Modelling composite shapes by Gibbs Random Fields

Boris Flach
Czech Technical University
flachbor@cmp.felk.cvut.cz

Dmitrij Schlesinger
Dresden University of Technology
dmytro.shlezinger@tu-dresden.de

Abstract

We analyse the potential of Gibbs Random Fields for shape prior modelling. We show that the expressive power of second order GRFs is already sufficient to express spatial relations between shape parts and simple shapes simultaneously. This allows to model and recognise complex shapes as spatial compositions of simpler parts.

1. Introduction

Motivation and goals. Recognition of shape characteristics is one of the major aspects of visual information processing. Together with colour, motion and depth processing it forms the main pathways in the visual cortex.

Experiments in cognitive science show in a quite impressive way, that humans recognise complex shapes by decomposition into simpler parts and interpreting the former as coherent spatial compositions of these parts [5]. Corresponding guiding principles for the decomposition where identified from these experiments as well as from research in computer vision (see e.g. [7]). The formulation of these principles relies however on the assumption that the objects are already segmented and thus concepts like convexity and curvature can be applied.

From the point of view of computer vision it is desirable to use shape processing and modelling in the early stages of visual processing. This allows to control e.g. segmentation directly by prior assumptions or by feedback from higher processing layers. This leads to the question whether composite shape models can be represented and learned in a topologically fully distributed way. The aim of the presented work is to study this question for probabilistic graphical models.

Related work. All mathematically well principled shape models for early vision can be roughly divided into the following two groups.

Global models treat shapes as a whole. Prominent representatives are variational models and level set methods in particular. A shape is described up to its pose by means of a level set function defined on the image domain. Cremers *et al*. have shown in [1] how to extend these models for scene segmentation. Recently we have shown how to use level set methods in conjunction with MRFs [2]. Global shape models are well suited e.g. for segmentation and tracking if the number of objects is known in advance and a good initial pose estimation is provided.

Semi global models consider shape characteristics in local neighbourhoods and go back to the ideas of G. Hinton on "product of experts" as well as of Roth and Black on "fields of experts" (see [4, 9] and citations therein). Mathematically these models are higher order GRFs of a certain type – additional auxiliary variables are used to express mixtures of local shape characteristics in usually overlapping neighbourhoods. Marginalisation over these auxiliary variables results in GRFs of higher order. The work of Kohli, Torr *et al*. [8, 6] demonstrates how to introduce such higher order Gibbs potentials directly and to use them for segmentation in hierarchical Conditional Random Fields (CRF). However, it is not clear how to learn the graphical structure for such models.

Contributions. We will show that Gibbs Random Fields of second order have already sufficient expressive power to model complex shapes as coherent spatial compositions of simpler parts. Obviously, these models have to have a significantly more complex graphical structure than just simple lattices. Moreover, the graphical structure itself becomes a parameter which has to be learnt together with the Gibbs potentials for each considered shape class.

From the application point of view these models have advantages especially in the context of scenes with an unknown number of similar objects (i.e. all objects are instances of a single shape class). Moreover, such models can be easily combined for scenes with instances of different shape classes.

2. The shape model

Probability distribution. We begin with the description of the prior part of our shape model. Let $D \subset \mathbb{Z}^2$ be a finite set of nodes $t \in D$, where each node corresponds to an image pixel. Let $A \subset \mathbb{Z}^2$ be a set of vectors used to define a neighbourhood structure on the set of nodes, i.e. a graph:

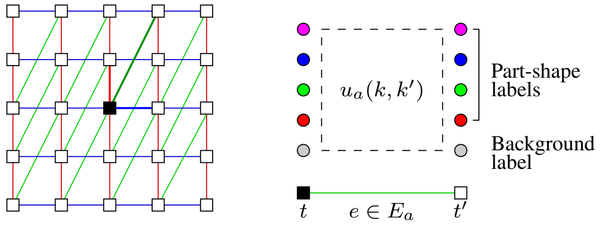

Figure 1. Left: example of a translational invariant graphical structure. Equivalence classes of edges E_a are coloured by different colours. The set A is represented by bold edges outgoing from the central node. Right: Gibbs potentials for an edge from E_a.

two nodes t and t' are connected by an edge if $t' - t = a \in A$. To avoid double edges we require $-A \cap A = 0$ (we use unary potentials as well). The resulting graph is obviously translational invariant and the elements of $a \in A$ define subsets $E_a \subset E$ of equivalent edges, where $e = (t, t') \in E_a$ if $t' - t = a$. An simple example is shown in Fig. 1.

Given a class of composite shapes, we denote the set of its parts enlarged by an extra element for the background by K. A shape-part labelling $y \colon D \to K$ is a mapping, that assigns either a shape-part label or the background label $y_t \in K$ to each node $t \in D$. A function $u_a : K \times K \to \mathbb{R}$, is defined for each difference vector $a \in A$. Its values $u_a(k, k')$ are called Gibbs potentials. A corresponding probability distribution is defined over the set of shape-part labellings as follows

$$p(y) = \frac{1}{Z(u)} \exp \sum_{a \in A} \sum_{tt' \in E_a} u_a\big(y_t, y_{t'}\big), \qquad (1)$$

where Z denotes the partition sum (we omit the unary terms for better readability). This p.d. is homogeneous (up to boundary effects) – all edges in an equivalence class E_a have the same potentials. Moreover, it can be shown that the homogeneous parametrisation is unique up to additive constants.

The appearance model is assumed to be a "simple" conditional independent model. The probability to observe an image $x \colon D \to C$ (C is some colour space) given a shape-part labelling y is

$$p(x \mid y) = \prod_{t \in D} p\big(x_t \mid y_t\big). \qquad (2)$$

In the light of the current popularity of CRFs it might well be asked, why we decided to favour a GRF here. Both variants are identical with respect to inference. Differences occur for learning. We can imagine that shape-part labellings can be used as latent variable layers for complex object segmentation models. Recently, empirical risk minimisation learning has been proposed for structured SVM models with latent variables [12]. This shows that learning of graphical models with latent variables is possible for both variants – GRFs and CRFs. However, since we want to study the expressive power of the model in its pure form, we need a prior p.d. and moreover, we want to be able to learn such models fully unsupervised, which is possible for GRFs but not for CRFs.

The inference task. Informally, the inference task can be understood as follows. Given an observation (i.e. an image), it is necessary to assign values to all hidden variables. We pose the segmentation task as a Bayesian decision task. Let y' be the true (but unknown) segmentation and $C(y, y')$ be a loss function, that assigns a penalty for each possible decision y. The task of Bayesian decision is to minimise the expected loss

$$R(y; x) = \sum_{y'} p(y' \mid x) C(y, y') \to \min_{y}. \qquad (3)$$

We use the number of misclassified pixels

$$C(y, y') = \sum_{t} \mathbb{I}\{y_t \neq y'_t\} \qquad (4)$$

as the loss function. It leads to the max-marginal decision

$$y^*_t = \max_k p\big(y_t = k \mid x\big) \quad \forall t \in D. \qquad (5)$$

Hence, it is necessary to calculate the marginal posterior probabilities for each node $t \in D$ and label $k \in K$. This task is infeasible for GRFs. Several approximation techniques based e.g. on belief propagation or variational methods have been proposed for it (see e.g. [11] for an overview). Unfortunately none of them guarantees convergence to the exact values of the sought-after marginal probabilities. To our knowledge, the only scheme which does it is sampling, which is however known to be slow [10].

Estimation of Gibbs potentials. The learning task comprises to estimate the unknown model parameters given a learning sample. We assume that the latter is a random realisation of i.i.d. random variables, so that the Maximum Likelihood estimator is applicable.

The following situations are distinguished depending on the format of the learning data. If the elements of the sample have the format (x, y) then the learning is called *supervised*. If instead, they consist of images only then the learning is called fully unsupervised. To cope with variants in-between as well, i.e. partial labellings y_V, we consider the elements of the training sample to be events of the type $\mathcal{B} = (x, y_V) = \{(x, y) \mid y_{|V} = y_V\}$.

We start with the learning of unknown potentials u. For simplicity we consider the case when only one event $\mathcal{B}$ is given as the training sample. According to the Maximum Likelihood principle, the task is

$$p(\mathcal{B}; u) = \sum_{y \in \mathcal{B}} p(y) p(x|y) \to \max_{u}. \qquad (6)$$

Taking the logarithm and substituting the model (1), (2) gives

$$L(u) = \log \sum_{y \in \mathcal{B}} \exp\Big[\sum_{a \in A} \sum_{tt' \in E_a} u_a\big(y_t, y_{t'}\big)\Big] p(x|y) - \\ \log\big(Z(u)\big) \to \max_u . \quad (7)$$

It is easy to see, that the derivative with respect to the potentials is a difference of expectations of certain random variables $n_a(k,k';y)$ with respect to the posterior and prior p.d.

$$\partial L / \partial u_a(k,k') = \mathbb{E}_{p(y|\mathcal{B};u)}\big[n_a(k,k';y)\big] - \\ \mathbb{E}_{p(y;u)}\big[n_a(k,k';y)\big]. \quad (8)$$

The random variables $n_a(k,k';y)$ are defined by

$$n_a(k,k';y) = \sum_{tt' \in E_a} \mathbb{I}\big\{y_t{=}k, y_{t'}{=}k'\big\} \quad (9)$$

and represent co-occurrences for label pairs (k,k') along the edges in E_a for a labelling y.

The exact calculation of the expectations in (8) is not feasible. Therefore, we propose to use a stochastic gradient ascent to maximise (7). The learning algorithm is an iteration of the following steps:

1. Sample $\tilde{y}$ and y according to the current a-posteriori probability $p(y|\mathcal{B};u)$ and a-priori probability $p(y;u)$ respectively.
2. Compute $n_a(k,k';\tilde{y})$ and $n_a(k,k';y)$ by (9) for each $a \in A$, $k,k' \in K$.
3. Replace the expectations in (8) by their realisations and calculate new potentials u.

For the sake of completeness we mention that the learning of the appearance models $p(c|k)$ can be done in a very similar manner. It is even simpler from the computational point of view because the normalising constant Z does not depend on these probabilities. Therefore only a-posteriori sampled labellings are needed to perform the corresponding stochastic gradient step.

Estimation of the interaction structure. A very important question not discussed so far is the optimal choice of the neighbourhood structure A. Unfortunately, no well founded answer to this question is known at present. One option is to use an abundant set of interaction edges, e.g. to assume that the set A consists of all vectors $A = \{a \in \mathbb{Z}^2 \mid |a_1|, |a_2| \leqslant d\}$ within a certain range. Despite of the computational complexity this would lead to models with high VC dimension and possibly – as a result – to weak discrimination. Therefore we use a greedy search for the interaction edges proposed by Zalesny in the context of texture modelling [13, 3]. Starting from the set $A = \{0\}$, i.e. a model with unary potentials, new edges are iteratively chosen and included into A as follows. For the current set A the optimal set of potentials $u_A^* \in U_A$ is determined as described in the previous subsection. Here U_A denotes the subspace of potentials on the edges in A (assuming that the Gibbs potentials are zero on all other edges). If a bigger neighbourhood A' is considered, then clearly, the gradient of the (log) likelihood with respect to $u_{A'}$ in the point u_A^* will be orthogonal to the subspace U_A. The proposal is to include the vector $a' \in A'$ with the largest gradient component

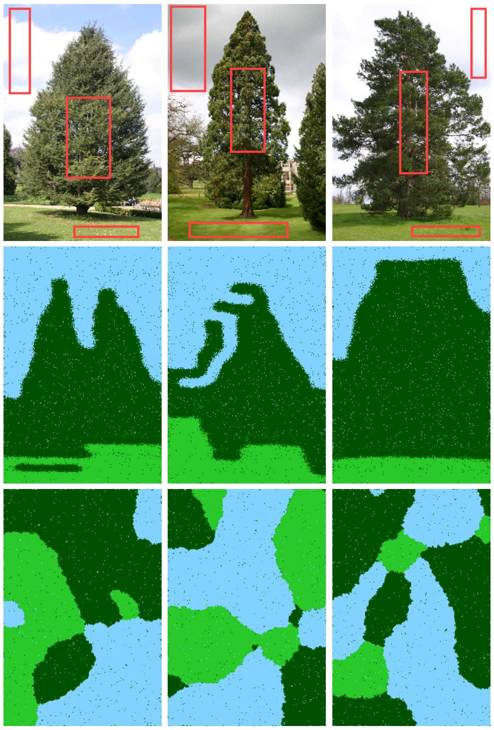

Figure 2. Modelling spatial relations between segments. The first row shows input images and regions with fixed segmentation. The middle and bottom row show labellings generated by the learned a-priori models (segment labels are coded by colour): the images in the middle row were generated by the model with full neighbourhood, whereas the images in the bottom row were generated by the baseline model.

$$a' = \operatorname*{arg\,max}_{a \in A' \setminus A} \sum_{k,k'} \big[n_a(k,k';\mathcal{B},u) - n_a(k,k';u)\big]^2. \quad (10)$$

3. Experiments

Modelling spatial relations between segments. The first experiment investigates the ability of the model (1), (2) to reflect spatial relations between segments, i.e. scene parts, which are too large for capturing their shape by a neighbourhood structure of reasonable size. We used the three images shown in the first row of Fig. 2 as training examples. Each scene should be segmented into three segments:

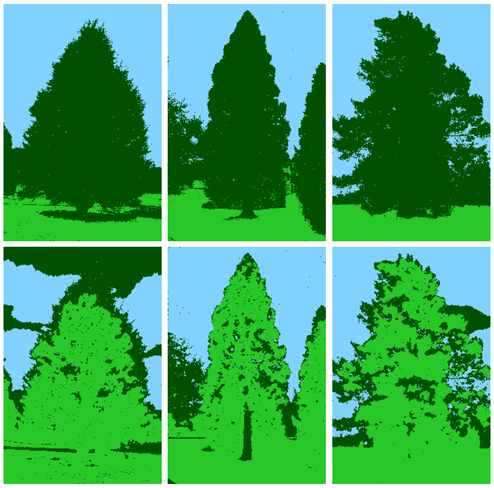

Figure 3. Segmentation results obtained after fully unsupervised learning of the appearance part of the model. Upper row – model with full neighbourhood, bottom row – baseline model.

$K = \{sky, trees, grass\}$. The appearance models $p(c|k)$ for the segments were assumed as mixtures of multivariate Gaussians (four per segment). A model with "full" neighbourhood structure – all vectors $\{a \in \mathbb{Z}^2 \mid |a_1|, |a_2| \le d\}$ with $d = 20$ was used in this experiment. A "simple" but anisotropic Potts model on the 8-neighbourhood was chosen as a baseline for comparison.

Semi-supervised learning was applied by fixing the segment labels in the rectangular areas shown by red rectangles during learning. Both the a-priori models (the potentials and the direction specific Potts parameters for the baseline model) and the appearance models (mixture weights, mean values and covariance matrices) were learned.

The difference of the models can be clearly seen by observing labellings generated a-priori by the learned models, i.e. without input images. Some of them are shown in the second and third row for the model with complex neighbourhood structure and the baseline model respectively. It can be seen, that the spatial relations between segments (like e.g. "above", "below" etc.) were correctly captured by the complex model, whereas it is clearly not the case for the Potts model.

The consequences can be clearly seen from the following experiment. We fixed the prior models obtained in the previous experiment for both variants and learned the parameters of the Gaussian mixtures completely unsupervised. Fig. 3 shows labellings (i.e. segmentations) sampled at the end of the learning process by the corresponding a-posteriori probability distributions for the complex a-priori model and the Potts a-priori model in the first and the second row respectively. The advantages of the complex model are clearly seen. These results can be explained as follows. There are twelve Gaussians in total to interpret the given images. For the learning process it is "hard to decide" which of the Gaussians belongs to which segment. Using the compactness assumption only, is obviously not enough to separate segments from each other. If the complex model is used instead, the learning process starts to generate labellings according to the a-priori probability distribution, i.e. labellings which reflect the correct spatial relations between the segments. This forces the unsupervised learning of the appearance models into the right direction.

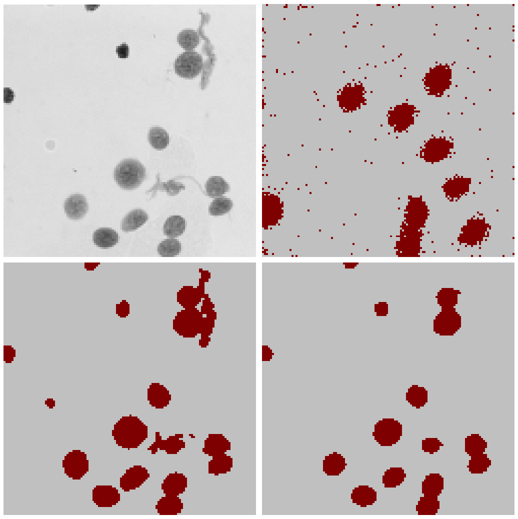

Figure 4. Modelling and segmentation of simple shapes. Upper left – input image, upper right – a labelling generated a-priory by the learned complex model. Final segmentations are shown in the bottom row: left – baseline model, right – complex model.

Modelling simple shapes. This group of experiments demonstrates the ability of the model to represent simple shapes as well as to perform shape driven segmentation. This experiment is prototypical e.g. for a class of image recognition tasks in biomedical research. Fig. 4 (upper left) shows a microscope image of liver cells with stained DNA. Thus, only the cell nuclei are visible. The task is to segment the image into two segments – "cells" (nearly circular shaped) and "background" (the rest including artefacts). Hence, two labels are used. The "full" neighbourhood structure with $d = 12$ was used (it approximately corresponds to the mean cell diameter). Again, we used a baseline model for comparison – a GRF with 4-neighbourhood and free potentials. The appearances for grey-values were assumed to be Gaussian mixtures (two per segment) in both models.

First, semi-supervised learning was performed (like in the previous experiment with trees) in order to learn the prior distributions for labellings as well as the appearances for both, the complex and the baseline model. A labelling generated a-priori by the learned complex model is shown in Fig. 4 (upper right). The final segmentations accord-

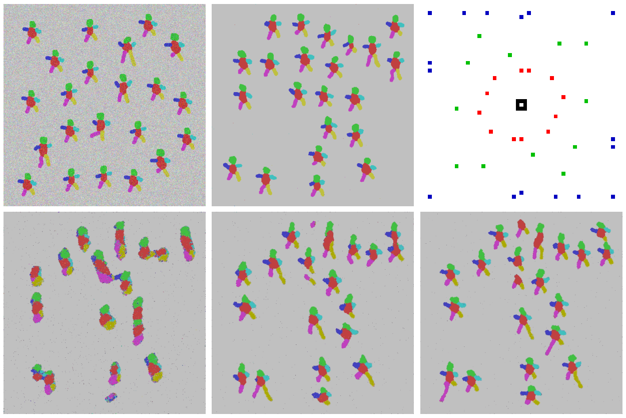

Figure 5. Composite shape modelling. Upper row from left: input image, labelling generated a-priori by the learned model, estimated interaction structure. Bottom row: labellings generated by models during learning.

ing to the max-marginal decision (5) are shown in the bottom row of the same figure. The differences are clearly seen. The shape prior modelled in the complex model led to the correct segmentation – the artefacts were segmented as background, whereas the baseline model produces a wrong segmentation because neither the appearance nor a simple ”compactness“ assumption allow to differentiate between cells and artefacts.

Modelling composite shapes. The previous experiments have shown that second order GRFs can model both, spatial relations between segments and simple shapes. Now we are going to demonstrate the capability of the model to capture both properties *simultaneously*. This opens the possibility to represent complex shapes as spatial compositions of simpler parts. To demonstrate this, we use an artificial example shown in Fig. 5 (upper left). It was produced manually and corrupted by Gaussian noise. Accordingly, the model was defined as follows. The label set K consists of seven labels, each one corresponding to a part of the modelled shape (as well as one for the background). The appearance models $p(c|k)$ for the labels are Gaussians with known parameters. In this experiment we applied the estimation of the interaction structure as described in section 2.

Fig. 5 (upper row, center) shows a labelling generated by the learned prior model. It is clearly seen that both, spatial relation between object parts and part shapes are captured correctly.

The bottom row of Fig. 5 displays labellings generated during the process of structure learning at time moments, when the current interaction structure learned so far was not yet capable to capture all needed properties. As it can be seen, the model was able to learn spatial relations between the segments more or less correctly even for a small numbers of edges (5 edges – bottom left). More relations are learned as the number of edges grows (bottom middle and right). Finally, 20 difference vectors were necessary to capture all relations (out of 1200 possible for the maximal range of $d = 24$).

Fig. 5 (upper right) shows the estimated neighbourhood structure. The endpoints of all edges from central pixel are marked by colours (the image is magnified for better visibility). A certain structure can be seen in this image. The 8-neighbourhood edges (black) reflect compactness and adjacency relations of the object parts. The learned potentials on these edges represent strong label co-occurrences. Most of the other vectors are responsible for the shapes of the parts. The potentials on the red edges express characteristic breadths, and the potentials on the green edges – characteristic lengths of the parts. The potentials on these edges mainly represent anti-correlations, forcing label values to change along certain directions. The blue pixels in the figure reflect relative positions of object parts.

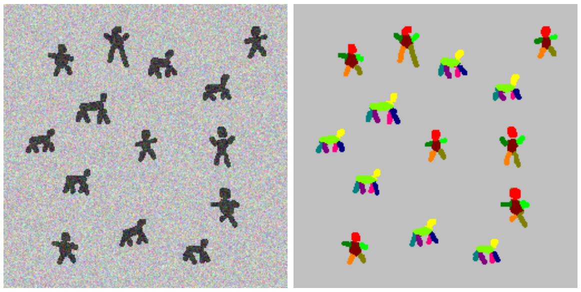

Figure 6. Shape segmentation and classification. Left – input image, right – segmentation (part-labels are encoded by colours).

Composite shape recognition. The final experiment demonstrates possibilities to combine composite shape models. The aim is to obtain a joint model which can be used for detection, segmentation and classification of objects in scenes populated by instances of different shape classes like e.g. the example in Fig. 6. As the appearance model can be learned in a fully unsupervised way, the most important question is, how to combine the prior models. We propose a method based on the following observation. Example images (or segmentations) are not needed for learning the model if the a-posteriori statistics $\bar{\Phi}_a(k,k') = \mathbb{E}_{p(y|\mathcal{B},u)}\big[n_a(k,k';y)\big]$ (see (8)) are known for all difference vectors $a \in A$ and label pairs (k,k'). The gradient of the likelihood reads then

$$\partial L/\partial u_a(k,k') = \bar{\Phi}_a(k,k') - \mathbb{E}_{p(y;u)}\big[n_a(k,k';y)\big]. \quad (11)$$

The parameter estimation scheme for the joint model consists therefore of two stages: (i) compose the desired a-posteriori statistics $\bar{\Phi}_a(k,k')$ for the joint model and (ii) learn the model according to (11) so that it reproduces this statistics. We consider the first stage in more detail for a simple example – two shapes as shown in Fig. 6. Let us

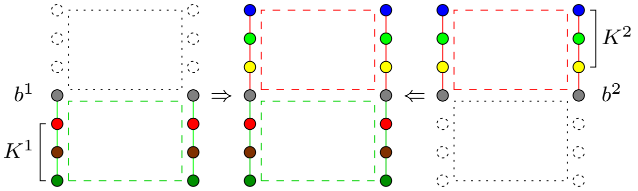

Figure 7. Estimation of the a-posteriori statistics for the joint model. Left and right: statistics for shape models. Middle: the joint model – statistics marked green and red are inherited from the components. Others are set to a small constant.

assume that the a-posteriori statistics are known for both shape models after their learning. The composition of the needed a-posteriori statistics for the joint model is illustrated in Fig. 7. The label set of the composed model is $K^1 \cup K^2 \cup b$, where K^1 and K^2 denote the label sets corresponding to the shape parts for the first and for the second shape type respectively, b is the common background label. The a-posteriori statistics for the joint model is obtained as a weighted mixture of the two original ones (extended in a suitable way to the new label set) and an additional uniformly distributed component. The latter is added in order to avoid zero probabilities (which would lead to obvious technical problems for the Gibbs Sampler). Given these statistics the joint model is learned according to (11).

For the experiment in Fig. 6 two composite shape models were learned separately. The test image in Fig. 6 (left) is a collage of both shape types. Note that the appearance of all shape parts is identical, so they are not distinguishable without the prior shape model. Fig. 6 (right) shows the final segmentation. It is seen that all objects were correctly segmented and recognised – although both composite shape classes share similarly shaped parts – they were not confused.

4. Conclusions

The notation of shape is often understood as an object property of global nature. We followed a different direction by modelling shapes in a distributed way. We have demonstrated that the expressive power of second order GRFs allows to model spatial relations of segments, simple shapes and moreover, both aspects *simultaneously* i.e. composite shapes which are understood as coherent spatial compositions of simpler shape parts.

We have shown that complex shapes can be recognized even in the situation, when their parts are not distinguishable by appearance. However, in our learning experiments we used training images, where they are distinguishable. Thus, an important question is, whether it is possible to perform unsupervised decomposition of complex shapes into simpler parts during the learning phase, i.e. to learn shape models from images, where the desired spatial relations between shape parts are not explicitly present . Another important issue is the learning of the interaction structure. It would be very useful to have a well grounded approach for this.

Acknowledgements. We would like to thank Georgy Gimel'farb (University of Auckland) for the fruitful and instructive discussions which have been particularly valuable with regard to structure learning.

One of us (B.F.) was supported by Czech Ministry of Education project MSM6840770012. D.S. was supported by Deutsche Forschungsgemeinschaft, Grant FL307/2-1. Both authors were partially supported by Grant NZL 08/006 of the Federal Ministry of Education and Research of Germany and the Royal Society of New Zealand.

References

[1] D. Cremers, N. Sochen, and C. Schnörr. A multiphase dynamic labeling model for variational recognition-driven image segmentation. *IJCV*, 66(1):67–81, January 2006. 2177

[2] B. Flach and D. Schlesinger. Combining shape priors and MRF-segmentation. In N. da Vitoria Lobo et al., editor, *S+SSPR 2008*, pages 177–186. Springer, 2008. 2177

[3] G. L. Gimel'farb. Texture modeling by multiple pairwise pixel interactions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(11):1110–1114, 1996. 2179

[4] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. 2177

[5] D. D. Hoffman. *Visual Intelligence: How We Create What We See*. W. W. Norton & Company, 2000. 2177

[6] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical CRFs for object class image segmentation. In *Proceedings IEEE 12. International Conference on Computer Vision*, 2009. 2177

[7] H. Liu, W. Liu, and L. J. Latecki. Convex shape decomposition. In *CVPR*, pages 97–104, 2010. 2177

[8] S. Ramalingam, P. Kohli, K. Alahari, and P. Torr. Exact inference in multi-label CRFs with higher order cliques. In *CVPR 2008*, pages 1–8, June 2008. 2177

[9] S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision*, 82(2):205–229, 2009. 2177

[10] A. D. Sokal. Monte Carlo methods in statistical mechanics: Foundations and new algorithms. Lectures notes, 1989. 2178

[11] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. 2178

[12] C.-N. J. Yu and T. Joachims. Learning structural SVMs with latent variables. In *International Conference on Machine Learning (ICML)*, 2009. 2178

[13] A. Zalesny and L. V. Gool. Multiview texture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2001*, pages 615–622. IEEE Computer Society, 2001. 2179