Extending Pictorial Structures for Object Recognition

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

The goal of this paper is to recognize various deformable objects from images. To this end we extend the class of generative probabilistic models known as *pictorial structures*. This class of models is particularly suited to represent articulated structures, and has previously been used by Felzenszwalb and Huttenlocher for pose estimation of humans. We extend pictorial structures in three ways: (i) likelihoods are included for both the boundary and the enclosed texture of the animal; (ii) a complete graph is modelled (rather than a tree structure); (iii) it is demonstrated that the model can be fitted in polynomial time using belief propagation.

We show examples for two types of quadrupeds, cows and horses. We achieve excellent recognition performance for cows with an equal error rate of 3% for 500 positive and 5000 negative images.

1 Introduction

One of the fundamental problems in the field of Computer Vision is object recognition. Indeed research in this area has seen a mini Renaissance [1, 3, 5, 8, 10, 12, 15], with much interest focussing on the recognition of object *categories* rather than individual objects, e.g. recognizing cows rather than a particular cow ('Daisy'). In this paper, we present a method to recognize objects and demonstrate with two types of quadrupeds: horses and cows. Fig. 1 shows an example of a horse being recognized using our approach.

When attempting to recognize object categories there might be substantial spatial and colour variation between individual instances of that category, e.g. the variation in the texture of the cows shown in Fig 8. Furthermore, pose, lighting and occlusion result in variability in appearance of an object instance. In order to manage this variability there is a broad agreement that object categories should be represented by a collection of

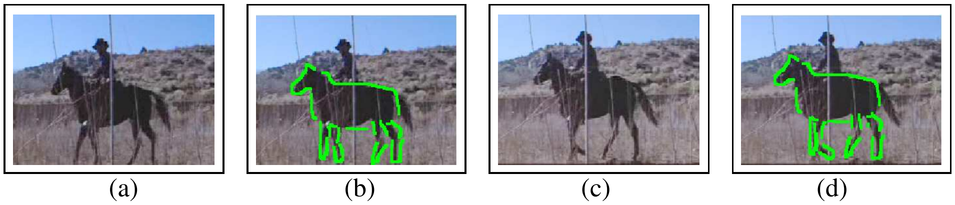


Figure 1: *Clint Eastwood's horse is recognized in frames from the movie "The Outlaw Josey Wales". Figures (a) and (c) show two frames of a shot. Figures (b) and (d) show the output of our horse category recognition method on these frames. The green lines show the outline of the parts detected.*

BMVC 2004 doi:10.5244/C.18.81

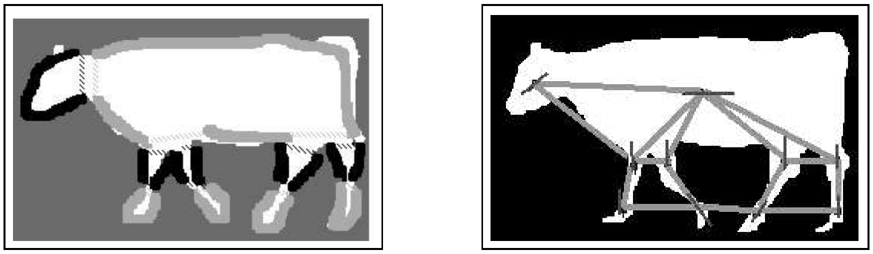


Figure 2: *Example pictorial structure of a cow. (a) Various parts p_i of the cow (e.g. head, torso and legs) (b) The black lines show the location and orientation of the parts and the grey lines show some of the connections between parts.*

spatially related parts each with its own appearance. This sort of approach dates back to the *pictorial structures* (PS) model introduced by Fischler and Elschlager three decades ago [6].

Recent work has shown excellent recognition results using PS for varied object categories [5]. Furthermore, Felzenszwalb and Huttenlocher [4] have developed computationally efficient methods for finding the best fit of the PS model in an image. Ramanan and Forsyth [12] have used PS to represent and detect animals walking or running in fronto-parallel planes in a video.

We extend the articulated pictorial structure of Felzenszwalb and Huttenlocher [4] in a number of ways. In particular, both the outline and the enclosed texture of the part are included in its appearance parameters and all parts are connected to each other to form a complete graph instead of a tree structure. A properly normalized measure of the probability of a part being present at a location is modelled using the PDF projection theorem [2]. Sections 2 and 3 describe the proposed model in detail. Section 4 describes how the model parameters are learnt and outlines a computationally efficient method for fitting the model to previously unseen images. Section 5 presents several recognition results. We summarize in section 6 and suggest some directions for future work.

2 Bayesian pictorial structures

Pictorial structures (PS) are compositions of 2D patterns, termed *parts*, under a probabilistic model for both the appearance and the spatial layout. A PS can be viewed as a Markov random field (MRF) with the sites of the MRF corresponding to parts such that the PS would provide a generative model for the object of interest. By generative we mean that given an image of an object, we can assign it a likelihood (possibly unnormalized).

Previous works differ in the way they divide the object into parts for the PS representation. While [12] finds the parts by locating parallel lines across a video, the methods described in [1, 5, 15] define parts as sub-regions of the object. We observe that for the connections between parts to truly represent the spatial layout, all points belonging to a part must always move together rigidly. Thus, we define the parts of a PS as rigidly moving components of the object. In the case of quadrupeds, this results in 10 parts: head, torso and 8 half limbs (see Fig. 2).

Each site takes one of n_L labels which encode the putative poses of the part. Let the label at the i^{th} site be $l_i = (x_i, y_i, \theta_i, \sigma_i, \varphi_i)$, where (x_i, y_i) is the location, θ_i is the orientation, σ_i is the scale and φ_i is equal to 0 or 1 depending on whether the part is occluded or not. For a given label l_i and data (image) $\mathbf{D}$, the i^{th} part maps to the set of pixels $\mathbf{D}_i \subset \mathbf{D}$. Let n_P be the number of parts. Given an image $\mathbf{D}$, the posterior distribution for the model parameters is given by

$$\Pr(\mathbf{a},\mathbf{l}|\mathbf{D}) = \frac{\Pr(\mathbf{D}|\mathbf{a},\mathbf{l})\Pr(\mathbf{a})\Pr(\mathbf{l})}{\Pr(\mathbf{D})}. \tag{1}$$

where $\mathbf{a}$ is the appearance parameters and $\mathbf{l} = \{l_1, l_2, ...l_{n_P}\}$. Let $\mathbf{a}_i$ be the appearance parameters for part p_i and $\mathbf{a}_{bg}$ be the appearance parameters for the background. By assuming that the parts do not overlap, we get

$$\Pr(\mathbf{D}|\mathbf{a},\mathbf{l}) = \prod_{i=1}^{i=n_P} \Pr(\mathbf{D}_i|\mathbf{a}_i)\Pr(\mathbf{D}'|\mathbf{a}_{bg}), \tag{2}$$

where $\mathbf{D}' = \mathbf{D} - \bigcup_i \mathbf{D}_i$. We can compute the likelihood ratio of the object being present in the image $\mathbf{D}$ to the object being absent as

$$\Rightarrow \frac{\Pr(\mathbf{D}|\mathbf{a},\mathbf{l})}{\Pr(\mathbf{D}|\mathbf{a}_{bg},\mathbf{l})} = \prod_{i=1}^{i=n_P} \frac{\Pr(\mathbf{D}_i|\mathbf{a}_i)}{\Pr(\mathbf{D}_i|\mathbf{a}_{bg})}. \tag{3}$$

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

$$\Pr(\mathbf{l}) \propto \exp\left(-\sum_{i=1}^{i=n_P}\sum_{j=1}^{j=n_P} \psi(l_i, l_j)\right). \tag{4}$$

Note that we use a completely connected MRF. The benefits of using a complete graph, instead of a tree structure used in [4], are demonstrated in section 5. In our approach, the pairwise potentials $\psi(l_i, l_j)$ are given by a Potts model, i.e.

$$\begin{aligned} \psi(l_i, l_j) &= 0,\ if\ \ valid\ \ configuration, \\ &= \texttt{const},\ \textit{otherwise}. \end{aligned} \tag{5}$$

In other words, all valid configurations are considered equally likely and have no cost. Valid configurations are learnt using training video sequences as described in § 4.1. Given an image, $\mathbf{D}$, the best fit of the model is found by maximizing

$$\Pr(\mathbf{a},\mathbf{l}|\mathbf{D}) \propto \prod_{i=1}^{i=n_P} \frac{\Pr(\mathbf{D}_i|\mathbf{a}_i)}{\Pr(\mathbf{D}_i|\mathbf{a}_{bg})} \exp\left(-\sum_{j\neq i} \psi(l_i, l_j)\right) \tag{6}$$

In our model, appearance parameters $\mathbf{a}$ model both the shape and texture of the parts. The next section describes how we model the likelihood of the parts of the PS.

3 Likelihood of parts

It is not immediately obvious how to estimate the likelihood ratio given in equation (3). Our approach is to extract a set of sufficient statistics for classification. A statistic $\mathbf{z}_i(\mathbf{D}_i)$ is a function of the image $\mathbf{D}_i$ and will be denoted simply as $\mathbf{z}_i$. If $\mathbf{z}_i$ is a sufficient statistic then by the PDF projection theorem [2]

$$\frac{\Pr(\mathbf{D}_i|\mathbf{a}_i)}{\Pr(\mathbf{D}_i|\mathbf{a}_{bg})} = \frac{\Pr(\mathbf{z}_i|\mathbf{a}_i)}{\Pr(\mathbf{z}_i|\mathbf{a}_{bg})}, \tag{7}$$

i.e. we hope that the features $\mathbf{z}_i$ are as good as the original data for detecting the object (an assumption implicit whenever features, rather than pixels, are used). Although it is difficult to prove the sufficiency requirement in most cases, near optimal performance can be obtained even if this requirement is not completely satisfied [14].

For this paper, we select two statistics, noting that others could be used. However, it will be seen later that these yielded good results. These two statistics $\mathbf{z}_i = (z_1(\mathbf{D}_i), z_2(\mathbf{D}_i))$ together model both the shape and appearance of each part p_i of the PS. The probability distributions for $\Pr(\mathbf{z}_i|\mathbf{a}_i)$ and $\Pr(\mathbf{z}_i|\mathbf{a}_{bg})$ are modelled as 2D normal distributions whose parameters are learnt as described in § 4.1.

Outline ($z_1(\mathbf{D}_i)$): In order to handle the variability in shape among members of an object class (e.g. cows), it is necessary to represent the part outline by a set of exemplar curves (see Fig. 3). Chamfer distances are computed for each exemplar for each pose l_i. The first statistic $z_1(\mathbf{D}_i)$ is the minimum of the truncated chamfer distances over all the exemplars of p_i at pose l_i. Truncated chamfer distance measures the similarity between two shapes $\mathscr{U} = (u_1, u_2, ... u_n)$ and $\mathscr{V} = (v_1, v_2, ... v_m)$. It is the mean of the distances between each point $u_i \in \mathscr{U}$ and its closest point in $\mathscr{V}$:

$$d_{cham} = \frac{1}{n} \sum_i \min\{\min_j ||u_i - v_j||, \tau_1\}, \tag{8}$$

where τ_1 is a threshold for truncation which reduces the effect of outliers and missing edges.

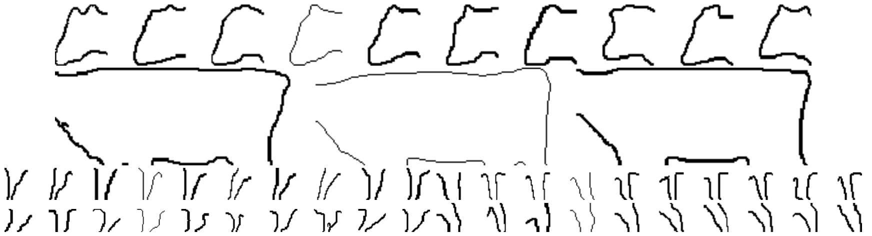


Figure 3: *Examples of exemplars for head, body and legs of a cow extracted from various instances of the object category.*

Texture ($z_2(\mathbf{D}_i)$): It might be thought that a representative set of textures could be learnt (similar to the representative set of exemplars for the outline). However, there is considerable variation in the texture of cows over breeds, e.g. Jersey, Ayrshire, Guernsey, and this means that at least one example of each breed must be included. Instead we use a weak model for the texture, and this has proved sufficient to aid in distinguishing foreground from background regions. We model the intensity values of the pixels belonging to the object as a Gaussian mixture model (GMM) of two Gaussians which captures the nature of texture for a cow – which is essentially either one or two colours (with image variation due to lighting, shadows etc.). To capture the intra-class variability in texture, multiple GMMs are used. In our experiments, we used 20 GMMs. The statistic $z_2(\mathbf{D}_i)$ is the maximum of the sum of the log of probabilities of the intensities of the pixels within the region enclosed by p_i over all GMMs.

The next section describes how the model parameters are learnt and how the maximum a posterior (MAP) estimate of the PS is found by maximizing equation (6).

4 Model implementation

The number of labels n_L has the potential to be very large. Consider discretization of (x,y,θ,σ) into 360×240 for (x,y) with 15 orientations and 7 scales at each location. This results in $9,072,000$ poses which causes some computational difficulty when determining the MAP estimate of the PS.

Felzenszwalb and Huttenlocher [4] advocate maintaining all labels and suggest an $O(n_L n_P)$ algorithm for finding the MAP estimate of the PS by restricting the pairwise potentials to normal distributions. However, this approach would no longer be computationally feasible as the number of parameters used to represent a pose l_i increase (e.g. 6 parameters for affine or 8 parameters for projective). We propose finding the best fit of the PS for an image $\mathbf{D}$ in two stages: (i) *part detection*, or finding putative positions for each part along with the corresponding likelihoods, and (ii) MAP *estimation* of the PS.

During part detection, we consider the same amount of discretization as in [4]. However, using a strong appearance model along with discriminative features allows us to consider only a small number of candidate poses, n_L, per part by discarding the poses with low likelihood. We found that using a few hundred poses per part, instead of the millions of poses used in [4], was sufficient. The MAP estimate of the PS is then found using an $O(n_L^2 n_P)$ algorithm which does not place any restrictions on the pairwise potentials. Even though we use a Potts model for pairwise potentials, the extension to any other parametric or non-parametric distribution is trivial. The algorithm is computationally feasible due to the small value of n_L. It is worth noting that, unlike our approach, [4] only finds the MAP estimate of the poses $\mathbf{l}$ without considering the appearance parameters $\mathbf{a}$ and thus, does not describe a truly generative model. We now describe how the parameters of the PS are learnt.

4.1 Learning model parameters

The exemplars for various parts of the PS (as shown in Fig. 3) and other model parameters are learnt using training video sequences. Rigidly moving parts are identified and valid configurations are learnt for each video sequence using the method described in [7]. Each video sequence also provides the intensity values of pixels belonging to the object. These are then used to learn the parameters of a GMM, which models the texture of the object, using the EM algorithm. In our experiments, 20 videos of 45 frames each were used.

The parameters of the normal distributions which model $\Pr(\mathbf{z}_i|\mathbf{a}_i)$ and $\Pr(\mathbf{z}_i|\mathbf{a}_{bg})$ are learnt by computing $\mathbf{z}_i$ for a number of positive and negative examples of p_i. Positive examples are provided by the training video sequences. We use windows over a few hundred background images obtained from the web as negative examples. We now describe the two stages of our approach in detail.

4.2 Part detection

The putative poses of the parts are found efficiently using a *tree cascade of classifiers*. The cascade efficiently discards poses with low likelihood as pointed out in [13]. When matching many similar templates to an image, a significant speed-up is achieved by forming a template hierarchy and using a coarse-to-fine search. The idea is to group similar templates together with an estimate of the variance of the error within the cluster, which is then used to define a matching threshold. The prototype is first compared to the image; only if the error is below the threshold are the individual templates within the cluster compared to the image. This clustering is done at various levels, resulting in a hierarchy, with the templates at the leaf level covering the space of all possible templates (see Fig. 4).

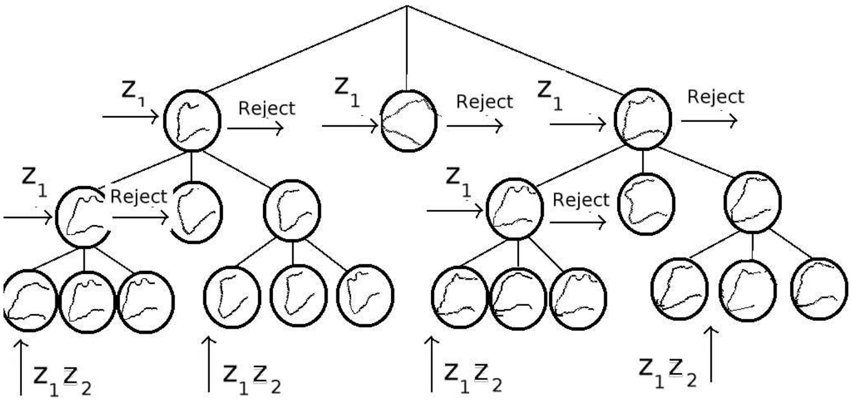


Figure 4: *The putative poses of the parts, e.g. the head, are calculated using a cascade of classifiers. A 3-level tree structure is used to prune away the bad poses by thresholding on the chamfer distance. The statistic $z_2(\mathbf{D}_i)$ is measured only at the third level of the tree since it is computationally expensive.*

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 [9] (a variation on Canny in which a confidence measure is computed from an ideal edge template). The statistic $z_1(\mathbf{D}_i)$ (truncated chamfer distance) is computed efficiently by using a distance transform of the edge image. This transformation assigns to each pixel in the edge image the minimum of τ_1 and the distance to its nearest edge pixel. Truncated chamfer distance is then calculated efficiently as shown in Fig. 5.

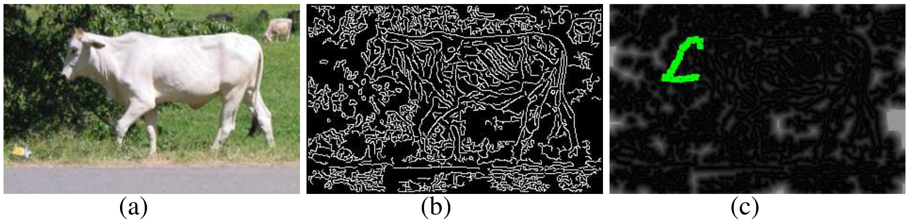

(a) (b) (c)

Figure 5: *(a) Original image of a cow in a cluttered scene. (b) Edgemap of the original image. (c) The distance transform of the edgemap along with an exemplar of the head. Truncated chamfer distance for the exemplar is calculated as the mean of the distance transform values at the exemplar point coordinates.*

The statistic $z_2(\mathbf{D}_i)$ is defined as

$$z_2(\mathbf{D}_i) = \max_k \Pr(\mathbf{D}_i|GMM_k), \tag{9}$$

where GMM_k is the k^{th} GMM representing the texture of the object. To calculate $z_2(\mathbf{D}_i)$, we use a row sum of the image $\mathbf{D}$. The row sum of the image for GMM_k is defined as:

$$RS_k(i,j) = RS_k(i,j-1) + \log(\Pr(\mathbf{D}(i,j)|GMM_k). \tag{10}$$

Fig. 6 shows how the sum of the probabilities of all pixels in one row of $\mathbf{D}_i$ is found efficiently. Summing over all rows of $\mathbf{D}_i$ provides us with a measure of $\Pr(\mathbf{D}_i|GMM_k)$.

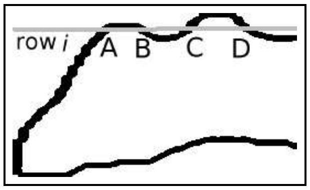


Figure 6: *Row i intersects the region* $\mathbf{D}_i$ *defined by an exemplar for the head of the cow at four points A, B, C and D. The sum of the probabilities of all pixels in row i for the* k^{th} GMM *is given by* $RS_k(B) - RS_k(A) + RS_k(D) - RS_k(C)$.

Despite using row sums, the statistic $z_2(\mathbf{D}_i)$ is computationally more expensive as it requires calculating $\Pr(\mathbf{D}_i|GMM_k)$, for all k. Since truncated chamfer distance ($z_1(\mathbf{D}_i)$) is sufficient to reject a large number of bad poses, $z_2(\mathbf{D}_i)$ is calculated only at the third level of the tree cascade (see Fig. 4).

The putative poses l_i of parts p_i are found by rejecting bad poses by traversing through the tree structure starting from the root node. The likelihoods $\Pr(\mathbf{D}_i|\mathbf{a}_i)$ are found using equation (7). Note that even though the parts do overlap, results indicate that the likelihoods obtained are close to the true likelihoods.

4.3 MAP estimation

A method to compute the MAP estimate of the PS which maximizes equation (6) is required. We use loopy belief propagation (LBP) to find the posterior probability of a part p_i having label l_i. LBP is a message passing algorithm proposed by Pearl [11]. It is a Viterbi-like algorithm for graphical models with loops.

The message that p_i sends to its neighbour p_j at iteration t is a vector of length n_L. The elements of this vector are given by:

$$m_{ij}^t(l_j) = \max_{l_i} \left(V(l_i, l_j) + B_i(l_i) + \sum_{s \in \mathcal{N}_i \backslash p_j} m_{si}^{t-1}(l_i) \right), \tag{11}$$

where $V(l_i, l_j) = \psi(l_i, l_j)$ and $B_i(l_i) = \log\left(\frac{\Pr(\mathbf{D}_i|\mathbf{a}_i)}{\Pr(\mathbf{D}_i|\mathbf{a}_{bg})}\right)$. All messages are initialized to 0, i.e. $m_{ij}^0(l_j) = 0$, for all i and j and are updated in parallel at each iteration.

The belief of a part p_i having label l_i after T iterations is given by

$$b_i(l_i) = B_i(l_i) + \sum_{j \in \mathcal{N}_i} m_{ji}^T(l_i)\ . \tag{12}$$

The termination criterion is that the rate of change of all beliefs falls below a certain threshold. The label l_i^* that maximizes $b_i(l_i)$ is selected for each part. Once the MAP estimate of $\mathbf{l}$ is found, it is further refined by searching over a small affine transformations around l_i^*. This allows us to account for slight variations in visual aspects of the quadruped. The affine transformation which results in the smallest chamfer distance is obtained by gradient descent. We briefly describe the main steps of our approach below.

Algorithm Recognition of objects from test images.

```
1. Given an image $\mathbf{D}$, compute the distance transform image.
2. Compute the row sums of the texture match scores.
3. Find the putative positions of parts by the tree cascade of classifiers [13].
4. Consider only those poses $l_i$ of part $p_i$ for which $\frac{\Pr(\mathbf{D}_i|\mathbf{a}_i)}{\Pr(\mathbf{D}_i|\mathbf{a}_{bg})} > \tau_2$ and define
   an MRF over the parts as described in § 3.
5. Run LBP on the MRF to obtain the most likely poses $l_i^*$ for the PS.
6. Search over affine transformations around $l_i^*$ and refine the pose estimates of the part to result in smaller chamfer score.
```

5 Recognition results

We show recognition results on two types of quadrupeds, cows and horses. For the moment we restrict this to side views of the animals, which is the current state of the art [3, 8]. The animal is said to be correctly recognized in an image if it is found at the right position and the total belief of the object at that position is greater than a certain threshold.

Fig. 7 illustrates the advantage of using a complete graph over a tree structure. Examples of recognition on test images of cows and horses are shown in Fig. 8 and 9. A number of variations of the algorithm were tested to investigate the contribution of various components. These variations were: (i) calculating the likelihoods with and without using the likelihood ratios in equation (3), (ii) using match scores for shape with and without considering texture, and (iii) using a tree structure with using a complete graph for connections between the parts. Fig. 10 illustrates these variations on ROC curves obtained for a total of 500 cow images and 5000 negative examples by varying the threshold on the total belief of the PS. It is evident from the ROC curves that each component that we added is beneficial. The recognition performance is excellent e.g. the equal error rate of the full model is 3%.

6 Summary and Conclusion

We have made a number of extensions to the pictorial structure framework. Methods for efficiently calculating highly discriminative features which include both shape and texture information are presented. The connections between parts forms a complete graph and belief propagation is used for efficient recognition. Results of our method when tested on many cow and horse images convincingly demonstrate its effectiveness.

The model needs to be extended to handle various visual aspects of a quadruped. Another extension is to improve the localization of detection by mapping each point in the exemplar of a part to its nearest edge pixel.

Acknowledgments

We are grateful to Derek Magee for providing the cow videos. 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] S. Agarwal and D. Roth. Learning a sparse representation for object detection. In *ECCV02*, page IV: 113 ff., 2002.

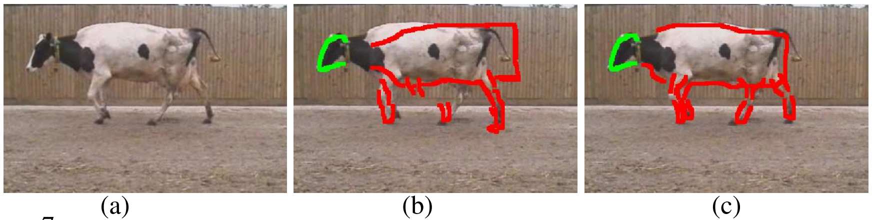


Figure 7: *(a) Original cow image. (b) Result obtained when using a tree structure for the configuration of parts. Both the back legs of our model are mapped to the same leg and the forelegs and the body are not detected properly. (c) Result obtained using a complete graph for the configuration. The connections between the two forelegs result in correct mapping of legs and body (see text).*

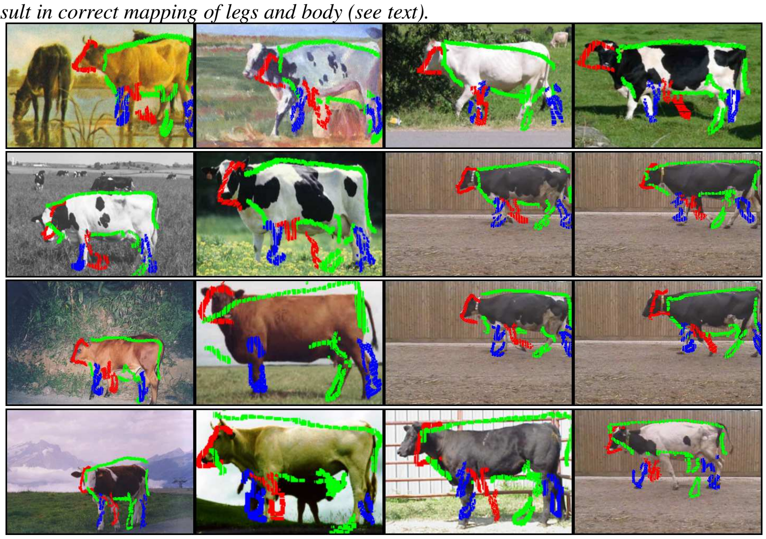


Figure 8: *Each row shows four cow images correctly recognized by our approach, with the outlines of the detected parts overlaid on them. The first image in the fourth row shows an example where a significant variation in visual aspect results in foreshortening of parts. Using affine refinement to minimize chamfer distances of detected parts overcomes this problem. However, it sometimes results in bad localization of parts, e.g. the torso in the last two images of the fourth row undergoes incorrect anisotropic scaling. The problems caused by a weak texture model are evident in the second image of the fourth row where the head of another cow is mistaken for the udder of the cow in front.*

[2] P.M. Baggenstoss. The pdf projection theorem and the class-specific method. *IEEE Transactions on Signal Processing*, 51:672–685, 2003.

[3] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV02*, page II: 109 ff., 2002.

[4] P.F. Felzenszwalb and D.P. Huttenlocher. Efficient matching of pictorial structures. In *CVPR00*, pages II: 66–73, 2000.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR03*, pages II: 264–271, 2003.

[6] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. *TC*, 22:67–92, January 1973.

[7] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Learning layered pictorial structures from video. Technical report, Oxford Brookes University, 2004.

[8] B. Leibe and B. Schiele. Interleaved object categorization and segmentation. In *BMVC03*, pages II: 264–271, 2003.

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

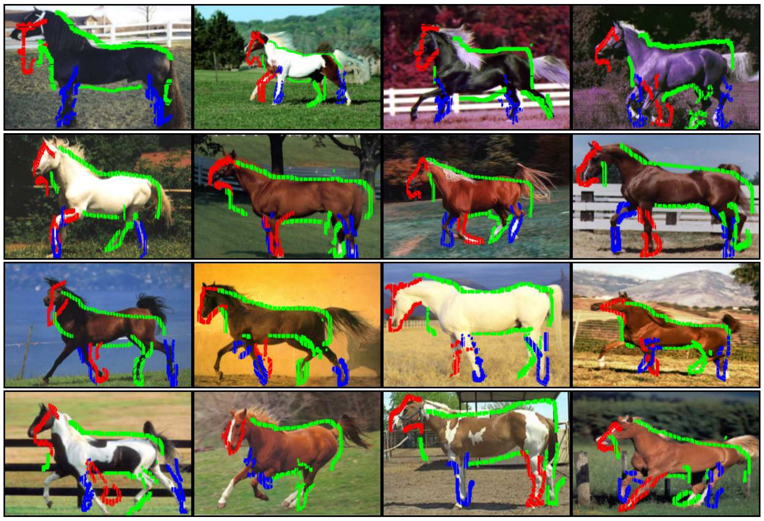
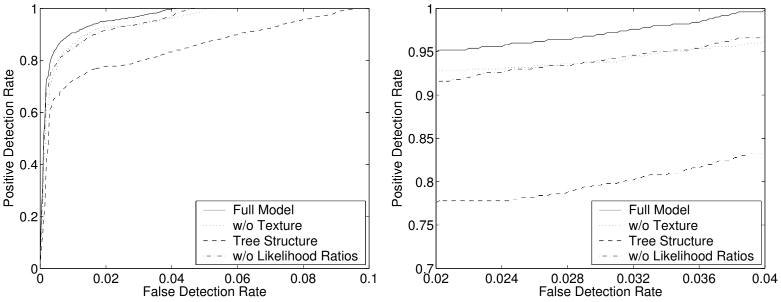


Figure 9: *Each row shows four horse images with the outlines of the detected parts overlaid on them. Certain poses not observed in the training sequences are not correctly identified even though the object is recognized, e.g. first two images of the third and fourth row. Again, using only chamfer distance for affine refinement of poses results in bad localization of torso in the last two images of the fourth row.*



(a) (b)

Figure 10: *Left:* ROC *curves obtained for cow recognition system. Right: Zoomed versions of part of the* ROC *curves. Results indicate that better recognition is obtained when texture information and likelihood ratios are used on complete graphs.*

[10] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV04*, pages I:69–82, 2004.

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

[12] D. Ramanan and D.A. Forsyth. Using temporal coherence to build models of animals. In *ICCV03*, pages 338–345, 2003.

[13] B. Stenger, A. Thayananthan, P.H.S. Torr, and R. Cipolla. Hand pose estimation using heirarchical detection. In *Intl. Workshop on Human-Computer Interaction*, pages 105–116, 2004.

[14] A. Thayananthan, R. Navaratnam, P.H.S. Torr, and R. Cipolla. Likelihood models for template matching using the pdf projection theorem. In *BMVC04*, 2004.

[15] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *CVPR00*, pages II: 101–108, 2000.