UDC 004.93'1

B. Flach, D. Schlesinger

Modelling Distributed Shape Priors by Gibbs Random Fields of Second Order

Проведен анализ потенциалов гиббсовских случайных полей для моделирования априорных вероятностных характеристик формы объекта. Показано, что при помощи этих полей второго порядка можно описывать как простые формы, так и пространственные отношения между ними, что позволяет распознавать сложные формы как сочетания более простых.

The potentials of Gibbs random fields are analyzed for the shape prior modeling. It is shown that the expressive power of second order GRFs is already sufficient to express simple shapes and spatial relations between them simultaneously. This allows to recognize complex shapes as spatial compositions of simple parts.

Проведено аналіз потенціалів гібсовських випадкових полів для моделювання апріорних ймовірнісних характеристик форми об'єкту. Показано, що завдяки цим полям другого порядку можна описувати як прості форми, так і просторові відношення між ними, що дозволяє розпізнавати складні форми як композицію більш простих.

1. Introduction

Motivation and goals

The 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 [7]. Corresponding guiding principles for the decomposition are identified from these experiments as well as from research in computer vision (see e.g. [9]). 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 modeling 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 [2] how to extend these models for scene segmentation. Recently we have shown how to use level set methods in conjunction with MRFs [4]. 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 [6, 11] 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. [10, 8] 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 the 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 be-

comes 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.

The structure of the paper is as follows. In section 2 we introduce the GRF model for composite shapes and discuss the inference and learning tasks. The latter means to learn the Gibbs potentials and the graphical structure itself. The section 3 gives experiments exploring the expressive power of the model − first we separately show its ability to express spatial relations between segments and its ability to model simple shapes. Then we demonstrate its capability to model composite shapes including structure learning. Finally, we show how to combine such models for the discrimination of 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: 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 $t' - t = a$. A 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 $y : D \to K$ 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:

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

where Z denotes the partition sum (we omit the unary terms for better readability). This p.d. is homogeneously parametrised – all edges in an equivalence class E_a have the same potentials.

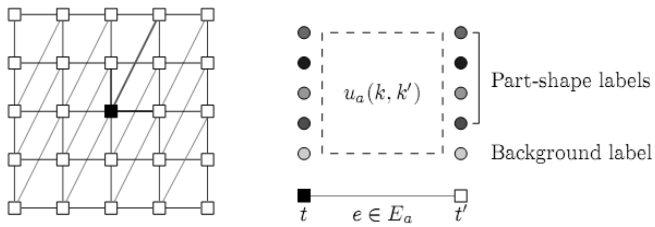

Fig. 1. Left: example of a translational invariant graphical structure. Equivalence classes of edges E_a are coloured by different colours. The set A is represented by bold edges outgoing from the central node. Right: Gibbs potentials for an edge from E_a.

Note that the parameters u_a of this model are unique up to additive constants for a given p.d. under fairly general assumptions − the only possible equivalent transformations (aka re-parametrisations) consist in adding a constant $\tilde{u}_a() = u_a() + \text{const}$. This will be shown in appendix A. Therefore, we assume from here that the Gibbs potentials for each $a \in A$ are normalised to sum to zero: $\sum_{k,k} u_a(k, k') = 0$.

It is important to notice that a homogeneously parametrised GRF on a finite domain $D \subset \mathbb{Z}^2$ is not necessarily homogeneous. A p.d. $p(y)$ for labellings $y : D \to K$ is called homogeneous if its marginals for congruent subsets coincide. This inhomogeneity, if present, usually reveals at the domain boundary. It is easy to verify that the converse is true at least for chains: a homogeneous Markov model on a finite chain admits a homogeneous parametrisation.

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

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

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 suggested for structured SVM models with latent variables [13]. This shows that the 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

$$C(y, y') = \sum_{t} 1\{y_t \neq y'_t\} \qquad (4)$$

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

$$y_t^* = \max_{k} p(y_t = k \mid x) \quad \forall t \in D. \qquad (5)$$

Hence, it is necessary to calculate the marginal posterior probabilities for each node $t \in D$ and label $k \in K$. Currently this task is infeasible for GRFs. Several approximation techniques based e.g. on belief propagation or variational methods have been suggested for this task (see e.g. [12] 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 [3].

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 $\mathcal{B} = (x, y_V) = \{(y, x) \mid y_{|V} = y_V\}$.

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

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

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

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

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

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

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

$$n_a(k, k'; y) = \sum_{tt' \in E_a} 1\{y_t = k, y_{t'} = k'\} \qquad (9)$$

and represent co-occurrences for label pairs (k, k') along the edges in E_a for a labelling y. Combining these random variables into a random vector Φ, the gradient of the log-likelihood can be written as

$$\nabla L(u) = \mathbb{E}_{p(y \mid \mathcal{B};u)}[\Phi] - \mathbb{E}_{p(y;u)}[\Phi]. \qquad (10)$$

The exact calculation of the expectations in (8) is not feasible. Therefore, we suggest 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 $p(y \mid \mathcal{B};u)$ 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 would like to mention that the learning of the appearance models $p(c \mid k)$ 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 it is not necessary to sample labellings according to the a-priori probability distribution $p(y)$. 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 $A = \left\{ a \in \mathbb{Z}^2 \,\middle|\, |a_1| \le d, |a_2| \le d \right\}$ 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. It is therefore important to investigate the possibility to identify the neighbourhood structure A from a given training sample. A possible variant of a corresponding formal task reads as follows. Given a training sample the task is to find the best neighbourhood structure A of given size $|A| = m$ according to the Maximum Likelihood principle $L(u_A, A) \to \max_{u_A, A}$. This task is however not feasible – an exhaustive search over all possible sets A would be computationally prohibitive, and, moreover, the likelihood can be calculated only approximatively. Therefore we rely on a greedy approximation which we will consider in two variants − one of them successively includes new elements into the neighbourhood structure starting from $A = \{0\}$ and the other successively removes elements from this structure starting from $A = \left\{ a \in \mathbb{Z}^2 \,\middle|\, |a_1| \le d, |a_2| \le d \right\}$.

For the first variant we use a greedy search for the interaction edges suggested by Zalesny and Gimel'farb in the context of texture modelling [14, 5]. Starting from the set $A = \{0\}$, i.e. a model with unary potentials, new edges are iteratively chosen and included into A as follows. First, the optimal set of potentials $u_A^* \in \mathcal{U}_A$ is determined for the current set A as described in the previous subsection. Here $\mathcal{U}_A$ denotes the subspace of potentials on the edges in A (we may assume 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 $\mathcal{U}_A$. The proposal is to include the vector $a' \in A' \setminus A$ with the largest gradient component

$$a' = \arg\max_{a \in A' \setminus A} \sum_{k,k'} [n_a(k,k';\mathcal{B},u) - n_a(k,k';u)]^2 . \quad (11)$$

Optionally the Kullback–Leibler divergence can be used instead of the Euclidean distance.

The second variant of structure estimation proceeds in opposite order. Starting with the neighbourhood structure $A = \left\{ a \in \mathbb{Z}^2 \,\middle|\, |a_1| \le d, |a_2| \le d \right\}$, elements of A are successively removed. The aim is to remove in each step the element with the smallest impact on the maximal likelihood

$$\max_{u_A} L(u_A) - \max_{u_{A \setminus a}} L(u_{A \setminus a}) \to \min_{a \in A}. \quad (12)$$

It is impossible to estimate this expression in the point $u_A^* = \arg\max_{u_A} L(u_A)$ using the gradient of the likelihood (like in the first variant) because of $\nabla L(u_A^*) = 0$. It is nevertheless possible to estimate this expression based on u_A^*. For the sake of simplicity we show this for the situation of super-

vised learning. The likelihood maximisation with respect to the Gibbs potentials reads

$$\max_{u_A}\{< \psi_A, u_A > - \log \sum_y \exp < \varphi_A(y), u_A >\} \quad (13)$$

for this case. Here we have used the following notations. The set of all Gibbs potentials $u_a(.,.)$, $a \in A$ is considered as a vector u_A. A realisation of the random vector Φ_A (see (10)) is denoted by $\varphi_A(y)$. Finally, ψ_A denotes the corresponding vector of statistics resulting from the training sample. Designating $\log Z(u_A)$ by $H(u_A)$, the expression in (13) is nothing but the Fenchel conjugate $H^*(\psi_A)$. It is known that for exponential families the latter can be written as

$$H^*(\psi_A) = \inf\{\sum_y p(y) \log p(y) \,|\, \mathbb{E}_p[\Phi_A] = \psi_A, \quad p \in \mathbb{E}\} \quad (14)$$

(see e.g. [12, 1]), where we denoted the expectation w.r.t. a probability distribution p by $\mathbb{E}_p$ and the set of all probability distributions on labellings y by $\mathcal{P}$. This means to find the p.d. with maximal entropy among *all* distributions having expectation ψ_A of the random vector Φ_A.

Removing an element a from the neighbourhood structure A can be equivalently expressed by the linear constraints $u_a \equiv 0$. Considering the task (13) with these additional constraints, it can be shown by the use of Fenchel duality (see e.g. [1]) that the corresponding conjugate function $\tilde{H}^*(\psi_A)$ can be written as

$$\tilde{H}^*(\psi_A) = \inf_{z_a} \inf_p \{\sum_y p(y) \log p(y) \,|\, \mathbb{E}_p[\Phi_A] = \psi_A + z_a, \quad p \in \mathbb{E}\}, \quad (15)$$

where z_a is an arbitrary vector of the subspace $\mathcal{U}_a$. Therefore, the difference in (12) is equal to $H^*(\psi_A) - \tilde{H}^*(\psi_A) = H^*(\psi_A) - H^*(\psi_A + z_a^*)$ and can be estimated by the gradient of H^* in ψ_A. The latter gradient is nothing but the vector of Gibbs potentials u_A^*.

The convex, lower semi-continuous function $H^*(\psi_A)$ is not differentiable in general. Therefore its sub-differential may consist of more than one subgradient u_A. This corresponds to the non-uniqueness of the Gibbs potentials. We have however shown that the Gibbs potentials are unique up to additive constants for the model class considered in this paper (see Appendix *A*).

Summarising, the difference in (12) can be estimated by $\|u_a\|$, that leads to the following greedy removal strategy for elements of the neighbourhood structure *A*. Given a current neighbourhood structure *A*, estimate the optimal Gibbs potentials u_A^* and remove the element $a \in A$ with the smallest value of $\|u_a\|$.

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 to capture 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: $K = \{sky, trees, grass\}$. 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 $\left\{a \in \mathbb{Z}^2 \,\middle|\, |a_1| \le d, |a_2| \le d\right\}$ 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 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.

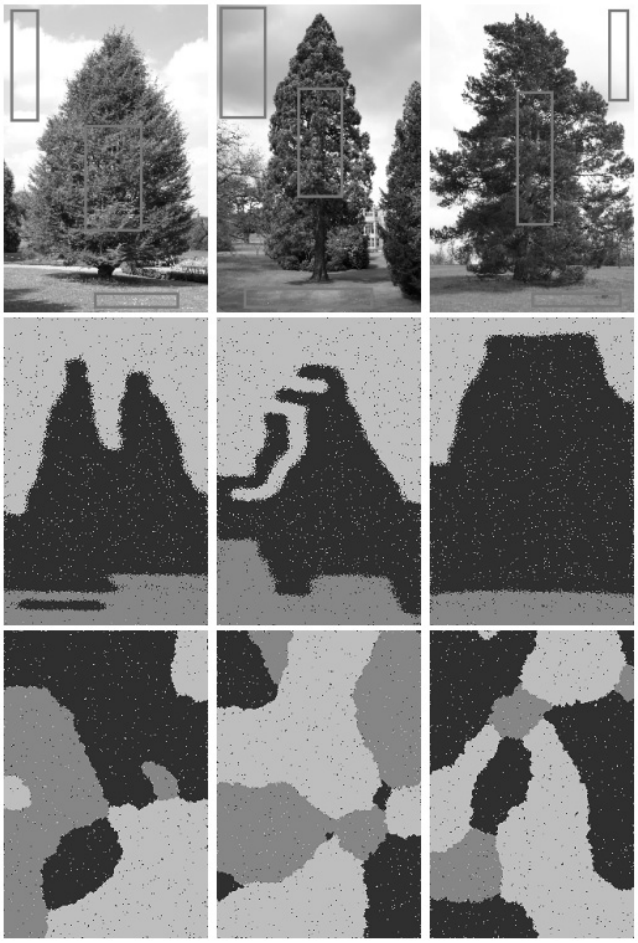

Fig. 2. Modelling spatial relations between segments. The first row shows input images and regions with fixed segmentation. The middle and bottom row show labellings generated by the learned a-priori models (segment labels are coded by colour): the images in the middle row were generated by the model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

The consequences can be clearly seen from the following experiment. We fixed the prior models obtained in the previous experiment (semi-supervised learning) for both variants (the complex prior and the Potts prior) 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 (obtained with the learned appearances) 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.

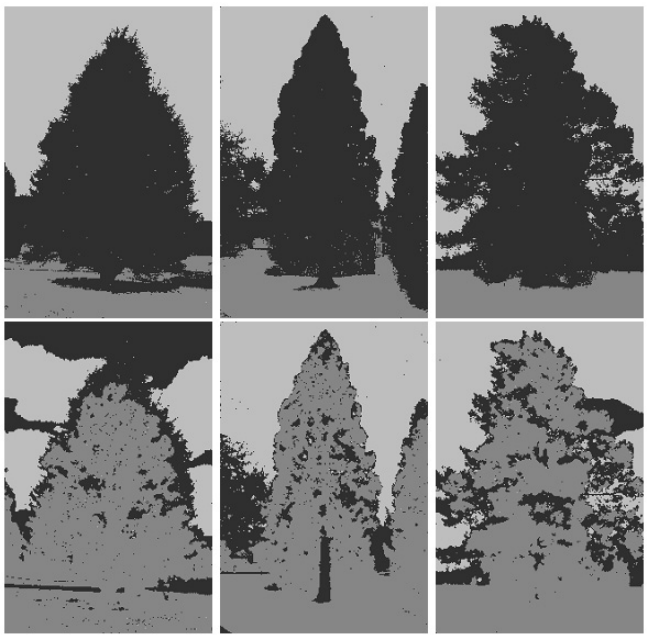

Fig. 3. Segmentation results obtained after fully unsupervised learning of the appearance part of the model. Upper row − model with full neighbourhood, bottom row − baseline model

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» (which have nearly circular shape) 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.

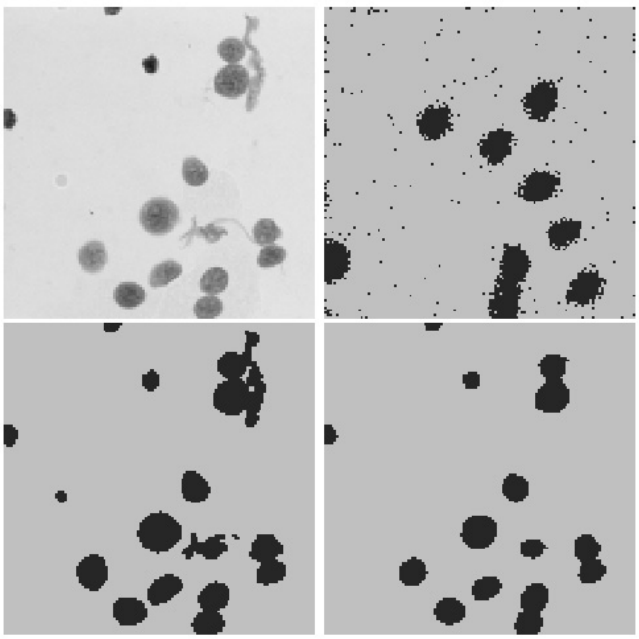

Fig. 4. Modelling and segmentation of simple shapes. Upper left − input image, upper right − a labelling generated a-priori by the learned complex model. Final segmentations are shown in the bottom row: left − baseline model, right − complex model

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 according to the max-marginal decision (see equation (5)) are shown in the bottom row of the same figure. The differences are clearly seen. The shape prior captured 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 nor even their combination allow to differentiate between cells and artefacts.

A structure estimation for simple shapes

In order to investigate the structure identifiability of shape models we have used an artificial model which generates simple «blobs». The neighbourhood structure consists of 8 elements. The group of the first four elements with coordinates (0,1), (0,–1), (1,1), and (–1,1) describes a standard 8-neighbourhood. The remaining four vectors are scaled versions of the first (scale factor 5). The Gibbs potentials on the short vectors are supermodular and express the correlation of the labels on the edges of this type

$$u(k,k') = \begin{cases} \alpha & \text{if } k = k', \\ -\alpha & \text{else.} \end{cases} \qquad (16)$$

The Gibbs potentials on the long edges consist of an submodular and a modular part $u(k,k') = u_1(k,k') + u_2(k,k')$, where the submodular part u_1 is just the negative version of the potentials on the short edges and expresses an anti-correlation of the labels on these edges. The modular part

$$u_2(k,k') = \begin{cases} \beta & \text{if } k = k' = 0, \\ -\beta & \text{if } k = k' = 1, \\ 0 & \text{else}, \end{cases} \qquad (17)$$

is used to influence the density of the blobs. A labelling (fragment) sampled by this model ($\alpha = 0.35$, $\beta = 0.5$) is shown in Fig. 5. Both heuristic approaches for structure estimation discussed in the previous section where applied for the supervised version, i.e. using a labelling generated by the known model as a learning sample.

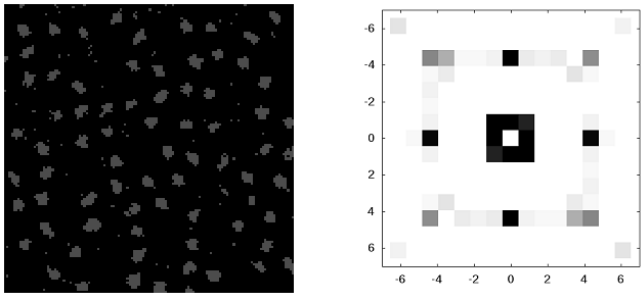

Fig. 5. Shape estimation for a simple shape model. Left − labelling generated by the known model, right − histogram of the estimated structures

The first approach − iterative growth of the structure − was run 40 times. The estimated structures resulting from these runs are shown in Fig. 5 as a grey-coded histogram. As a stochastic gradient ascend is used for the learning of the potentials, each run may result in a different structure. The histogram shows however, that the structure estimation is essentially correct. All trials of the second approach − iterative shrinking of the neighbourhood structure − resulted much to our surprise in one and the same estimated structure − the correct one.

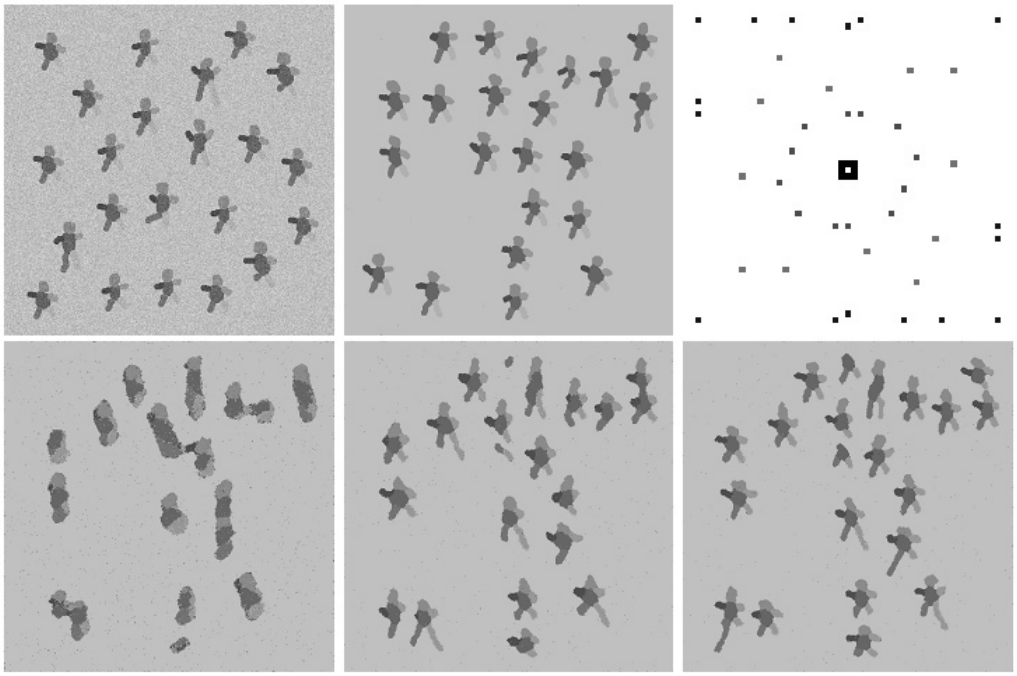

Fig. 6. Composite shape modelling. Upper row from left: input image, labelling generated a-priory by the learned model, estimated interaction structure. Bottom row: labellings generated by models during learning

We conclude from these experiments that the neighbourhood structure of a shape model is identifiable (at least in principle) from labellings generated by the model.

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. 6 (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 $p(c \mid k)$ for the labels are Gaussians with known parameters. In this experiment we applied the growth variant for the estimation of the interaction structure as described in section 2.

Fig. 6 (upper row, center) shows a labelling generated by the learned prior model. It is clearly seen that both, spatial relations between object parts and part shapes are captured correctly.

The bottom row of Fig. 6 displays labellings generated during the process of structure learning at time moments, when the 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. 6 (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.

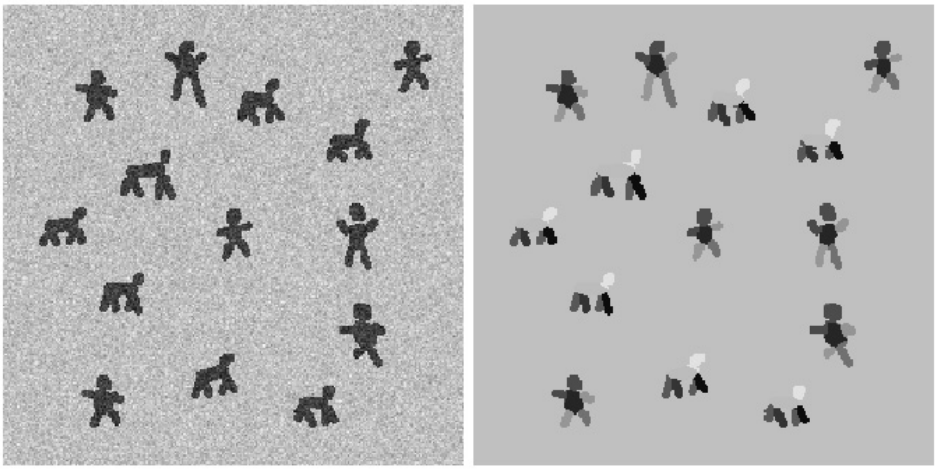

Fig. 7. Shape segmentation and classification. Left − input image, right − segmentation (part-labels are encoded by different grey values)

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. 7. We conclude from the previous experiments, that the appearance model can be re-learned in a fully unsupervised way if the prior shape model is discriminative. Hence, the most important question is, how to combine the prior models. We propose a method for this that is based on the following observation. It is not necessary to have an example image (or an example segmentation) in order to learn the model if the aposteriori statistics

$$\bar{\Phi}_a(k,k') = \mathbb{E}_{p(y|\mathcal{B},u)}[\Phi_a(k,k')] \qquad (18)$$

for all difference vectors $a \in A$ and label pairs (k,k') are known − the gradient of the likelihood (equation (8)) reads then

$$\partial L / \partial u_a(k,k') = \bar{\Phi}_a(k,k') - \mathbb{E}_{p(y;u)}[n_a(k,k';y)]. \qquad (19)$$

Let us consider this in a bit more detail for a simple example − just two shapes like in Fig. 7. Let us assume that the both models are learned, i.e. both the potentials and statistics are known for both models and for all difference vectors a. Obviously, it is not easy to combine the potentials of both shape models in order to obtain new ones for a model that generates such collages. It is however very easy to estimate the needed aposteriori statistics for the joint model given the aposteriori statistics for both shape models. Summarizing, the scheme to obtain the parameters of the joint model consists of two stages:

1. compute the aposteriori statistics for the joint model and

2. learn the model according to (19) so that it reproduces this statistics.

As the second stage is standard, we consider the first one in more detail. Let us denote the label sets corresponding to the shape parts by K^1 and K^2 for the first and for the second shape type respectively. Let b^1 and b^2 be the background labels in the corresponding shape models and b be the background label in the joint one. Consequently, the label set of the latter is $K^1 \cup K^2 \cup b$ (see the middle part of Fig. 8).

First of all we enlarge the label sets of each shape model by labels that are not present in this model but present in the joint one. Thereby the statistics for the new introduced labels (for all difference vectors a) are set to zero (see Fig. 8, left and right). Informally said, these extended aposteriori statistics correspond to the situations that the joint model is learned on examples, in which only labels of one particular shape are present. The aposteriori statistics for the joint model is then obtained as a weighted mixture of the two extended ones and an additional uniformly distributed component. The latter is added in order to avoid zero probabilities (which would lead to obvious techni-

cal problems for the Gibbs Sampler). Summarising, the aposteriori statistics of label pairs for a difference vector a of the joint model is:

$$\bar{\Phi}_a(k,k') \sim \begin{cases} w_1 \cdot \bar{\Phi}_a^1(k,k') + w_0 & \text{if } k \in K^1 \text{ and } k' \in K^1, \\ & \quad k \in K^1 \text{ and } k' = b, \\ & \quad k = b \text{ and } k' \in K^1, \\ w_2 \cdot \bar{\Phi}_a^2(k,k') + w_0 & \text{if } k \in K^2 \text{ and } k' \in K^2, \\ & \quad k \in K^2 \text{ and } k' = b, \\ & \quad k = b \text{ and } k' \in K^2, \\ w_1 \cdot \bar{\Phi}_a^1(b^1,b^1) + & \\ \quad + w_2 \cdot \bar{\Phi}_a^2(b^2,b^2) + w_0 & \text{if } k = b \text{ and } k' = b \\ w_0 & \text{otherwise.} \end{cases} \quad (20)$$

with some weights $w_0 \ll w_1 \approx w_2$, where the indices 1 and 2 correspond to the particular shape model. Given these statistics the joint model is learned according to (19).

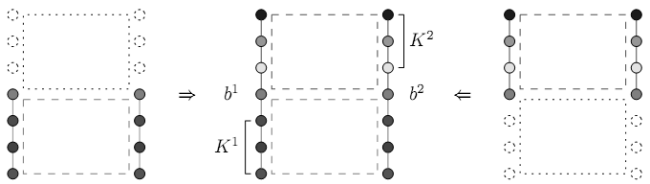

Fig. 8. Estimation of the aposteriori statistics for the joint model. Left and right: extended statistics for shape models. Middle: the joint model − statistics marked green and red are inherited from the components. Others are set to a small constant

For the experiment in Fig. 7 two composite shape models were learned separately. The test image in Fig. 7 (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. 7 (right) shows the final segmentation. It is seen that all objects were correctly segmented and recognised − although both composite shape classes share some 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.

Acknowledgments

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 the Czech Ministry of Education project 1M0567. D.S. was supported by the 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.

1. *Borwein J.M.*, *Lewis A.S.* Convex Analysis and Nonlinear Optimization of CMS Books in Mathematics. Springer, 2000.
2. *Cremers D.*, *Sochen N.*, *Schnörr C.* A multiphase dynamic labeling model for variational recognition-driven image segmentation // IJCV 66(1). – P. 67–81 (January 2006).
3. *Flach B.*, *Schlesinger D.* Combining shape priors and MRF-segmentation // da Vitoria Lobo et al., (ed) S+SSPR 2008. – P. 177–186. Springer (2008).
4. *Gimel'farb G.L.* Texture modeling by multiple pairwise pixel interactions // IEEE Trans. Pattern Anal. Mach. Intell. – P. 1110–1114 (1996).
5. *Hinton G.E.* Training products of experts by minimizing contrastive divergence. Neural Computation. – 14(8). – P. 1771–1800 (2002).
6. *Hoffman D.D.* Visual Intelligence: How We Create What We See. W.W. Norton & Company (2000).
7. *Associative* Hierarchical CRFs for Object Class Image Segmentation / *L. Ladicky*, *C. Russell*, *P. Kohli*, *P.H. Torr* // Proc. IEEE 12. Intern. Conf. on Computer Vision (2009).
8. *Liu H.*, *Liu W.*, *Latecki L.J.* Convex shape decomposition // *CVPR.* – P. 97–104 (2010).

9. *Ramalingam, S., Kohli, P., Alahari, K., Torr, P.* Exact inference in multi-label CRFs with higher order cliques // CVPR. – 2008. – P. 1–8 (2008).
10. *Roth S., Black M.J.* Fields of Experts // Intern. J. of Computer Vision. – **82**(2). – P. 205–229 (2009).
11. *Sokal A.D.* Monte Carlo Methods in Statistical Mechanics: Foundations and new Algorithms. Lectures notes (1989).
12. *Wainwright M.J., Jordan M.I.* Graphical models, exponential families, and variational inference // Foundations and Trends in Machine Learning **1**(1–2). – P. 1–305 (2008).
13. *Yu C.N.J., Joachims T.* Learning Structural SVMs with Latent Variables // Intern. Conf. on Machine Learning (ICML) (2009).
14. *Zalesny A., Gool L.V.* Multiview Texture Models // Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. – 2001. – P. 615–622. IEEE Computer Society (2001).

Appendix **A**

Equivalent transforms for homogeneously parametrised GRFs

As we have already seen, the probability distribution (1) for shape part labellings y can be equivalently written as

$$p(y) \sim \exp\left[\sum_k u_0(k) n_0(k;y) + \sum_{a\in A'}\sum_{kk'} u_a(k,k') n_a(k,k;y)\right], \quad (21)$$

where $A' = A \setminus \{0\}$. We call two parametrisations u, $\tilde{u}$ equivalent, if the corresponding probability distributions are identical. It follows that the difference $v = u - \tilde{u}$ of equivalent potentials fulfils

$$V(y) = \sum_k v_0(k) n_0(k;y) + -\sum_{a\in A'}\sum_{kk'} v_a(k,k') n_a(k,k';y) = \text{const}. \quad (22)$$

We will conclude that all functions v_a are constant under fairly general conditions. We perform the proof in two steps. First we show that the pairwise functions v_a, $a \neq 0$ are modular and can be written as a sum of unary functions. In a second step we will conclude the claimed statement under fairly general conditions for the graph (*D*, *E*).

Let us consider an arbitrary non-zero vector $a \in A$ of the neighbourhood structure and an arbitrary edge $(tt') \in E_a$. Let k_1, k_2 be two arbitrary labels in the node *t* and k_1', k_2' be two arbitrary labels in the node t'. Let $y_{11}, y_{12}, y_{21}, y_{22}$ be four labellings with respective values (k_1, k_1'), (k_1, k_2'), (k_2, k_1'), (k_2, k_2') on the nodes t, t' such that they coincide on all other vertices. We consider the equation

$$V(y_{11}) + V(y_{22}) - V(y_{12}) - V(y_{21}) = 0. \quad (23)$$

It is easy to see that this equation reduces to

$$v_a(k_1, k_1') + v_a(k_2, k_2') - -v_a(k_1, k_2') - v_a(k_2, k_1') = 0. \quad (24)$$

This holds for arbitrary four-tuples of labels and it follows that the function v_a is modular and can be written as a sum of two unary functions

$$v_a(k, k') = \tilde{v}_a(k) + \tilde{v}_{-a}(k'). \quad (25)$$

These arguments can be applied for every element $a \in A'$. Consequently, $V(y)$ can be written as

$$V(y) = \sum_k v_0(k) n_0(k;y) + + \sum_{a\in A'}\sum_k [v_a(k) n_a(k;y) + v_{-a}(k) n_{-a}(k;y)], \quad (26)$$

where we have omitted the tildes. Note that $n_a(k;y) = \sum_{k'} n_a(k,k';y)$ denotes the number of vertices with an outgoing edge of type *a* for which the labelling *y* has the value *k*. Therefore in general $n_0(k;y) \neq n_a(k;y)$.

Let us consider an arbitrary vertex *t* and two labellings $y, \tilde{y}$ which coincide on all vertices but *t*. It follows from $V(y) - V(\tilde{y}) = 0$ that

$$v_0(k) + \sum_{\substack{a\in A' \\ t+a\in D}} v_a(k) + \sum_{\substack{a\in A' \\ t-a\in D}} v_{-a}(k) = \text{const}. \quad (27)$$

We assign a vector $z(t)$ with dimension $2|A|-1$ to every vertex $t \in D$ with components

$$z_0(t) = 1,\ z_a(t) = \begin{cases} 1 & \text{if } t+a \in D, \\ 0 & \text{else} \end{cases} \quad \text{and} \quad z_{-a}(t) = \begin{cases} 1 & \text{if } t-a \in D, \\ 0 & \text{else.} \end{cases} \quad (28)$$

If the domain *D* contains a subset of nodes *t* such that their vectors $z(t)$ span the whole vector space of dimension $2|A|-1$, then, clearly, considering equation (27) for each of them, we obtain

$$v_0(k) = \text{const}. \quad (29)$$

$$v_a(k) = \text{const}. \quad (30)$$

$$v_{-a}(k) = \text{const}. \quad (31)$$

for all $a \in A$.