcycle types. (d) Samples from the FSA pedestrian and face models as $\mathbf{v}$ moves on a 1D line in latent space. Notice how, for example, $\mathbf{v}$ affects the size of the forehead and the length of the hair.

to necks as background and motorbike spokes as foreground. Even though the FSA model does not have CRF-style pixelwise dependency terms, its performance is comparable to that of state-of-the-art methods for these datasets.

6 Discussion and future work

In this paper we have presented a novel probabilistic model of objects that learns by simultaneously segmenting all images in the training dataset. The model is parts-based and factorial: each of the parts can be modelled independently of the others. The model's descriptors for shape and appearance are particularly well suited to highly variable datasets of images. We have demonstrated that FSA can perform as intended across a range of datasets, and that its latent representation can be used to accomplish a variety of common computer vision tasks.

We are currently investigating the ways in which FSA's latent representation of part shape can be used for the fine-grained visual categorisation task, where the goal is to distinguish between, *e.g.*, species of animals and plants or car and motorcycle types.

We would like to extend the model in a number of ways. In FSA, we can model each part independently, both in terms of shape and appearance. We would like to consider an extension in which additional latent variables explicitly encode for independent rigid transformations (such as scaling, translation and rotation) of the parts. It is also of interest to consider alternative shape models, *e.g.* restricted Boltzmann machines or contour-based models. Additionally, FSA represents the statistics of part appearance in a way that ignores the spatial structure of the pixels within parts. We would like to investigate how more structured models of texture can be used to represent part appearances. Finally, wish to find out if efficient algorithms can be developed to automatically determine suitable choices of L and H.

Table 2: Average segmentation accuracies. Here we report the accuracy of the algorithm as the average percentage of correctly labelled pixels across all the test images.

	Weizmann	Caltech4			
	Horses	Cars	Faces	Motorbikes	Airplanes
GrabCut [1]	83.9%	45.1%	83.7%	82.4%	84.5%
Borenstein *et al.* [4]	93.6%	-	-	-	-
LOCUS [29]	93.1%	91.4%	-	-	-
Arora *et al.* [2]	-	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%

Acknowledgments. We are grateful for financial support from the Carnegie Trust and the SORSAS scheme for AE. This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. This publication only reflects the authors' views.

References

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. ClassCut for unsupervised class segmentation. In *Proceedings of the 11th European conference on Computer vision: Part V*, pages 380–393, 2010.

[2] Himanshu Arora, Nicolas Loeff, David Forsyth, and Narendra Ahuja. Unsupervised Segmentation of Objects using Efficient Learning. *IEEE Conference on Computer Vision and Pattern Recognition 2007*, pages 1–7, 2007.

[3] Irving Biederman. Recognition-by-components: A theory of human image understanding. *Psychological Review*, 94:115–147, 1987.

[4] Eran Borenstein, Eitan Sharon, and Shimon Ullman. Combining Top-Down and Bottom-Up Segmentation. In *CVPR Workshop on Perceptual Organization in Computer Vision*, 2004.

[5] Yuri Boykov and Marie-Pierre Jolly. Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D images. In *International Conference on Computer Vision 2001*, pages 105–112, 2001.

[6] Matthew Brand. An entropic estimator for structure discovery. In *Advances in Neural Information Processing Systems 11*, pages 723–729, 1999.

[7] Taylan Cemgil, Wojciech Zajdel, and Ben Krose. A Hybrid Graphical Model for Robust Feature Extraction from Video. In *IEEE Conference on Computer Vision and Pattern Recognition 2005*, pages 1158–1165, 2005.

[8] Timothy Cootes, Gareth Edwards, and Christopher Taylor. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:681–685, 2001.

[9] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition 2004, Workshop on Generative-Model Based Vision*, 2004.

[10] Rob Fergus, Pietro Perona, and Andrew Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition 2003*, pages 264–271, 2003.

[11] Martin Fischler and Robert Elschlager. The Representation and Matching of Pictorial Structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973.

[12] Brendan Frey, Nebojsa Jojic, and Anitha Kannan. Learning appearance and transparency manifolds of occluded objects in layers. In *IEEE Conference on Computer Vision and Pattern Recognition 2003*, pages 45–52, 2003.

[13] Daniel Graham and Nigel Allinson. *Face Recognition: From Theory to Application*, volume 163, chapter Characterizing Virtual Eigensignatures for General Purpose Face Recognition, pages 446–456. 1998.

[14] Xuming He, Richard S. Zemel, and Debajyoti Ray. Learning and incorporating top-down cues in image segmentation. In *European Conference on Computer Vision 2006*, volume 1, pages 338–351, 2006.

[15] Geoffrey Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, July 2006.

[16] Manuel Marin Jimenez, Nicolas Perez de la Blanca, and M. Angeles Mendoza. RBM-based Silhouette Encoding for Human Action Modelling. In *International Conference on Pattern Recognition 2010*, pages 979–982, 2010.

[17] Nebojsa Jojic and Yaron Caspi. Capturing Image Structure with Probabilistic Index Maps. In *IEEE Conference on Computer Vision and Pattern Recognition 2004*, pages 212–219, 2004.

[18] Nebojsa Jojic, Alessandro Perina, Marco Cristani, Vittorio Murino, and Brendan Frey. Stel component analysis: Modeling spatial correlations in image class structure. In *IEEE Conference on Computer Vision and Pattern Recognition 2009*, pages 2044–2051, 2009.

[19] Anitha Kannan, John Winn, and Carsten Rother. Clustering Appearance and Shape by Learning Jigsaws. In *Advances in Neural Information Processing Systems 19*, pages 657–664, 2006.

[20] Ashish Kapoor and John Winn. Located Hidden Random Fields: Learning Discriminative Parts for Object Detection. In *European Conference on Computer Vision 2006*, pages 302–315, 2006.

[21] Pawan Kumar, Philip Torr, and Andrew Zisserman. OBJ CUT. In *IEEE Conference on Computer Vision and Pattern Recognition 2005*, pages 18–25, 2005.

[22] Nicolas Le Roux, Nicolas Heess, Jamie Shotton, and John Winn. Learning a generative model of images by factoring appearance and shape. Technical Report MSR-TR-2010-7, Microsoft Research Cambridge, 2010.

[23] Bastian Leibe, Ales Leonardis, and Bernt Schiele. Combined Object Categorization and Segmentation With An Implicit Shape Model. In *ECCV Workshop on Statistical Learning in Computer Vision*, 2004.

[24] Iain Murray, Ryan Prescott Adams, and David J.C. MacKay. Elliptical slice sampling. *Journal of Machine Learning Research*, 9:541–548, 2010.

[25] Michael Oren, Constantine Papageorgiou, Pawan Sinha, Edgar Osuna, and Tomaso Poggio. Pedestrian Detection Using Wavelet Templates. In *IEEE Conference on Computer Vision and Pattern Recognition 1997*, pages 193–99, 1997.

[26] David Ross and Richard Zemel. Learning Parts-Based Representations of Data. *Journal of Machine Learning Research*, 7:2369–2397, 2006.

[27] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. “GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH) 2004*, 23:309–314, 2004.

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

[29] John Winn and Nebojsa Jojic. LOCUS: Learning object classes with unsupervised segmentation. In *International Conference on Computer Vision 2005*, pages 756–763, 2005.