Segmentation using Deformable Spatial Priors with Application to Clothing

Basela S. Hasan scbsh@leeds.ac.uk David. C. Hogg d.c.hogg@leeds.ac.uk School of Computing University of Leeds Leeds, UK

Abstract

We present a method for segmenting the parts of multiple instances of a known object category exhibiting large variations in projected shape and colour. The method builds on an existing MRF formulation incorporating a prior shape model and colour distributions for the constituent parts. We propose a novel shape model consisting of a deformable spatial prior probability for the part-label at each pixel. We also make a simple extension to the MRF formulation to deal simultaneously with multiple objects within a global optimisation. Finally, we evaluate the method for the task of segmenting individual items of clothing in images depicting groups of people, and demonstrate improved performance against the state of the art for this task.

1 Introduction

Image segmentation is a longstanding problem in computer vision with many potential applications. The formulation of this problem as maximum a posteriori probability (MAP) inference over a Markov Random Field (MRF) is both elegant and effective. Typically, the MRF is configured to favour contiguous regions with the same labelling, and consistency between the label at each pixel and prior intensity distributions for foreground and background regions. Boykov and Jolly show how to solve this labelling problem efficiently by reformulating as finding a minimum graph-cut [4]. Finding the min-cut for a given graph can be found by solving an equivalent max-flow problem [4]. The min-cut/max-flow technique provides a globally optimal solution.

In an extension to this method, Rother *et al.* [$[\]$] treat the colour distribution for the foreground as a latent property that is optimised along with the labelling in the proposed "GrabCut" method. The background distribution is estimated from a user-defined window surrounding the target foreground object. The problem remains that of finding the MAP solution, but now ranging over the space of possible labellings and foreground colour distributions. The method is iterative and finds a local minimum: first initialise the colour distributions from predetermined regions inside the user-defined window; find the optimal segmentation for these initial distributions; then re-estimate the foreground colour distribution from the labelled pixels. This procedure is repeated until convergence. Vicente *et al.* [$[\]$] propose a non-iterative optimization of segmentation and appearance that is shown to outperfom the iterative approach adopted in [$[\]$].



Figure 1: Clothing segmentation. For the input image in (a), face detections are overlaid in green. The image in (b) shows the results obtained by the proposed method for the different structures assuming everyone is wearing a suit. (c) shows the regions belonging to each person. The parts map in (d) is overlaid in (e) using the colours blue, green, yellow, and brown to represent the shirt, jacket, tie, and face & skin labelled pixels for each person image respectively.

In a further development (OBJ CUT) [1], the foreground object is assumed to be an instance of a known object category, so that prior shape information can be exploited as a top-down influence, whilst remaining within the MRF framework. The idea is to augment the posterior probability to be maximized with an additional term that reflects the conformity of the labelling to a prior shape model - in this case a (layered) pictorial structure $[\Box, \Box]$. Using the OBJ CUT formulation, Rihan et al. [1] proposed a method for face segmentation using a simple elliptical shape model, Bray *et al.* [5] utilize a stick-man like shape prior to simultaneously recover a segmentation and 3D pose estimate. In order to avoid local minimum solutions, Lempitsky *et al.* [**L**] show how to use a branch and bound search to find globally optimal colour distributions and shape parameters for a given segmentation problem. In work that partly motivated our proposed approach, Ramanan [1], Lee et al. [1] and Winn and Jojic [22] use a 2D *spatial prior* on the foreground/background probabilities for each pixel as their shape model. The method of Winn and Jojic [22] goes one step further in allowing smooth deformations to the array of probabilities. In the current paper, we adopt the OBJ CUT MRF formulation and use a spatial prior as the shape model. Our novel contribution is threefold: (1) we use a spatial prior with a category specific deformation function, ranging over multiple labels corresponding to the different parts of an object; (2) we deal jointly with multiple overlapping object instances within the same image, integrating this into a global optimisation within the same MRF framework; (3) we demonstrate an improvement using this approach on the state of the art for the problem of clothing segmentation from images of groups of people.

The use of deformable object models became popular with the development of the socalled Active Appearance Model (AAM) $[\square]$ in which the projected shape of an object instance is represented by the position of a small number of landmark points (e.g. the corners of the mouth and eyes) and the appearance is represented in a deformed texture map with landmarks in canonical position. The allowable variations in the position of landmarks and textures are represented as linear subspaces, learnt from training examples in which landmark points have been identified by hand (but see $[\square]$). We use this general approach to model allowable deformations, not of image textures but of a map of prior probabilities for the labels at each location in relation to an object-centred frame of reference within the image plane.

We evaluate our proposed method on the problem of segmenting items of clothing from

single images as shown in figure 1. The existing methods tackling this problem can be classified into three main approaches: i) naïve clothing-detection methods, ii) segmentation based methods, and iii) model based methods.

In $[\square, \square]$ a face detector is used to hypothesise the clothing region under the face. The rectangle prediction does not fit a person's torso and hence suffers from occlusion and background. Using the average of a set of hand-labelled clothing, Sivic *et al.* $[\square]$ improved the predicted clothing region. The average mask is more like the upper-body clothing, but still does not fit all the clothing shapes. This leads to similar problems as with a rectangular prediction.

Typically, previous work on clothing-segmentation has relied on using a graph-cut technique $[\square]$. Hu *et al.* $[\square]$ automate the clothing segmentation process using foreground (clothing) and background (non-clothing) estimation. Guided by face detection, torso detection, and skin detection, clothing/non-clothing seed pixels are placed on a person image. As the torso detection is based on dominant clothing colour, the segmentation framework fails to cope with the wide variations of clothing styles and colours. In the work of Gallagher *et al.* $[\square]$, the clothing/non-clothing models in the graph cut are initialised using a learnt spatial prior. Under the assumption of having multiple images of the same person wearing the same clothing results are improved and proved to be effective in clothing recognition compared to a naive approach.

To tackle more clothing styles, Chen *et al.* [**G**] learn a number of clothing components such as collars, sleeves, etc. from artist drawings of person images. An And-Or graph representation is built to generate different clothing configurations. Their approach of using colour and edge image cues requires a plain background, non-patterned clothing, and high resolution images.

We start (Section 2) by outlining the MRF formulation of the problem, broadly following the approach of OBJ CUT [1], adapted to deal with multiple object instances. In Section 3 we describe the deformable spatial prior and the way in which this is obtained from training data. In Section 4 we evaluate the approach in the domain of clothing segmentation and conclusions are drawn in section 5.

2 MRF Formulation

We are given an input image $D = \{d_1, d_2, ..., d_N\}$, where d_i is the observed RGB colour at pixel *i*, and *J* object hypotheses $\{o_1, o_2, ..., o_J\}$ for a known category of object. Each object is assumed to have the same *K* different parts $\{p_1, p_2, ..., p_K\}$. A candidate segmentation is an assignment of a single label to each pixel in the input image $L = \{l_1, l_2, ..., l_N\}$. The set of possible labels is the product set of objects $\{o_1, o_2, ..., o_J\}$ and parts $\{p_1, p_2, ..., p_K\}$, augmented by a label representing background denoted by (o_0, p_0) . Thus each label is a pair (o, p). The segmentation task is to find the labelling *L* such that each pixel is assigned a correct label as being part of an object or in the background. The segmentation task is posed as inference in a MRF model. Given prior information about the shape of each object $S = \{s_1, s_2, ..., s_J\}$ and the colour model for each part of each object denoted by $\Theta = \{\theta_{11}, ..., \theta_{1K}, \theta_{21}, ..., \theta_{2k}, ..., \theta_{J1}, ..., \theta_{JK}\}$, we seek the solution \hat{L} which maximises the posterior probability for *L*:

$$\hat{L} = \underset{L}{\operatorname{argmax}}(P(L|D, S, \Theta)) \tag{1}$$



Figure 2: Training models. Suit prior in (a) shows an average spatial prior models learnt for shirt, jacket and tie parts (appear in red, green and blue respectively). The shape model in (b) is built from annotated images with landmarks as illustrated in (c). The points 1 - 6 are manually annotated at left eye, right eye, near-neck left shoulder, near-neck right shoulder, neck centre, and stomach bottom point respectively. The red rectangle represents a fixed prediction of people's clothing. The rectangle parameters are learnt using training clothing data. The corner points numbered 7 - 10 are automatically located with respect to the face detection which appears in green rectangle.

Using Bayes' theorem and the usual spatial Markov assumption, the posterior probability for *L* can be written as:

$$P(L|D, S, \Theta) \propto P(D|L, S, \Theta) \times P(L|S, \Theta)$$
 (2)

$$= \prod_{i} P(d_i|l_i, S, \Theta) \times \prod_{i} P(l_i|S) \times \prod_{i' \in \mathcal{M}} P(l_i|l_{i'})$$
(3)

where \mathcal{N} denotes an 8-neighbourhood. This probability is cast as an energy minimisation problem, where the probabilities above correspond directly to the individual energy terms below:

$$-\log P(L|D, S, \Theta) \propto \alpha \sum_{i} \phi(l_{i}, d_{i}, \Theta) + \beta \sum_{i} \rho(l_{i}, S) + (1 - (\alpha + \beta)) \sum_{i, i' \in \mathscr{N}} \psi(l_{i}, l_{i'}, d_{i, i'})$$
(4)

where the constants α and β weight the importance of each term.

• $\phi(l_i = (o_i, p_k), d_i, \Theta)$ is the data dependent term:

$$\phi(l_i = (o_j, p_k), d_i, \Theta) = \sum_m \left((d_i - \mu_{jk}^m)^t (Cjk^m)^{-1} (d_i - \mu_{jk}^m) \right)$$
(5)

It favours assigning a label *l* to a pixel *i* which has high likelihood under the corresponding Gaussian Mixture Model (GMM) colour model defined by $\Theta = \{\mu_{jk}^m, C_{jk}^m\}$, where j = k = 0 or $1 \le j \le J, 1 \le k \le K$, and $1 \le m \le M_{ij}$ where *m* is the number of GMM components.

- $\rho(l_i, S)$ is the shape prior. It penalises assigning label to pixels that do not fit the target shape. The details of this term are described in the following section.
- $\psi(l_i, l_{i'}, d_{i,i'})$ is the smoothness term. It encourages segmentations with coherent labelling of neighbouring pixels and it takes the form:

$$\psi(l_i, l_{i'}, d_{i,i'}) = |l_i - l_{i'}| . exp(\frac{-||d_i - d_{i'}||^2}{2\sigma^2}) \frac{1}{dist(i, i')}$$
(6)

where dist(i,i') gives the Euclidean distance between *i* and *i'*. The constant σ is a user-defined parameter.



Figure 3: The primary three modes of variance of shape model and the corresponding effect on suit priors. The shape model is shown in yellow as connected vertices which represent the eyes, the body land marks and the rectangle predicting the clothing region. Using the parameters of this simple shape model, deformed versions of the average suit priors are generated. The different deformations are overlaid with the corresponding shape model instances. The spatial probability map for the mean shape is built using Delaunay triangulation in red based on landmarks ((a) middle). Sampled probability information ((a) right) are sampled using triangulation corresponds to ((a) middle).

An important issue for MAP estimation within the MRF framework is finding an efficient inference procedure. Global optimisation for problems involving multiple labels is known to be NP-hard. An approximate solution can be found by the α -expansion algorithm which finds a strong local minimum solution [**G**]. In OBJ CUT the authors demonstrate that the form of the shape model is such that the MAP labelling and shape parameters can be found using an EM procedure. A Branch and Bound procedure is used in [**IS**] to find the globally optimal solution. To maximise the posterior probability for L over the shape parameter space, we use a branch and bound search adapted from [**IS**] and [**B**]. For each portion S of the shape space selected in this search, an approximate lower bound on the energy (equation 4) is obtained by the method in [**IS**], modified to use a spatial prior that gives an upper bound over S on the pixelwise probabilities. Since this upper bound is approximate, the solution is not guaranteed to be globally optimal.

- Step 1: Initialise a label for each pixel $l \in \{(o_0, p_0)\} \cup (\{o_1, \dots, o_J\} \otimes \{p_1, \dots, p_K\})$ given a shape and a reference frame for each object detection.
- Step 2: Build Gaussian Mixture Model (GMM) colour models for background and each part, using currently assigned pixel labels.
- Step 3: Update labels to maximise the posterior given current shape and colour models.
- Step 4: Repeat steps 2–3 until convergence.



Figure 4: Suit hypothesised region using deformable priors. The person image in figure 2, is superimposed with the deformed priors of the first row shown in figure 3 (a) respectively. Note that the suit region is best initialized in (c).

3 Deformable Spatial Prior - $\rho(L,S)$

Our shape prior is based on an object-centred frame of reference and consists of a spatial array of prior probabilities for part labels, coupled with a parameterised deformation function of this array. This is equivalent to the spatial prior in [22] with the addition of a deformation function to adapt the array of prior probabilities to shape variations of the target object category. In our chosen domain, these shape variations arise through out-of-plane rotation of people, differences in body shape, and variations in the shape of clothing. We assume for our experiments that everyone is wearing the same combination of clothing (i.e. a shirt, jacket and tie), so that deformations are sufficient to cope with much of the variation. This is a limiting assumption that needs relaxing in future - for example, by allowing structure variation prior to deformation. Figure 2 (a) shows the canonical spatial prior for three items of clothing worn together (shirt, jacket and tie). At each pixel location, the colour represents the prior probability of shirt (red), jacket (green), tie (blue) and background labels - thus for example, RGB = (0.3, 0.2, 0.2) represents the distribution (0.3, 0.2, 0.2, 0.3). In our proposed method, this fixed spatial prior is allowed to deform according to some latent parameters. These parameters are learnt from a simple shape model according to figure 2 (b). The deformation function is defined in terms of (linear) offsets of a set of landmark points in canonical position:

$$\hat{m} = \bar{m} + \Phi b \tag{7}$$

where \bar{m} is the set of canonical positions and b is a parameter vector that gives different transformations. As in the AAM, we learn the canonical positions \bar{m} and matrix Φ from a training dataset of labelled persons all wearing suits. Example of an image annotated with landmarks is shown in figure 2 (c). Each image has 10 landmarks. Six are manually annotated (left eye, right eye, near-neck left shoulder, near-neck right shoulder, neck centre, and stomach bottom point). Four are fixed relative to the face position and scale (the four corners of the 2D array). Given the vector of landmark positions for the set of training instances, the mean of these vectors becomes the canonical set of positions \bar{m} and principal axes from principal component analysis become Φ (see [\Box] for details). Figure 3 shows the effect of varying separately each of the first three shape parameters in turn between ± 2 standard deviations from the mean value. Landmarks are connected by yellow sticks.

The aim at this stage is to generate a deformable spatial prior which represents allowable shape instances. Figure 3 shows examples of new prior models superimposed with the corresponding shape instances. As in the AAM, we use a Delaunay triangulation to deform the canonical probability map. Unlike the AAM, the 'texture map' (in our case an array of

Method	Balanced Accuracy			
	Shirt	Jacket	Tie	Clothing
Segmentation using fixed priors	67%	81%	75%	82%
Segmentation using deformed priors	72%	83%	80%	83%

Table 1: The Balanced-Accuracy of segmentations compared to ground-truth.



Figure 5: For each input image in (a), Results using fixed prior in (b) are improved in (c) using the deformable prior.

prior probabilities) does not vary, although allowing such variation would be an interesting direction for future work. The conditional probability for a label l at location i given a shape model s_j is given by:

$$P(l_i|s_j) = Q(l_i|T_j(x_i) + \bar{m} + \Phi s_j)$$
(8)

where $T_j(x_i)$ is the transformation giving the scale and position of the instance. For our purpose, scaling and translation are defined by face dimensions and location provided by a face detector. In a normalized-scale coordinate space, $Q(l_i|T_j(x_i) + \bar{m} + \Phi s_j)$ is the retrieved probability information for each location *i* in the deformed version based on the location map defined by $\bar{m} + \Phi s_j$. In the middle image in figure 3 (a) , using triangulation (red lines), we learn the probability vector for the mean shape. For any shape instance defined by the parameter *s*, we warp the landmarks of the shape to match the mean shape and build the location correspondence to sample the probability information $P(l_i|s_j)$ from the learnt probability map between the corresponding triangles (e.g. figure 3 (a) last image). The main benefit of allowing fixed spatial prior to deform is to better initialize structure regions as presented in the example image shown in figure 4 and that would improve the accuracy of the segmentation results as demonstrated in the next section.

4 Experimental Results

We evaluated the proposed method for the task of segmenting all people's upper-body items of clothing from images in which people are typically overlapping one another (e.g. Figure 6). The aim is to assign separate labels to the shirt, jacket and tie for each individual whose face is detected. We collected a dataset of 117 photos depicting 200 instances of someone wearing a suit (42 people in total). Using the MRF formulation in equation 2 with labels



Figure 6: Clothing-part segmentation results using deformed priors. The images show that the proposed segmentation framework is able to segment people's clothing parts in different poses.

for shirt, jacket, tie, face and skin of each individual plus background, our method simultaneously segments the clothing parts of all people in the photo, dealing with occlusion when people are close together. We have hand-labelled masks for each person to evaluate the accuracy of our segmentation result and for clothing matching evaluation, clothing identities ground truth are assigned to only unambiguous cases of people's clothing. We used 10-fold cross-validation, in each iteration: models are trained on 180 examples and tested on 20 examples. We learn a canonical spatial prior of size 100×100 . The deformation is constrained to be within ± 2 standard deviations, accounting for around 80% of the total variance.

The proposed segmentation framework outperforms the baseline. Images in figure 5, show a comparison with the baseline. Figure 6 shows more results for cases of multiple people and occlusions. The clothing prediction/segmentation tells us what pixels in the image are considered as clothing. We use *Balanced-Accuracy* [21] for the correct labelling compared to the ground truth segmentations. The results reported in Table 1 are averages over the validation sets.

To further evaluate the utility of the segmentation obtained, we consider the task of reidentifying individuals by matching their clothing between photographs. We compare the use of a composite descriptor containing RGB histograms for each item of clothing obtained from our segmentations, with a baseline method in which a single histogram is computed over the whole upper-body. Histograms are compared using the χ^2 distance measure. Matching people's clothing is posed as a "same vs. different" classification task. The results as an ROC and PR curves shown in figure 7 are combined over all the test sets. These demonstrate a significant improvement for using the spatial information from the segmented clothing parts over the using whole clothing region.



Figure 7: Performance evaluation for the proposed clothing parts segmentation. We compared to the baseline which teats people's clothing a whole. ROC and PR curves show that spatial information about people's clothing is of significant importance for the clothing recognition.

5 Conclusions

We have presented a hybrid novel framework to deform fixed prior models of different parts based on the parameters of a simple shape model. We demonstrated the efficiency of our approach on clothing part segmentation. We have also proposed a generalized fully automatic version of Grabcut which simultaneously segments the clothing part of all detected people in a photo, dealing with occlusion when people are close together. Experimental results show that: (i) segmentation accuracy results can be improved using deformable spatial priors and (ii) peoples' clothing is a rich source of multiple cues which have significant importance in clothing recognition even with simple colour descriptors. For the clothing recognition problem, we intend to use better clothing descriptors capturing texture and patterns in addition to colour.

References

- D. Anguelov, K. Lee, S. B. Gokturk, and B. Sumengen. Contextual Identity Recognition in Personal Photo Albums. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pages 1–7, 2007.
- [2] Y. Boykov and V. Kolmogorov. An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004. ISSN 0162-8828.
- [3] Y. Boykov, O. Veksler, and R. Zabih. Fast Approximate Energy Minimization via Graph Cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, I:377– 384, 1999.
- [4] Y. Y. Boykov and M. Jolly. Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images. *Computer Vision, IEEE International Conference on*, 1:105–112, 2001.
- [5] M. Bray, P. Kohli, and P. H. S. Torr. PoseCut: Simultaneous Segmentation and 3D Pose Estimation of Humans using Dynamic Graph-cuts. In ECCV (2), pages 642–655, 2006.

- [6] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu. Composite Templates for Cloth Modeling and Sketching. In CVPR '06: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 943–950, 2006.
- [7] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:681–685, 2001. ISSN 0162-8828.
- [8] D. Cremers, F. R. Schmidt, and F. Barthel. Shape priors in variational image segmentation: Convexity, Lipschitz continuity and globally optimal solutions. In *CVPR*, 2008.
- [9] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient Matching of Pictorial Structures. Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, 2: 2066, 2000. ISSN 1063-6919.
- [10] A. C. Gallagher and T. Chen. Clothing Cosegmentation for Recognizing People. In Proceedings of CVPR, pages 1–8. IEEE Computer Society, 2008.
- [11] Z. Hu, H. Yan, and X. Lin. Clothing Segmentation Using Foreground and Background Estimation Based on the Constrained Delaunay Triangulation. *Pattern Recogn.*, 41(5): 1581–1592, 2008.
- [12] M. P. Kumar, P. H. S. Torr, and A. Zisserman. Learning Layered Pictorial Structures from Video. In *Proceedings of the Indian Conference on Computer Vision, Graphics* and Image Processing, pages 158–163, 2004.
- [13] M. P. Kumar, P. H. S. Torr, and A. Zisserman. OBJ CUT. In CVPR (1), pages 18–25, 2005.
- [14] K.C. Lee, D. Anguelov, B. Sumengen, and S.B. Gokturk. Markov Random Field Models for Hair and Face Segmentation. In *Proceedings of IEEE Conference On Automatic Face and Gesture Recognition*, pages 1–6, 2008.
- [15] V. Lempitsky, A. Blake, and C. Rother. Image Segmentation by Branch-and-Mincut. In Computer Vision-Eccv 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV, pages 15–19, 2008.
- [16] V. S. Petrovic, T. F. Cootes, A. M. Mills, C. J. Twining, and C. J. Taylor. Automated Analysis of Deformable Structure in Groups of Images. pages 1060–1069, 2007.
- [17] D. Ramanan. Using Segmentation to Verify Object Hypotheses. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:1–8, 2007.
- [18] J. Rihan, P. Kohli, and P. H. S. Torr. OBJCUT for Face Detection. In *ICVGIP*, pages 576–584, 2006.
- [19] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": Interactive Foreground Extraction using Iterated Graph Cuts. ACM Transactions on Graphics, 23:309–314, 2004.
- [20] J. Sivic, C. L. Zitnick, and R. Szeliski. Finding People in Repeated Shots of the Same Scene. In *Proceedings of the British Machine Vision Conference*, pages III:909–918, 2006.

- [21] M. Sokolova, N. Japkowicz, and S. Szpakowicz. Beyond Accuracy, F-Score and ROC: A Family of Discriminant Measures for Performance Evaluation. AI 2006: Advances in Artificial Intelligence, pages 1015–1021, 2006.
- [22] Y. Song and T. Leung. Context-Aided Human Recognition: Clustering. In *Proceedings* of ECCV06, pages III: 382–395, 2006.
- [23] S. Vicente, V. Kolmogorov, and C. Rother. Joint optimization of segmentation and appearance models. pages 755–762, 2009.
- [24] J. Winn and N. Jojic. LOCUS: Learning Object Classes with Unsupervised Segmentation. *Computer Vision, IEEE International Conference on*, 1:756–763, 2005. ISSN 1550-5499.