# OBJ CUT

M. Pawan Kumar P.H.S. Torr
Dept. of Computing
Oxford Brookes University
{pkmudigonda,philiptorr}@brookes.ac.uk
http://wwwcms.brookes.ac.uk/~{pawan,philiptorr}

A. Zisserman
Dept. of Engineering Science
University of Oxford
az@robots.ox.ac.uk
http://www.robots.ox.ac.uk/~vgg

## Abstract

*In this paper we present a principled Bayesian method for detecting and segmenting instances of a particular object category within an image, providing a coherent methodology for combining top down and bottom up cues. The work draws together two powerful formulations: pictorial structures (*PS*) and Markov random fields (*MRF*s) both of which have efficient algorithms for their solution. The resulting combination, which we call the Object Category Specific* MRF*, suggests a solution to the problem that has long dogged* MRF*s namely that they provide a poor prior for specific shapes. In contrast, our model provides a prior that is global across the image plane using the* PS*. We develop an efficient method,* OBJ CUT*, to obtain segmentations using this model. Novel aspects of this method include an efficient algorithm for sampling the* PS *model, and the observation that the expected log likelihood of the model can be increased by a single graph cut. Results are presented on two object categories, cows and horses. We compare our methods to the state of the art in object category specific image segmentation and demonstrate significant improvements.*

## 1. Introduction

Image Segmentation has seen renewed interest in the field of Computer Vision, in part due the arrival of new efficient algorithms to perform the segmentation [5], and in part due to the resurgence of interest in object recognition [2, 8, 9]. Segmentation fell from favour partly due to an excess of papers attempting to solve ill posed problems with no means of judging the result. Interleaved object recognition and segmentation [4, 16] is both well posed and of practical use. Well posed in that the result of the segmentation can be quantitatively judged e.g. how many pixels have been correctly and incorrectly assigned to the object. Of practical use because (a) the more accurately the image can be segmented the more accurate the recognition results will be, and (b) image editing tools can be designed that provide a "power assist" to cut out applications like 'Magic Wand', e.g. I know this is a horse, please segment it for me, without the pain of having to manually delineate the boundary.

Markov Random Fields (MRFs) provide a useful model of images for segmentation and their prominence has been increased by the availability of efficient publically available code for their solution. The work of Boykov and Jolly [5] strikingly demonstrates that with the minimum of user assistance objects can be rapidly segmented. However samples from the Gibbs distribution defined by the MRF very rarely give rise to realistic shapes and on their own MRFs are ill suited to segmenting objects. What is required is a way to inject prior knowledge of object shape into the MRF. Within this paper we derive a Bayesian way of doing this in which the prior knowledge is provided by a Pictorial Structure (PS). Pictorial Structures [6] and the related Constellation of Parts model [8] have proven to be highly successful for the task of object recognition. We cast the problem of object category specific segmentation as that of estimating an MRF (representing bottom up information) which is influenced by a set of latent variables, the PS (representing top down information), encouraging the MRF to resemble the object. Unlike MRFs, which model the prior using pairwise potentials, the PS model provides a prior over the shape of the segmentation that is global across the image plane.

In contrast to previous approaches [4, 16], our method provides a principled probabilistic approach which can deal with large object deformations and articulations without having to resort to the computational inefficiency of hundreds of exemplars. The basis of our method are two new theoretical/algorithmic contributions: (1) we make the (not obvious) observation that the expectation of the log likelihood of an MRF with respect to some latent variables can be efficiently optimized with respect to the labels of the MRF by a single graph cut optimization; (2) we provide a highly efficient algorithm for marginalizing or optimizing the latent variables when they are a PS following a Potts model.

The paper is organized as follows. In Section 2 the probabilistic MRF model of the image is described in broad terms. Section 3 gives an overview of an efficient method for solving this probabilistic MRF model for figure-ground labellings. In section 4 the layered pictorial structures (LPS) model is

described, which extends the PS model so that it handles partial self occlusion. How to get an initial estimate of the LPS is given in section 5. The OBJ CUT algorithm is described in section 6. Results are shown for two object categories, namely cows and horses, and a comparison with other methods is given in section 7.

## 2. Object Category Specific MRF

In this section we describe the MRF model that forms the basis of the paper. We build upon and formally specify previous work on segmentation providing a Bayesian graphical model for work that has previously been specified in terms of energy functions [5]. Notably there are two issues to be addressed in this section (i) how to encourage the segmentation to follow edges within the image, (ii) how to encourage the segmentation to look like an object.

Given an image $\mathbf{D}$ containing an instance of a known object category, e.g. cows, we wish to segment the image into *figure*, i.e. pixels belonging to the object, and *ground*, i.e. the background. Taking a Bayesian perspective, we define a set of binary labels, $\mathbf{m}$, one label $m_x$ for each pixel $x$, that optimizes the posterior probability given by the Gibbs distribution

$$p(\mathbf{m}|\mathbf{D}) = \frac{p(\mathbf{D}|\mathbf{m})p(\mathbf{m})}{p(\mathbf{D})} = \frac{1}{Z_{\mathbf{m}}}\exp(-\Psi(\mathbf{m})), \quad (1)$$

where $Z_{\mathbf{m}}$ is the normalizing constant (or partition function). The energy is defined by the summation of clique potentials:

$$\Psi_1(\mathbf{m}) = \sum_x \left(\phi(\mathbf{D}|m_x) + \sum_y \psi(m_x, m_y)\right), \quad (2)$$

where $y$ is a neighbouring pixel of $x$. The likelihood term $\phi(\mathbf{D}|m_x)$ is the emission model for one or more pixels and is given by

$$\phi(\mathbf{D}|m_x) = \begin{cases} -\log(p(x \in \text{figure}|\mathcal{H}_{obj})) & \text{if} \quad m_x = 1 \\ -\log(p(x \in \text{ground}|\mathcal{H}_{bkg})) & \text{if} \quad m_x = 0, \end{cases} \quad (3)$$

where $\mathcal{H}_{obj}$ and $\mathcal{H}_{bkg}$ are the RGB distributions for foreground and background respectively. The prior $\psi(m_x, m_y)$ takes the form of an Ising model:

$$\psi(m_x, m_y) = \begin{cases} P & \text{if} \quad m_x \neq m_y, \\ 0 & \text{if} \quad m_x = m_y. \end{cases} \quad (4)$$

In the MRFs used for image segmentation, a contrast term is used to favour pixels with similar colour having the same label [3, 5]. This is done by reducing the cost within the Ising model for two labels being different in proportion to the difference in intensities of their corresponding pixels e.g. by $\gamma(x,y) = \lambda \exp\left(\frac{-g^2(x,y)}{2\sigma^2}\right)\frac{1}{dist(x,y)}$, where $g^2(x,y)$ measures the difference in the RGB values of pixels $x$ and $y$ and $dist(x,y)$ gives the spatial distance between $x$ and $y$ [3, 5]. However, this has previously not been given a proper Bayesian formulation which we now address. It can not be included in the prior, for the prior term cannot include the data. Rather it leads to a pair wise linkage between neighbouring labels and their pixels as shown in the graphical model given in figure 1. The energy function of this MRF is of the form

$$\Psi_2(\mathbf{m}) = \sum_x \left(\phi(\mathbf{D}|m_x) + \sum_y \left(\phi(\mathbf{D}|m_x, m_y) + \psi(m_x, m_y)\right)\right) \quad (5)$$

The contrast term of the energy function is given by:

$$\phi(\mathbf{D}|m_x, m_y) = \begin{cases} -\gamma(x,y) & \text{if} \quad m_x \neq m_y, \\ 0 & \text{if} \quad m_x = m_y. \end{cases} \quad (6)$$

MRF-based segmentation techniques which use MINCUT [13] have achieved excellent results [3, 5] with manual initialization. However, due to the lack of a shape model, these methods do not work so well for automatic segmentation. We would like to use the power of the MinCut algorithm for interleaved object recognition and segmentation. In some sense the result of the recognition will replace the user interventions. In order to achieve this we introduce a stronger shape model to the MRF and marry the two together. This shape model will supply a set of latent variables, $\mathbf{\Theta}$, which will favour segmentations of a specific shape, as shown in the graphical model depicted in figure 1. We call this new MRF model the Object Category Specific MRF, which has the following energy function:

$$\begin{aligned} \Psi_3(\mathbf{m}, \mathbf{\Theta}) = \quad & \textstyle\sum_x \quad (\phi(\mathbf{D}|m_x) + \phi(m_x|\mathbf{\Theta}) + \\ & \textstyle\sum_y \quad (\phi(\mathbf{D}|m_x, m_y) + \psi(m_x, m_y))) \end{aligned} \quad (7)$$

with posterior $p(\mathbf{m}, \mathbf{\Theta}|\mathbf{D}) = \frac{1}{Z_3}\exp(-\Psi_3(\mathbf{m}, \mathbf{\Theta}))$. The function $\phi(m_x|\mathbf{\Theta})$ is chosen so that if we were given an estimate of the location and shape of the object, then pixels falling near to that shape would more likely have object label and vice versa. It has the form:

$$\phi(m_x|\mathbf{\Theta}) = -\log p(m_x|\mathbf{\Theta}). \quad (8)$$

In this paper, we choose to define $p(m_x|\mathbf{\Theta})$ as

$$p(m_x = \text{figure}|\mathbf{\Theta}) = \frac{1}{1 + \exp(\mu * d(x, \mathbf{\Theta}))} \quad (9)$$

and $p(m_x = \text{ground}|\mathbf{\Theta}) = 1 - p(m_x = \text{figure}|\mathbf{\Theta})$, where $d(x, \mathbf{\Theta})$ is the distance of a pixel $x$ from the shape defined by $\mathbf{\Theta}$ (being negative if inside the shape). The parameter $\mu$ determines how much the points outside the shape are penalized compared to the points inside the shape. Note energy function $\Psi_3(\mathbf{m}, \mathbf{\Theta})$ can still be minimized via MINCUT [13].

In this paper we combine the Contrast Dependent MRF with the layered pictorial structures (LPS) model [15] (our extension of the pictorial structure, PS model [6]), however we

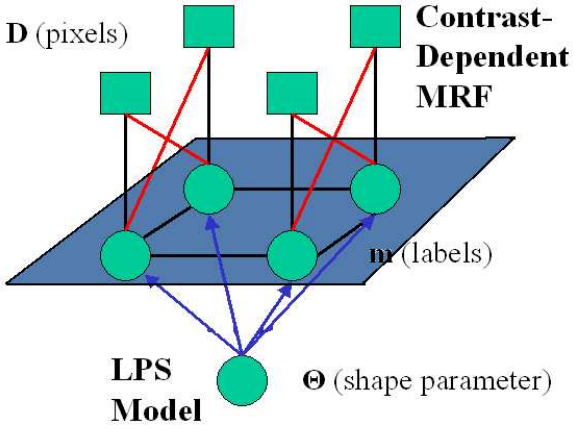


Figure 1. *Graphical model representation of the Object Category Specific* MRF. *The connections introducing the contrast term are shown in red. Note that some of these connections (going diagonally) are not shown for the sake of clarity of the image. The labels* $\mathbf{m}$ *lie in a plane. Together with the pixels shown above this plane, these form the contrast-dependent* MRF *used for segmentation. In addition to these, Object Category Specific* MRF *makes use of an underlying shape parameter in the form of an* LPS *model (shown lying below the plane). The* LPS *model guides the segmentation towards a realistic shape closely resembling the object of interest.*

observe that the methodology below is completely general and could be combined with any sort of latent shape model.

The optimal figure-ground labelling should be obtained by integrating out the latent variable $\mathbf{\Theta}$. The surprising result of this paper is that this rather intractable looking integral can in fact be optimized by a simple and computationally efficient set of operations. In order to do this, we need to demonstrate two things: (i) That given an estimate of $\mathbf{m}$ we can sample efficiently for $\mathbf{\Theta}$. This we shall demonstrate for the case of LPS and in §5.1 we describe a new algorithm for efficient calculation of the marginal distribution for a non regular Potts model (complementing the result of Felzenszwalb and Huttenlocher [7]). (ii) That given the distribution of $\mathbf{\Theta}$ we can efficiently optimize $\mathbf{m}$ so as to increase the posterior. For a MRF this is not immediately obvious, however we shall demonstrate this in the next section.

## 3. Roadmap of the Solution

For the problem of segmentation the parameters $\mathbf{m}$ are of immediate interest and the EM framework provides a natural way to deal with the latent parameters $\mathbf{\Theta}$ [11] by treating them as missing data. The log posterior density of $\mathbf{m}$ is given by

$$\log p(\mathbf{m}|\mathbf{D}) = \log p(\mathbf{\Theta}, \mathbf{m}|\mathbf{D}) - \log p(\mathbf{\Theta}|\mathbf{m}, \mathbf{D}), \quad (10)$$

where $p(\mathbf{\Theta}, \mathbf{m}|\mathbf{D}) = \frac{1}{Z_3}\exp(-\Psi_3(\mathbf{m}, \mathbf{\Theta}))$. The EM framework iteratively refines the estimate of $\mathbf{m}$ by marginalizing the latent parameters $\mathbf{\Theta}$. Given the current guess of the labelling $\mathbf{m}'$, we treat $\mathbf{\Theta}$ as a random variable with the distribution $p(\mathbf{\Theta}|\mathbf{m}', \mathbf{D})$. Averaging over $\mathbf{\Theta}$ yields

$$\log p(\mathbf{m}|\mathbf{D}) = \mathcal{E}(\log p(\mathbf{\Theta}, \mathbf{m}|\mathbf{D})) - \mathcal{E}(\log p(\mathbf{\Theta}|\mathbf{m}, \mathbf{D})), \quad (11)$$

where $\mathcal{E}$ is the averaging over $\mathbf{\Theta}$ under the distribution $p(\mathbf{\Theta}|\mathbf{m}', \mathbf{D})$.

The key result of EM is that second term on the right side of equation (11) is minimized when $\mathbf{m} = \mathbf{m}'$. Thus, choosing a new labelling $\mathbf{m}$ which maximizes

$$\mathcal{E}(\log p(\mathbf{\Theta}, \mathbf{m}|\mathbf{D})) = \int (\log p(\mathbf{\Theta}, \mathbf{m}|\mathbf{D}))p(\mathbf{\Theta}|\mathbf{m}', \mathbf{D})\partial\mathbf{\Theta} \quad (12)$$

increases the posterior $p(\mathbf{m}|\mathbf{D})$. This expression is called $Q(\mathbf{m}|\mathbf{m}')$, the expected complete-data log-likelihood, in the EM literature.

In §5.1 it will be shown that we can efficiently sample from a PS which suggests a sampling based solution to maximizing (12). Let the set of $s$ samples be $\mathbf{\Theta}_1 \ldots \mathbf{\Theta}_s$, with weights $p(\mathbf{\Theta_i}|\mathbf{m}', \mathbf{D}) = w_i$, then the corresponding minimization can be written as

$$\hat{\mathbf{m}} = \arg\min_{\mathbf{m}} \sum_{i=1}^{i=s} w_i \Psi_3(\mathbf{m}, \mathbf{\Theta}_i). \quad (13)$$

This is the key equation of our approach. Section 6 describes an efficient method for minimizing the energy function (13). We observe that this energy function is a weighted linear sum of the energies $\Psi_3(\mathbf{m}, \mathbf{\Theta})$ which, being a linear combination, can also be optimized using MINCUT [13]. This demonstrates the interesting result that for Markov random fields, with latent variables, it is computationally feasible to optimize $Q(\mathbf{m}|\mathbf{m}')$.

The EM algorithm often converges to a local minima of the energy function and its success depends on the initial labelling $\mathbf{m}^0$ (i.e. the labelling $\mathbf{m}'$ at the first iteration). In the last section a generative graphical model for pixel by pixel segmentation was set up. However, it would be computationally extravagant to attempt to minimize this straight off. Rather an initialization stage is adopted in which we get a rough estimate of the object's posterior extracted from a set of image features $\mathbf{Z}$, defined in § 4.1. Image features (such as textons and edges) can provide high discrimination at low computational cost. We approximate the initial distribution $p_0(\mathbf{\Theta}|\mathbf{m}, \mathbf{D})$, as $g(\mathbf{\Theta}|\mathbf{Z})$, where $\mathbf{Z}$ are some image features chosen to localize the object in a computationally efficient manner. Thus, the weights $w_i$ required to evaluate equation (13) on the first EM iteration are obtained by sampling from the distribution $g(\mathbf{\Theta}|\mathbf{Z})$, defined in Section 4.

The next section describes the LPS model in detail. In the remainder of the paper, we describe an efficient method to obtain the samples from the posterior of a PS model required for the marginalization in equation (13), and the OBJ CUT algorithm which re-estimates the labelling $\mathbf{m}$ by minimizing equation (13). These methods are applicable to any articulated object category which can be modelled using LPS. We demonstrate the results on two quadrupeds, namely cows and horses.

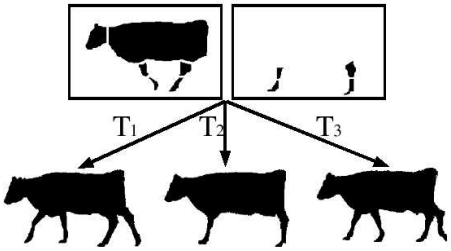


Figure 2. *Layered pictorial structure of a cow. The various parts belonging to layer 2 are shown in the top left and right image respectively. Pairwise potentials defined for every pair of parts as shown in equation (16) only allows valid configurations of a cow. Three such configurations are shown in the bottom row.*

# 4. Layered Pictorial Structures

Pictorial structures (PS) are compositions of 2D patterns, termed *parts*, under a probabilistic model for their shape, appearance and the spatial layout. When calculating the likelihood of model parameters, a typical assumption under the PS model is that the parts do not (partially) occlude each other [14]. This makes them unsuitable for segmentation which requires an accurate estimate of the poses of the parts.

In the layered pictorial structures (LPS) model introduced in [15] (and in a similar model described in [1]), in addition to shape and appearance, each part $p_i$ is also assigned a layer number $l_i$ which determines its relative depth. Several parts can have the same layer number if they are at the same depth. A part $p_i$ can partially or completely occlude part $p_j$ if and only if $l_i > l_j$. The parts of an LPS are defined as rigidly moving components of the object. In the case of side views of quadrupeds, this results in 2 layers containing a total of 10 parts: head, torso and 8 half limbs (see Fig. 2). The parts are obtained as described in § 4.2.

An LPS can also be viewed as an MRF with the sites of the MRF corresponding to parts. Each site takes one of $n_L$ part labels which encode the putative poses of the part. Let the part label at the $i^{th}$ site be $\mathbf{t}_i = (x_i, y_i, \theta_i, \sigma_i)$, where $(x_i, y_i)$ is the location, $\theta_i$ is the orientation, $\sigma_i$ is the scale.

For a given part label $\mathbf{t}_i$ and image $\mathbf{D}$, the $i^{th}$ part corresponds to the set of pixels $\mathbf{D}_i \subset \mathbf{D}$ which are used to calculate features $\mathbf{z}_i$. Let $n_P$ be the number of parts, $\mathbf{s}_i$ and $\mathbf{a}_i$ be the shape and appearance parameters for part $p_i$ and $\mathbf{T} = \{\mathbf{t}_1, ...\mathbf{t}_{n_P}\}$. Assuming that $\mathbf{D}_i$ does not include pixels accounted for by $p_j, l_j > l_i$, we get

$$p(\mathbf{Z}|\mathbf{\Theta}) = \prod_{i=1}^{i=n_P} p(\mathbf{z}_i|\mathbf{a}_i, \mathbf{s}_i), \tag{14}$$

where $\mathbf{Z} = \{\mathbf{z}_1 \dots \mathbf{z}_{n_P}\}$ are the image features.

LPS, like PS, are characterized by pairwise only dependencies between the sites. These are modelled as a prior on the part labels $\mathbf{T}$:

$$p(\mathbf{T}) \propto \exp\left(-\sum_{i=1}^{i=n_P}\sum_{j=1, j\neq i}^{j=n_P} \alpha(\mathbf{t}_i, \mathbf{t}_j)\right). \tag{15}$$

Note that we use a completely connected MRF. In our approach, the pairwise potentials $\alpha(\mathbf{t}_i, \mathbf{t}_j)$ are given by a Potts model, i.e.

$$\alpha(\mathbf{t}_i, \mathbf{t}_j) = \begin{cases} d_1 & \text{if valid configuration} \\ d_2 & \text{otherwise,} \end{cases} \tag{16}$$

where $d_1 < d_2$. In other words, all valid configurations are considered equally likely and have a smaller cost. A configuration is valid provided the relative shape parameters of the two poses lie within a box, i.e. if $\mathbf{t}_{ij}^{min} \leq |\mathbf{t}_i - \mathbf{t}_j| \leq \mathbf{t}_{ij}^{max}$, where $\mathbf{t}_{ij}^{min} = \{x_{ij}^{min}, y_{ij}^{min}, \theta_{ij}^{min}, \sigma_{ij}^{min}\}$ and $\mathbf{t}_{ij}^{max} = \{x_{ij}^{max}, y_{ij}^{max}, \theta_{ij}^{max}, \sigma_{ij}^{max}\}$ are learnt using training video sequences as described in § 4.2. The posterior of the model parameters is given by

$$g(\mathbf{\Theta}|\mathbf{Z}) \propto \prod_{i=1}^{i=n_P} p(\mathbf{z}_i|\mathbf{a}_i, \mathbf{s}_i) \exp\left(-\sum_{j\neq i} \alpha(\mathbf{t}_i, \mathbf{t}_j)\right) \tag{17}$$

We now describe how we model the likelihood of the parts of the LPS.

## 4.1. Feature Likelihood for Parts

We define the features $\mathbf{Z}$ extracted from the pixels $\mathbf{D}$. Here we use two types of features $\mathbf{z}_i(\mathbf{D}_i) = (z_1(\mathbf{D}_i), z_2(\mathbf{D}_i))$ for the shape and appearance of the part respectively. The likelihood based on the whole data is approximated as

$$p(\mathbf{z}_i|\mathbf{a}_i, \mathbf{s}_i) = p(z_1|\mathbf{s}_i)p(z_2|\mathbf{a}_i) \tag{18}$$

where $p(z_1|\mathbf{s}_i) = \exp(-z_1)$ and $p(z_2|\mathbf{a}_i) = \exp(-z_2)$.

**Outline ($z_1(\mathbf{D}_i)$):** In order to handle the variability in shape among members of an object class (e.g. horses), it is necessary to represent the part outline by a set of exemplar curves. Chamfer distances are computed for each exemplar for each pose $\mathbf{t}_i$. The first feature $z_1(\mathbf{D}_i)$ is the minimum of the truncated chamfer distances over all the exemplars of $p_i$ at pose $\mathbf{t}_i$. Truncated chamfer distance measures the similarity between two shapes $\mathcal{U} = (u_1, u_2, ...u_n)$ and $\mathcal{V} = (v_1, v_2, ...v_m)$. It is the mean of the distances between each point $u_i \in \mathcal{U}$ and its closest point in $\mathcal{V}$:

$$d_{cham} = \frac{1}{n}\sum_i \min\{\min_j ||u_i - v_j||, \tau_1\}, \tag{19}$$

where $\tau_1$ is a threshold for truncation which reduces the effect of outliers and missing edges. Edge orientation is included by computing the chamfer score only for edges with similar orientation, in order to make the distance function more robust [10]. We use 8 orientation groups for edges.

**Texture ($z_2(\mathbf{D}_i)$):** Similar to the outline of a part, we represent the texture of an object by a set of examples. We use the VZ classifier [21] which obtains a texton dictionary by clustering the vectorized raw intensities of $N \times N$ neighbourhood of each pixel in the exemplars. In this paper, we

use $N = 3$. The exemplars are then modelled as a histogram of pixel texton labellings [17]. The feature $z_2(\mathbf{D}_i)$ is defined as the minimum $\chi^2$ distance of the histogram of texton labellings for $\mathbf{D}_i$ with the histogram modelling the exemplars. We now describe how the LPS parameters are learnt so as to handle intra-class variability in shape and appearance.

### 4.2. Learning the LPS

The various parameters of the LPS model are learnt using the method described in [15] which divides a scene in a video into rigidly moving components and provides the segmentation of each frame. We use this approach on 20 cow videos of 45 frames each[1] which gives us multiple shape exemplars for each part required to compute the feature $z_1$ and multiple texture examples for calculating $z_2$ along with the layer numbers of all parts. Furthermore, this provides us with an estimate of $|\mathbf{t}_i - \mathbf{t}_j|$, for each frame and for all pairs of parts $p_i$ and $p_j$, which is used to compute the parameters $\mathbf{t}_{ij}^{min}$ and $\mathbf{t}_{ij}^{max}$ that define valid configurations.

To obtain the shape exemplars and texture examples for horses, we use 20 segmented images of horses[2]. A point to point correspondence is established over the outline of a cow from a training video to the outlines of the horses using shape context with continuity constraint [20]. Using this correspondence and the learnt parts of the cow, the parts of the horse are determined. The part correspondence thus obtained maps the parameters $\mathbf{t}_{ij}^{min}$ and $\mathbf{t}_{ij}^{max}$ that were learnt for cows to horses. In the next section, we describe an efficient algorithm for matching the LPS model to the image.

## 5. Sampling the LPS

Given an image, our objective is to match the LPS model to the image to obtain samples from the distribution $g(\mathbf{\Theta}|\mathbf{Z})$. We achieve this in two stages: (i) *Initialization*, where we fit a PS model of the object to a given image $\mathbf{D}$ as described in [14] without considering the layer numbers of the parts, and (ii) *Refinement*, where the initial estimate is refined by learning an RGB distribution for the object and background and compositing the parts in descending order of their layer number before sampling for the LPS parameters. We develop a novel algorithm for efficient sampling which generalizes the method described in [7] to non-grid based MRFs.

### 5.1. Initial estimation of poses

We find the initial estimate of the poses of the LPS for an image $\mathbf{D}$ in two stages: (i) *part detection*, or finding putative poses for each part along with the corresponding likelihoods and, (ii) *estimating posteriors* of the putative poses.

**Part detection:** The putative poses of the parts are found using *tree cascade of classifiers* as described in [14]. In our experiments, we constructed a 3-level tree by clustering the templates using a cost function based on chamfer distance. We use 20 exemplars per part, with discrete rotations between $-\pi/4$ and $\pi/4$ in intervals of $0.1$ radians and scales between $0.7$ and $1.3$ in intervals of $0.1$.

The edge image of $\mathbf{D}$ is found using edge detection with embedded confidence [18]. The feature $z_1(\mathbf{D}_i)$ (truncated chamfer distance) is computed efficiently by using a distance transform of the edge image. The feature $z_2(\mathbf{D}_i)$ is computed only at level 3 of the tree cascade by efficiently determining the nearest neighbour of the histogram of texton labelling of $\mathbf{D}_i$ among the histogram of texture examples using the method described in [12].

The putative poses $\mathbf{t}_i$ of parts $p_i$ are found by rejecting bad poses by traversing through the tree structure starting from the root node. The likelihoods $p(\mathbf{D}_i|\mathbf{a}_i, \mathbf{s}_i)$ are approximated by feature likelihoods $p(\mathbf{z}_i|\mathbf{a}_i, \mathbf{s}_i)$ shown in equation (18).

**Estimating posteriors:** A method to compute the posteriors of the putative poses is required. We use loopy belief propagation (LBP) to find the posterior probability of $p_i$ having a part label $\mathbf{t}_i$. LBP is a message passing algorithm proposed by Pearl [19]. It is a Viterbi-like algorithm for graphical models with loops.

The message that $p_i$ passes to its neighbour $p_j$ at iteration $t$ is a vector of length equal to the number of discrete part labels $n_L$ of $p_j$ and is given by:

$$m_{ij}^t(\mathbf{t}_j) \leftarrow \sum_{\mathbf{t}_i} p(\mathbf{z}_i|\mathbf{a}_i, \mathbf{s}_i) \exp(-\alpha(\mathbf{t}_i, \mathbf{t}_j)) \prod_{s \neq i, s \neq j} m_{si}^{t-1}(\mathbf{t}_i). \tag{20}$$

The beliefs (posteriors) after $T$ iterations are calculated as:

$$b^T(\mathbf{t}_i) = p(\mathbf{z}_i|\mathbf{a}_i, \mathbf{s}_i) \prod_{s \neq i} m_{si}^T(\mathbf{t}_i). \tag{21}$$

The algorithm is said to have converged when the rate of change of all beliefs falls below a certain threshold. The time complexity of this algorithm is $O(n_P n_L^2)$ where $n_P$ is the number of parts in the LPS and $n_L$ is the number of putative poses per part. This makes sampling infeasible for large $n_L$ which is the case with smaller parts of the LPS model such as the half-limbs. Thus, we develop an efficient novel algorithm for LBP for the case where the pairwise potentials are given by a Potts model as shown in equation (16). The algorithm exploits the fact that the number of pairs of part labels $n'_L$, one for each of the two parts $p_i$ and $p_j$, which form a valid configuration is much smaller than the total number of such pairs, $n_L^2$, i.e. $n'_L \ll n_L^2$. Note that a similar method is described in [7] which takes advantage of fast convolutions using FFT. However, it is restricted to MRFs with regularly discretized labels, i.e. the labels lie on a grid, which is not true for putative poses of parts of the LPS.

Let $\mathcal{C}_i(\mathbf{t}_j)$ be the set of part labels of $p_i$ which form a valid pairwise configuration with $\mathbf{t}_j$. The part labels $\mathcal{C}_i(\mathbf{t}_j)$ are computed just once before running LBP.

[1] Courtesy Derek Magee, University of Leeds

[2] Courtesy Eran Borenstein, Weizmann Institute of Science

Figure 3. *The first column shows the initial estimate obtained for poses of parts of a cow in two images (see § 5.1). The half-limbs tend to overlap since layer numbers are not used. Refined estimates of the poses obtained using the* RGB *distribution of foreground and background together with the* LPS *model are shown in the second column (see § 5.2). The parts are shown overlaid on the image. The third column shows the segmentation obtained using the* OBJ CUT *algorithm (see § 6).*

We define

$$T(i,j) = \sum_{\mathbf{t}_i} p(\mathbf{z}_i|\mathbf{a}_i,\mathbf{s}_i) \prod_{s\neq i, s\neq j} m_{si}^{t-1}(\mathbf{t}_i), \qquad (22)$$

which is independent of the part label $\mathbf{t}_j$ of $p_j$ and needs to be calculated only once before $p_i$ passes a message to $p_j$. It is clear from equation (20) that if no part label $\mathbf{t}_i$ forms a valid configuration with $\mathbf{t}_j$, then the message $m_{ij}(\mathbf{t}_j)$ is simply $\exp(-d_2)T(i,j)$. To compute the contribution of the labels $\mathbf{t}_i \in \mathcal{C}_i(\mathbf{t}_j)$ in computing $m_{ij}(\mathbf{t}_j)$ we define

$$S(i,\mathbf{t}_j) = \sum_{\mathbf{t}_i \in \mathcal{C}_i(\mathbf{t}_j)} p(\mathbf{z}_i|\mathbf{a}_i,\mathbf{s}_i) \prod_{s\neq i, s\neq j} m_{si}^{t-1}(\mathbf{t}_i), \qquad (23)$$

which is computationally inexpensive to calculate since $\mathcal{C}_i(\mathbf{t}_j)$ consists of very few part labels. The message $m_{ij}^t(\mathbf{t}_j)$ is calculated as

$$m_{ij}^t(\mathbf{t}_j) \leftarrow \exp(-d_1)S(i,\mathbf{t}_j) + \exp(-d_2)(T(i,j) - S(i,\mathbf{t}_j)). \qquad (24)$$

Our method speeds up LBP by a factor of nearly $n_L$. Extension to Generalized Potts model is trivial. The beliefs computed using LBP allow us to determine the MAP estimate which provides the initial estimate of the poses of the parts. Fig. 3 (column 1) shows the initial estimate obtained for two cow images. The initial estimate is refined using the LPS model as described below.

### 5.2. Layerwise refinement

Once the initial estimate of the parts is obtained, we refine it by using the colour of the object and background together with the LPS model. The colour of the object and background are represented as histograms $\mathcal{H}_{obj}$ and $\mathcal{H}_{bkg}$ of RGB values learnt using the initial estimate. The feature $z_2(\mathbf{D}_i)$ is now defined such that

$$p(z_2|\mathbf{a}_i) = \prod_{x\in\mathbf{D}_i} \frac{p(x|\mathcal{H}_{obj})}{p(x|\mathcal{H}_{bkg})}. \qquad (25)$$

The refined estimate of the poses are obtained by compositing the parts of the LPS in descending order of their layer numbers as follows. When considering layer $l_i$, putative poses of the parts $p_j$ belonging to $l_i$ are found using the tree cascade of classifiers around the initial estimate of $p_j$. In our experiments, we consider locations which are at most at a distance of 15% of the size of the object as given by the initial estimate. When computing the likelihood of the part at a given pose, pixels which have already been accounted for by a previous layer are not considered. The posteriors over the putative poses is computed using the efficient LBP algorithm.

### 5.3. Sampling the LPS

One might argue that if the MAP estimate of the poses has a very high posterior compared to other configuration of poses, then equation (13) can be approximated using only the MAP estimate $\mathbf{\Theta}^*$ instead of the samples $\mathbf{\Theta}_1 \ldots \mathbf{\Theta}_s$. However, we found that this is not the case especially when the RGB distribution of the background is similar to that of the object. Fig. 3 (column 2) shows the MAP estimate of the refined poses of the parts using the initial estimate shown in Fig. 3 (column 1). Note that the legs of the first cow in Fig. 3 (column 2) are detected incorrectly since the parts of the background have roughly the same colour as the cow. Thus, it is necessary to use multiple samples of the LPS model.

We describe the method for sampling for 2 layers. The extension to an arbitrary number of layers is trivial. To obtain a sample $\mathbf{\Theta}_i$, parts belonging to layer 2 are considered first and the posterior over their putative poses is computed using LBP. The posterior is then sampled for poses of parts, one part at a time, such that the pose of the part being sampled forms a valid configuration with the poses of the parts previously sampled. The process is repeated to obtain multiple samples $\mathbf{\Theta}_i$ which do not include the poses of parts belonging to layer 1. This method of sampling is efficient since $\mathcal{C}_i(j)$ are pre-computed and contain very few part labels. The best $n_S$ samples, with the highest belief, are chosen.

To obtain the poses of parts in layer 1 for sample $\mathbf{\Theta}_i$, we fix the poses of parts belonging to layer 2 as given by $\mathbf{\Theta}_i$ and calculate the posterior over the poses of parts in layer 1 using LBP. We sample this posterior for poses of parts such that they form a valid configuration with the poses of the parts in layer 2 and with those previously sampled. As in the case of layer 2, multiple samples are obtained and the best $n_S$ samples are chosen. The process is repeated for all samples $\mathbf{\Theta}_i$ for layer 2, resulting in a total of $n_S^2$ samples.

However, since computing the likelihood of the parts in layer 1 for each $\mathbf{\Theta}$ is inefficient, we approximate by using only those poses whose overlap with layer 2 is below a threshold $\tau$. Fig. 4 shows some of the samples obtained using the above method for cows shown in Fig. 3. We now describe the OBJ CUT algorithm for segmentation.

## 6. Estimation, OBJ CUT

Given an image $\mathbf{D}$ containing an instance of a known object category, and the samples $\mathbf{\Theta}_1 \ldots \mathbf{\Theta}_s$ of the LPS parameters, we wish to obtain the segmentation of the object, i.e. infer labels $\mathbf{m}$. We now present the OBJ CUT algorithm which

Figure 4. *Posteriors over the putative poses of parts are calculated using* LBP. *The posterior is then sampled to obtain instances of the object (see § 5.3) . The half-limbs are detected correctly in some samples.*

provides reliable segmentation using both (a) modelled and (b) unmodelled deformations of articulated object categories.

**Modelled deformations.** These are taken into account by the LPS model which uses multiple shape exemplars for each part and allows for all valid configurations of the object category using pairwise potentials $\alpha(\mathbf{t}_i, \mathbf{t}_j)$. The various samples $\mathbf{\Theta}_i$ localize the parts of the object in the image. They also provide us with refined estimates of the histograms $\mathcal{H}_{obj}$ and $\mathcal{H}_{bkg}$ which model the appearance of the figure and ground.

**Unmodelled deformations.** These are accounted for by merging pixels surrounding the object which are similar in appearance to the object. Only those pixels which lie in a 'band' surrounding the outline of the object are considered. The width of the band is 10% of the size of the object as specified by the sample of the LPS. Points lying inside the object are given preference over the points surrounding the object. As is the case with MRF-based segmentations, boundaries are preferred around image edges.

The segmentation is obtained by minimizing equation (13) using the MINCUT algorithm. The various terms in equation (13) are defined as follows. The weights $w_i$ are approximated as $w_i \approx g(\mathbf{\Theta}_i|\mathbf{Z})$. The data likelihood term $\phi(\mathbf{D}|m_x)$ is computed using equation (3). The contrast term is given by equations (4) and (6). The function $\phi(m_x|\mathbf{\Theta}_i)$ is defined by equation (8). Table 1 summarizes the main steps of obtaining the segmentation using the OBJ CUT algorithm.

The figure-ground labelling $\mathbf{m}$ obtained as described above can be used iteratively to refine the segmentation using the EM algorithm. However, we found that this does not result in a significant improvement over the initial segmentations as the samples $\mathbf{\Theta}_1 \ldots \mathbf{\Theta}_s$ do not change much from one iteration to the other. We present several results of the OBJ CUT algorithm in the next section and compare it with a state-of-the-art method and ground truth.

1. Given an image $\mathbf{D}$, an object category is chosen, e.g. cows or horses.
2. The corresponding LPS model is matched to $\mathbf{D}$ to obtain the samples $\mathbf{\Theta}_1 \ldots \mathbf{\Theta}_s$.
3. The objective function given by equation (13) is determined by computing $\Psi_3(\mathbf{m}, \mathbf{\Theta}_i)$ and using $w_i \approx g(\mathbf{\Theta}_i|\mathbf{Z})$.
4. The objective function is minimized using a single MINCUT operation to obtain the segmentation $\mathbf{m}$.

Table 1: *The* OBJ CUT *algorithm*

## 7. Results

We present the segmentation results obtained using OBJ CUT for cows and horses. In all our experiments, we set the values of the parameters as $P = \lambda = 0.1$, $\sigma = 5$ and $\mu = 0.2$. Fig 3 (column 3) shows the results of the OBJ CUT algorithm for two cow images. Figure 5 shows the segmentation of various images of cows and horses. These images were manually segmented to obtain ground truth for comparison. For the cow images, out of the 125,362 foreground pixels and 472,670 background pixels present in the ground truth, 120,127 (95.82%) and 466,611 (98.72%) were present in the segmentations obtained. Similarly, for the horses images, out of the 79,860 foreground pixels and 151,908 background pixels present in the ground truth, 71,397 (89.39%) and 151,185 (99.52%) were present in the segmentations obtained. In the case of horses, most errors are due to unmodelled mane and tail parts. Results indicate that, by considering both modelled and unmodelled deformations, excellent results were obtained by OBJ CUT.

Figure 6 shows a comparison of the segmentation results obtained when using OBJ CUT with a state-of-the-art method for object category specific segmentation described in Leibe and Schiele [16]. Note that a similar approach was described in [4]. The OBJ CUT algorithm provides better segmentations using significantly smaller number of exemplars by exploiting the ability of MINCUT for providing excellent segmentations using a good initialization obtained by LPS.

## 8. Summary and Conclusions

We presented a new model, called the Object Category Specific MRF, which combines the LPS and the MRF model to perform object category specific segmentation. Reliable segmentation is obtained using OBJ CUT, which considers both modelled deformation, provided by the LPS, and unmodelled deformation. An efficient method for LBP is developed which, together with the MINCUT, makes OBJ CUT computationally feasible. The method needs to be extended to handle multiple visual aspects of an object category and to deal with partial occlusion.

**Acknowledgments.** We thank Dan Huttenlocher for fruitful discussions on efficient LBP. This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

## References

[1] A. Agarwal and B. Triggs. Tracking articulated motion using a mixture of autoregressive models. In *ECCV*, pages III:54–65, 2004.

[2] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV*, page IV: 113 ff., 2002.

[3] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr. Interactive image segmentation using an adaptive GMMRF model. In *ECCV*, pages Vol I: 428–441, 2004.

[4] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*, page II: 109 ff., 2002.

[5] Y. Boykov and M. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. In *ICCV*, pages I: 105–112, 2001.

Figure 5. *Segmentation results. The first two images in each row show some of the samples of the* LPS *model. The segmentation obtained using the Object Category Specific* MRF *is shown in the last column. Most of the errors were caused by the mane (in the case of horses) and tail (which were not a part of the* LPS *model) and parts of the background which were close and similar in colour to the object.*

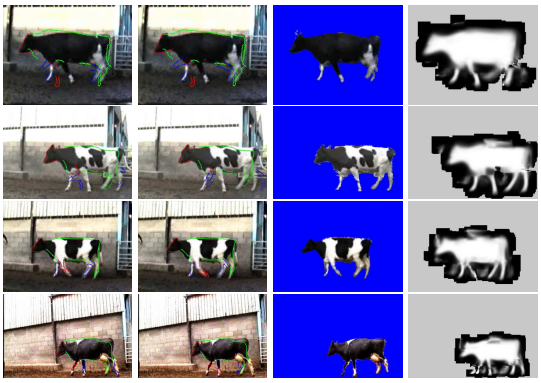

Figure 6. *Comparison with Leibe and Schiele. The first two images of each row show some of the samples obtained by matching the* LPS *model to the image. The third image is the segmentation obtained using the* OBJ CUT *algorithm and the fourth image is the segmentation obtained using [16]. Note that* OBJ CUT *provides a better segmentation of the torso and head without detecting extra half limbs.*


[6] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *CVPR*, pages II: 66–73, 2000.

[7] P. Felzenszwalb and D. Huttenlocher. Fast algorithms for large state space HMMs with applications to web usage analysis. In *NIPS*, 2003.

[8] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, pages II: 264–271, 2003.

[9] V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation by image exploration. In *ECCV*, pages Vol I: 40–54, 2004.

[10] D. Gavrilla. Pedestrian detection from a moving vehicle. In *ECCV*, pages II: 37–49, 2000.

[11] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman and Hall, 1995.

[12] J. Goldstein, J. Platt, and C. Burges. Indexing high-dimensional rectangles for fast multimedia identification. Technical Report MSR-TR-2003-38, Microsoft Research, 2003.

[13] V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts. *IEEE PAMI*, 26(2):147–159, 2004.

[14] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Extending pictorial structures for object recognition. In *BMVC*, pages II: 789–798, 2004.

[15] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Learning layered pictorial structures from video. In *ICVGIP*, pages 148–153, 2004.

[16] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC*, pages II: 264–271, 2003.

[17] T. Leung and J. Malik. Recognizing surfaces using three-dimensional textons. In *ICCV*, pages 1010–1017, 1999.

[18] P. Meer and B. Georgescu. Edge detection with embedded confidence. *PAMI*, 23:1351–1365, December 2001.

[19] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffman, 1998.

[20] A. Thayananthan, B. Stenger, P. Torr, and R. Cipolla. Shape context and chamfer matching in cluttered scenes. In *CVPR*, pages I: 127–133, 2003.

[21] M. Varma and A. Zisserman. Texture classification: Are filter banks necessary? In *CVPR*, pages II:691–698, 2003.