Multimed Tools Appl (2017) 76:9169–9194
DOI 10.1007/s11042-016-3513-0

Conditional random field with the multi-granular contextual information for pixel labeling

Jie Zhao[1,2] · **Gang Xie**[1] · **Jiwan Han**[3]

Received: 21 August 2015 / Revised: 14 March 2016 / Accepted: 4 April 2016 /
Published online: 26 April 2016

Abstract To make full use of the contextual information object recognition and scene understanding, a multi-granular context conditional random field (MGCCRF) model is presented to combine context information in a variety of scales. It is efficiently implemented through extending the pairwise clique to the multi-granular context windows. In the fine-granular context window, the label consistency of similar features can be obtained with the probability of the label transferring between two adjacent pixels. At the same time, the spatial relationships among different classes in the coarse-granular context window are explicated in details. To train the MGCCRF model, a piecewise training method with the bound optimization algorithm is designed to improve the performance. Experiments on two real-world image databases show that compared with other methods, the modified conditional random field model is more competitive and effective in terms of the quantitative and qualitative labeling performance.

Keywords Conditional Random Field (CRF) · Contextual information · Multi-granular context · Pixel labeling

✉ Gang Xie
xiegang@tyut.edu.cn

Jie Zhao
tydxcomputer@163.com

[1] College of Information Engineering, Taiyuan University of Technology, No.79 West Yingze Street, Taiyuan 030024 Shanxi, China

[2] Department of Computer Engineering, Taiyuan college, No.18 South Dachang Road, Taiyuan 030032 Shanxi, China

[3] National Plant Phenomics Center, Aberystwyth University, Aberystwyth SY23 3EE, UK

1 Introduction

Pixel labeling approaches work with a predefined set of class labels that dictates the categories of objects and types of scenes through assigning a semantic label to each pixel. Pixel labeling plays an important role in scene understanding and object recognition so that the labeling framework can encode the complex relationship between the visual appearance of a scene and the underlying semantic labels. Labeling requires context information. Since Conditional Random Field (CRF) model has intrinsic ability to incorporate the context information in both labels and observed images in a principled manner, CRF is a popular method for pixel labeling [7, 9, 15].

The problem of labeling the semantic classes to the pixels in the image is a challenging task due to ambiguities in the appearance of the visual data. For example, the sky and the water patches may locally look very similar due to the flat blue area. The red box in Fig. 1 is the fine-granular contexts that contain a few neighbor pixels wide. If we use only the local contexts, the water and the sky cannot be clearly distinguished. However, the coarse-granular contextual information is obtained in the large-scale window as shown in the green box of Fig. 1. This type of coarse-granular context refers to the relative spatial configurations between the objects. It shows the fact that the boats tend to be in the water and airplanes in the sky so that the visual ambiguity between the sky and the water can be solved. So the contex information at different levels can help alleviate this problem significantly.

In general, an image contains the useful information for labeling at several levels. We thus use the multi-granular contextual information from different levels to improve the labeling performance. The key contribution of this paper is a framework which provides an approach to incorporate the fine-granular and the coarse-granular context information into a single model. The proposed model uses the fine-granular contextual information to produce the continuous object surface and preserve the accurate object boundaries. Meanwhile, it can also adopt the coarse-granular contextual information to improve the performance of object recognition. To solve the problems of the image labeling, the piecewise training method with the bound optimization algorithm is utilized to develop a parallel training method for the proposed model.

Section 2 of this paper presents a brief discussion of related work. Section 3 is about the pixel labeling model. Section 4 introduces the training and inference method for the proposed model. Section 5 describes the experiments and the analysis concerned and Section 6 shows the conclusion.

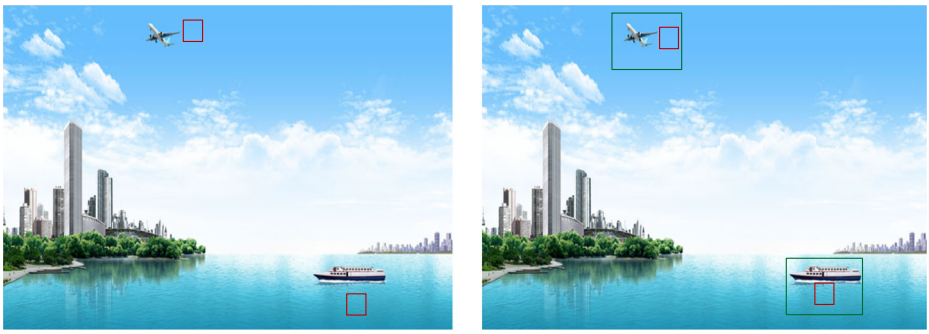


Fig. 1 Context information at different ranges

2 Related work

Typically, image labeling in the form of extracted feature vectors and the training semantic labels are used by machine learning algorithms with an attempt to automatically apply annotations to new images [19]. The semantic labeling approaches for image analysis and understanding have proceeded along with many separate trajectories.

One of image semantic labeling approaches is regarded as a type of multi-label classification which focuses on providing a high-level summary or categorization of the image context by using a few labels. Multi-label image classification is a supervised learning problem that an instance may be associated with multiple labels. Nasiereling et al. [16] presented a clustering based multi-label classification framework. The proposed framework comprises an initial clustering phase that breaks the original training set into several disjoint clusters of data, and then trains a multi-label classifier from the data of each cluster. On the other hand, the labels can be propagated on holistic image similarity. Chen et al. [3] proposed to construct graph on label level to reveal correlation. Liu et al.[13] introduced a label similarity matrix to provide a semi-supervised learning algorithm. These methods do not separate the similarity among different labels. In order to involve other labels when propagating a certain label, methods based on image decomposition have been explored [2, 29, 33]. However, the automated solution is still far from satisfaction because of the limitation with the process of segmentation. Bao et al. [1] considered to implicitly decompose the label representation on feature level to avoid the explicit image segmentation process. In fact, two images with the similar visual contents may correspond to quite different semantic concepts. Yu et al. [32] proposed a multi-label classification framework based on the neighborhood rough sets to reduce the bias between visual similarity and semantic similarity. By introducing the concept of upper and lower approximations of neighborhood rough set model, the framework can find all the possibly related labels of the given instance and then confirm the final labels according to the information of the neighborhood of the given instance. In general, such approaches are concerned with the task of assigning a few semantic labels to a given image without explicitly identifying the locations of objects in the image.

The second category is generally founded on pixel semantic labeling techniques which aim to locate the discrete objects in an image. The pixel-labeling approaches work with a predefined set of class labels that dictate the categories of objects [18, 25]. While the semantic information of the image is described by the appropriate labels, the detailed object outlines are also provided at the pixel level. The pixel-labeling approaches to image understanding and analysis are our main focus in this paper. The main research direction in the pixel-labeling approaches is toward application of statistical methods that solve the labeling problem where the correct label has to be assigned to each pixel through capturing the full interaction between pixels. Thus most methods for pixel labeling use a probabilistic model which provides a formal framework for encoding the complex relationship between the visual appearance of a scene and the underlying semantic labels.

As one of the popular models for gridded image-like data, Markov Random Field (MRF) framework is a classical probabilistic approach for modeling to fuse the low-level image statistics and high-level contextual information. Wilson et al. [28] incorporated differences between neighboring sites into likelihood to capture local contextual

information, and assumed that the differences follow simple normal distribution. While in [17], the differences are demonstrated as non-Gaussian and heavy-tailed, and then are modeled by two parametric families, i.e., Bessel K form and generalized Laplacian. This is the fact that neighborhood relationships encoded in the MRF are a relatively weak cue, stronger information such as relative location and containment relationships should be included in the MRF. For example, Posner et al. [21] used MRF to model the expected relationships between patch labels both spatially and temporally, thus capturing some of the strong structural relationships between parts of a typical urban scene. In [11, 20, 22], several MRF frameworks which utilized the observed data at a given site and its parents, have also been developed to capture the dependencies in observed data. Although these extended MRF frameworks have abilities to capture the contextual information in observed data, they also make simplified assumptions to get some sort of factored approximation of likelihood for computational tractability. For most of the real-world applications, this assumption is too simplistic.

The Conditional Random Field (CRF), another probabilistic graphical model, avoids the problem of explicit modeling likelihood in MRF framework and has intrinsic ability to incorporate the contextual information in both labeling and observed images in a principled manner. In recent years, many researchers focus on modifying the CRF model for the pixel labeling. For example, He et al. [6] generalized the standard form of feature functions used in CRF to use hidden variables, each encoding a learned pattern within a subset of label variables. The proposed model is a product combination of individual models to respect the relationship between the objects at both local and global scales. Thus a wide variety of patterns of labels at different scales are represented by the features, and the features all interact at the label layer. To encode the context at different scales, the different hierarchical structures contribute to the image labeling in the CRF model. Russell et al. [24] proposed the hierarchical random field model that allows integration of features computed at different levels of the quantization hierarchy. Huang et al. [8] introduced the hierarchical two-stage CRF model which combines the ideas used in both parametric and nonparametric image labeling methods. Roig et al. [23] proposed the hierarchical random field for part based model which incorporates relations among sets of parts. Yang et al. [30] presented the hierarchical CRF model which aggregates evidence from local to global level by using multi-scale mean shift segmentation. These hierarchical CRF models addressed the combination of global and local features to improve the performance. Moreover, a tree conditional random field framework [15] is used to allow for more complex dependencies by using multiple labels per node, and mixtures of trees. Since the use of context has been well documented for image labeling, these models are expressive than independent predictors, and they will lead to more accurate label predictions.

In general, how to effectively incorporate the contextual information into the labeling model is always a challenge for the research direction of computer vision. We will discuss to address the problem in this paper through a novel method. The pairwise potential of CRF framework is directly extended to two kinds of potentials to incorporate the multi-granular contextual information in order to generate more reasonable labeling results. Simultaneously, estimating for the labeling parameters is performed by the piecewise training model with the bound optimization algorithm. In the parallel way, the performance efficiency can be improved.

3 The proposed CRF framework

Let the observed data from an input image be given by $\mathbf{y} = \{y_i\}_{i \in S}$, where $y_i = [y_{i1}, y_{i2}, \cdots, y_{id}]$ denotes a feature vector that encodes appearance based features from the ith site. $S = [1, 2, \cdots, n]$ is the set of all the image sites, n is the number of the pixels in the image, and d is the dimension of visual features. Denote $\boldsymbol{L}$ as the set of all possible labels associated with an observation, where $|\boldsymbol{L}|$ is the number of label classes. So our goal is to assign a label $x_i \in \boldsymbol{L}$ to each site $i \in S$ in a way of the discriminative CRF framework. Therefore, the corresponding labels of all sites are given by $\mathbf{x} = \{x_1, x_2, \cdots, x_n\}$.

Compared with generative models including MRF, CRF models the contextual dependencies in a probabilistic discriminative framework that directly considers the posterior distribution over labels given observations. Then CRF relaxes the strong independence assumption and captures neighborhood interactions among observation. In the MRF framework, the prior $P(\mathbf{x})$ is usually modeled by Gibbs distribution. Given an image $\mathbf{y}$, we are interested in obtaining the conditional distribution $P(\mathbf{x}|\mathbf{y})$ of the true labels. If the posterior distribution of label field modeled directly by Gibbs distribution, $(\mathbf{x}, \mathbf{y})$ is said to be a CRF [34]. We write formulation of CRF as multiplicatively combining component conditional distributions that capture contextual information at different levels c:

$$P(\mathbf{x}|\mathbf{y}, \theta) = \frac{1}{Z(\mathbf{y}, \theta)} \prod_{c \in C} \psi_c(\mathbf{x_c}, \mathbf{y}, \theta) \tag{1}$$

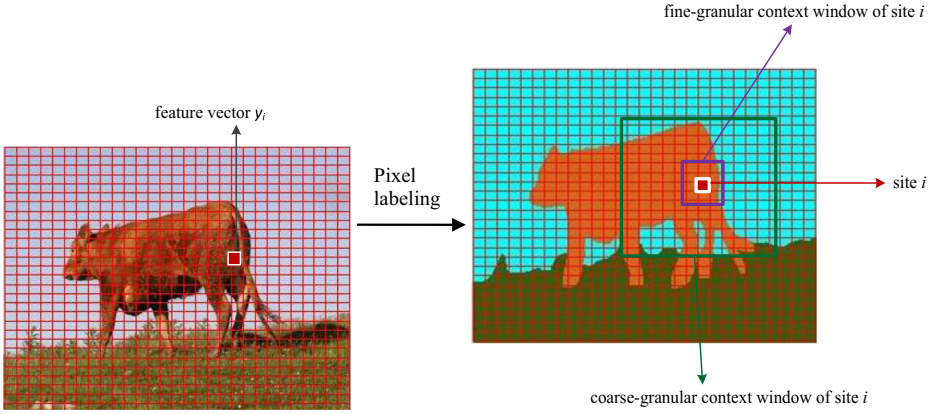
where c is a clique and C is the set that consists of all cliques. $Z(\mathbf{y}, \theta) = \sum_{\mathbf{x}} \prod_{c \in C} \psi_c(\mathbf{x_c}, \mathbf{y}, \theta)$ is the partition function and ψ_c is the potential defined on clique c with parameter θ. Our task is to learn a mapping from images to labels by assigning a label to each pixel upon the visual features, hoping the labels are as close to the ground truth as possible. This problem can be formulated naturally under the multi-granular context CRF (MGCCRF) framework: i) The fine-granular contextual information represents the short range interactions among a few neighbor sites in order to label smoothly for pixelwise labeling and keep the geometric consistency among parts of an object, and ii) The coarse-granular contextual information is the long range interactions encoded by the relative spatial co-occurrence between different semantic labels.

With the form of CRF in (1), MGCCRF considers the CRF framework with only unary and pairwise clique potentials. Thus the pairwise clique is extended to incorporate the multi-granular contextual information into this model. In this paper, the multi-granular context windows are defined to capture more contextual information and improve the labeling results. In the fine-granular context window, the label consistency of similar features can be captured. Simultaneously, the spatial relationships between classes are obtained in the coarse-granular context window. The example of the multi-granular context windows in MGCCRF is shown in Fig. 2.

The remaining issue of formulating MGCCRF model to pixel labeling is how to define two kinds of potentials, i.e. unary and pairwise clique potentials, while pairwise potential is decomposed into two parts according to the granularity of the contextual information.

3.1 Unary potential

The unary potential is used to model and discriminate observations for single image site. Each site corresponds to a single image pixel in the unary potential.



(a) Observed data (b) Labeling result

Fig. 2 The example of the multi-granular context windows. **a** Observed data. **b** Labeling result

Generalizing the binary form in [12] to multiclass problems, we model the unary potential as

$$\prod_{i\in S}\phi_i(x_i,\mathbf{y},\mathbf{w}) = \prod_{\substack{i\in S\\ k\in L}}\delta(x_i=k)\log p(x_i=k|\mathbf{y},\boldsymbol{\lambda}) \tag{2}$$

where $\delta(x_i=k)$ is 1 if $x_i=k$ or 0 otherwise, and $p(x_i=k|\mathbf{y},\boldsymbol{\lambda})$ is an arbitrary domain-specific discriminative classifier. This form of unary potential gives us the desired flexibility to integrate different applications preferring different types of local classifiers in a single framework. To model $p(x_i=k|\mathbf{y},\boldsymbol{\lambda})$, we generalize the logistic regression classifier to a softmax function and model the multinomial logistic regression (MLR)

$$\begin{aligned} p\left(x_i=k\middle|\mathbf{y},\boldsymbol{\lambda}\right) &= \frac{\exp\left(\lambda_k^T y_i\right)}{\sum_{k=1}^{|L|}\exp\left(\lambda_k^T y_i\right)} \\ &= \frac{\exp\left(\sum_d \lambda_{kd}^T y_{id}\right)}{\sum_{k=1}^{|L|}\exp\left(\sum_d \lambda_{kd}^T y_{id}\right)} \end{aligned} \tag{3}$$

where $\lambda_k=[\lambda_{k1},\lambda_{k2},\cdots,\lambda_{kd}]$ is the parameter vector for the kth class, and $\boldsymbol{\lambda}$ denotes a $|L|$-dimensional vector produced from concatenating the vectors $\{\lambda_k, k=1,2,\cdots,|L|\}$.

3.2 Pairwise potential

The pairwise potential predicts how the labels at two sites will interact given the observations. To define pairwise potential, we mainly focus on its ability to encode the large range context information. The pairwise clique is the neighborhood system of the site in the MGCCRF model. It is extended to the fine-granular neighbor set η_i^1 and the coarse-granular neighbor set

η_i^2. Figure 3a shows the structure of η_i^1 which is represented by site i and its eight adjacent neighbors. The coarse-granular context window is divided into eight cells $\{R_i^1, R_i^2, R_i^3, R_i^4, R_i^5, R_i^6, R_i^7, R_i^8\}$, and each cell is regarded as a whole neighbor site. Figure 3b demonstrates the structure of η_i^2 which is the set of the regions centered around the site i. So the neighborhood system is defined as the multi-granular context windows to capture more contextual information. According to the multi-granular contextual information, we model the pairwise potentials. In the following sections, the two pairwise potentials are introduced elaborately.

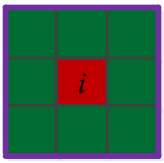
3.2.1 Fine-granular context

In the fine-granular neighborhood, the pairwise potential intends to represent the interaction relationships between a pair of sites in neighbor set η_i^1. We model the pairwise potential by using the method similar with the unary potential

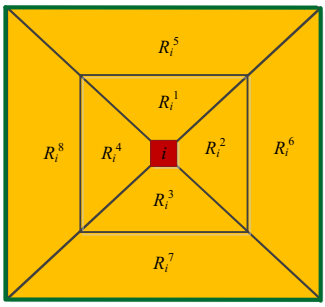
$$\prod_{i\in S,\, j\in\eta_i^1} \phi_{ij}\left(x_i, x_j, f_{ij}, \boldsymbol{\alpha}\right) = \prod_{\substack{i\in S,\, j\in\eta_i^1 \\ k\in L}} \delta(x_i = k)\delta(x_j = l)p\left(x_i = k, x_j = l \middle| f_{ij}, \boldsymbol{\alpha}\right) \quad (4)$$

where $\boldsymbol{\alpha}$ is the parameter vector in the fine-granular neighborhood, and f_{ij} is a feature vector of site pair (i, j) obtained by concatenating all elements of two vectors y_i and y_j. The feature vector f_{ij} is regarded as the fine-granular context descriptor for a pair of sites (i, j). $p(x_i = k, x_j = l|f_{ij}, \boldsymbol{\alpha})$ denotes the statistic results of the labels between two adjacent sites. The labels vary smoothly on the surface of an object, but change dramatically at the object boundary. In order to implement the smoothness of pixel labels and reduce the computational burden, we only take account of the consistency of the semantic labels between two adjacent sites in the fine-granular context window. Therefore, we use the transferring characteristic of the labels in the homogeneous sites to represent the pairwise potential of the fine-granular neighborhood so that the pairwise potential can be simpler and more effective.

At first, we build the eight-connected neighborhood as η_i^1 in both the labels and the observed data for the site i. Then, let α denote the label smoothing parameter for the fine-granular contextual information, and (i, j) be a pair of sites in η_i^1. If the label x_i of the site i is the *kth* class and the label x_j of the site j is the *lth* class, which means $x_i \neq x_j$, the label smoothing parameter α_{kl} is set to 0. We only consider $\{\alpha_{kk}\}_{k\in\{1,2,\cdots,L\}}$ that relates to the pair of sites with



(a) The fine-granular neighbor set η_i^1



(b) The coarse-granular neighbor set η_i^2

Fig. 3 The structures of the pairwise clique. **a** The fine-granular neighbor set η_i^1. **b** The coarse-granular neighbor set η_i^2

the same semantic label so that the fine-granular contextual information can reflect the consistency of local label classes. Therefore, the probability of the label transferring from the site i to its adjacent site j is calculated by the following equation:

$$p\left((x_i,x_j)\triangleq k\middle|f_{ij},\alpha,j\in\eta_i^1\right)=\begin{cases}\dfrac{\exp\left(\alpha_{kk}^T f_{ij}\right)}{1+\sum_{k=1}^{|L|}\exp\left(\alpha_{kk}^T f_{ij}\right)}, & if \quad k\le|L|\\ \dfrac{1}{1+\sum_{k=1}^{|L|}\exp\left(\alpha_{kk}^T f_{ij}\right)}, & if \quad k=|L|+1\end{cases} \tag{5}$$

where $(x_i,x_j)\triangleq k$ denotes the fact $x_i=x_j=k$ if $k\le|L|$, while $(x_i,x_j)\triangleq k$ denotes the fact $x_i\neq x_j$ if $k=|L|+1$. η_i^1 is the fine-granular neighborhood system that is represented by the eight-connected neighborhood of the site i as shown in Fig. 3a. Obviously, the transferring probability of the same label can be calculated through the generalized MRL classifier with $|L|+1$ classes.

According to the transferring characteristic of the labels in the homogeneous sites, the pairwise potential of the fine-granular neighborhood can be reformulated as follows

$$\prod_{i\in S,j\in\eta_i^1}\phi_{ij}\left(x_i,x_j,f_{ij},\alpha\right)=\prod_{\substack{i\in S,\,j\in\eta_i^1\\ k\in L}}\log p\left((x_i,x_j)\triangleq k\middle|f_{ij},\alpha,j\in\eta_i^1\right) \tag{6}$$

Consequently the pairwise potential with the fine-granular context can keep the label smooth on an object surface through the statistics of the consistency of local label classes. This leads to good results at the object boundaries.

3.2.2 Coarse-granular context

We describe the coarse-granular context in the large-range neighborhood to impose the spatial interaction to improve the recognition of the objects. The neighbor regions adjacent to the site i are divided into eight sub-regions $R_i^o(o=1,2\cdots,8)$ as shown in Fig. 3b. Each sub-region is regarded as a whole neighbor site for the site i. Thus arbitrary neighbor site is not a pixel but a region within the coarse-granular context window. This type of contextual information can explore not only the co-occurrence of two semantic classes but also the spatial relative location between the semantic classes.

To facilitate the expression, the coarse-granular neighbor site R_i^o of the site i is abbreviated as o. In the coarse-granular neighborhood η_i^2, we take into account the co-occurrence relationship among the different semantic labels. Thus, the pairwise potential between the site i and its coarse-granular neighbor site o is defined as the following equation through a generalized Ising model

$$\prod_{i\in S,o\in\eta_i^2}\phi_{io}(x_i,x_o,h_{io},\boldsymbol{\beta})=\prod_{\substack{i\in S,\,o\in\eta_i^2\\ m,n\in L}}\beta_{mn}\mu_{io}^n\delta(m\neq n) \tag{7}$$

where $\boldsymbol{\beta}$ is the parameter for the co-occurrence of semantic labels. We define h_{io} as the coarse-granular context descriptor for the neighbor site o of the site i, which describes the spatial co-occurrence contextual information of different classes in the coarse-granular neighborhood o. μ_{io}^n

is the *nth* element of h_{io}, and μ_{io}^n represents the maximum of the *nth* class likelihood maps when there is the label n in the site o, i.e., $h_{io} = \{\mu_{io}^n, n = 1, 2, \cdots, |L|\}$.

In general, the Eq. (1) about the proposed pixel labeling model is rewritten as

$$P(\mathbf{x}|\mathbf{y}, \theta) = \frac{1}{Z(\mathbf{y}, \theta)} \prod_{i \in S} \phi_i(x_i, \mathbf{y}, \boldsymbol{\lambda}) \prod_{i \in S, j \in \eta_i^1} \phi_{ij}\left(x_i, x_j, f_{ij}, \boldsymbol{\alpha}\right) \prod_{i \in S, o \in \eta_i^2} \phi_{io}(x_i, x_o, h_{io}, \boldsymbol{\beta}) \quad (8)$$

where $\theta = \{\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ denotes the set of parameters involved in the MGCCRF model.

4 Model training and inference

In this section, we first describe the method to train a model and choose the parameters of the unary and pairwise potentials, *i.e.*, $\theta = \{\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$. Supposing $\tilde{C}$ is the set of the selected cliques in the training samples, which contains the unary clique $\tilde{C}_1$, the fine-granular pairwise clique $\tilde{C}_2^1$ and the coarse-granular pairwise clique $\tilde{C}_2^2$. If $\tilde{\mathbf{x}}$ denotes the labels of the training samples and $\tilde{\mathbf{y}}$ denotes the feature vectors, then the training samples can be expressed as $\{\tilde{\mathbf{x}}, \tilde{\mathbf{y}}\} = \{\tilde{\mathbf{x}}_c, \tilde{\mathbf{y}}_c\}_{c \in \tilde{C}}$. We train the proposed model based on maximum likelihood parameter estimation. The objective function of the maximum log-likelihood is as follows

$$\begin{aligned} J(\theta) &= \log P\left(\tilde{\mathbf{x}} \middle| \tilde{\mathbf{y}}; \theta\right) \\ &= \sum_{c \in \tilde{C}} \log \psi_c\left(\tilde{\mathbf{x}}_c, \tilde{\mathbf{y}}, \theta\right) - logZ\left(\tilde{\mathbf{y}}, \theta\right) \end{aligned} \quad (9)$$

However, the exact estimation of θ is intractable in general due to the combinatorial size of the label space in the computing partition function. In principle, the partition function $Z(\tilde{\mathbf{y}}, \theta)$ can be approximated by Markov Chain Monte Carlo (MCMC) sampling technique. However, the method is prohibitively impractical in computation. Another related approach is to estimate the parameters locally. Pseudo-likelihood estimation is a classical local estimation method. The piecewise training framework [26, 27] retains the computational efficiency of pseudo-likelihood, and even has much better accuracy. At the same time, the piecewise method is available for the CRF model with multi-granular contextual information due to the dependency between the context descriptor in the pairwise potential and the pixel classifier in the unary potential. In this paper, the piecewise training framework is adopted to choose the parameters of the model.

4.1 Piecewise training

As discussed in [26], piecewise training method can minimize the upper bound on the log partition function. If the partition function $Z(\tilde{\mathbf{y}}, \theta)$ is indexed by the divided piece in the model, then $Z(\tilde{\mathbf{y}}, \theta) \leq \sum_c Z_c(\tilde{\mathbf{y}}, \theta_c)$ where c belongs to the set of the selected cliques $\tilde{C}$, θ_c are the parameters of the cth piece and $Z_c(\theta_c)$ is the partition function for a model containing only the cth piece. Replacing $Z(\tilde{\mathbf{y}}, \theta)$ with $\sum_c Z_c(\theta_c)$ in the objective function, we get a lower bound on the conditional likelihood, which is maximized during piecewise learning. The above conclusion indicates the intuition of piecewise training demonstrated in [27]. If each factor $\psi_c(\tilde{\mathbf{x}}_c, \tilde{\mathbf{y}})$ of the objective function can on its own accurately predict $\tilde{\mathbf{x}}_c$ from $\tilde{\mathbf{y}}$, the prediction of the

global model will also be accurate. So the objective function in (9) is approximated in piecewise training framework as

$$J_{PT}(\theta) = \sum_{c \in \tilde{C}} \log \frac{\psi_c\left(\tilde{\mathbf{x}}_c, \tilde{\mathbf{y}}, \theta\right)}{\sum_{\mathbf{x}_c} \psi_c\left(\mathbf{x}_c, \tilde{\mathbf{y}}, \theta\right)} \tag{10}$$

where $c \in \tilde{C}$ is a graph factor composed of a set of sites and $\tilde{C}$ is the set of all graph factors. In this paper, the MGCCRF model is divided into pieces corresponding to the different terms in (8) so that a piece of the model is a factor of the objective function. Then c is just a clique in the set $\tilde{C} = \left\{\tilde{C}_1, \tilde{C}_2^1, \tilde{C}_2^2\right\}$. Consequently, the factor $\psi_c(\tilde{\mathbf{x}}_c, \tilde{\mathbf{y}}, \theta)$ of objective function is exactly the potential defined on clique c.

Figure 4 shows the factor graph of the MGCCRF model. Each black square represents a term in (8) and each circle represents a latent variable. Terms are connected to all variables that they depend on. The coarse-granular context descriptor has the important dependence on the coarse-granular neighbor site. However, it is not directly regarded as a term of Eq. (8). So there is the dotted line between the coarse-granular context descriptor and the coarse-granular neighbor site.

Each of these pieces is then trained independently, as if it was the only term in the conditional model. For example, if applying piecewise training to the MGCCRF model of Fig. 4, the parameters are estimated by maximizing the conditional likelihood in each of the three models in Fig. 5. In each case, only the factors in the model that contain the relevant parameter are retained. The objective function of the log-likelihood can be further rewritten under the piecewise training framework

$$\begin{aligned} J_{\mathrm{PT}}(\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = & \sum_{i \in \tilde{C}_1} \log \frac{\exp\left\{\phi_i\left(\tilde{x}_i, \tilde{\mathbf{y}}, \boldsymbol{\lambda}\right)\right\}}{\sum_{x_i} \exp\{\phi_i(x_i, \mathbf{y}, \boldsymbol{\lambda})\}} + \sum_{(i,j) \in \tilde{C}_2^1} \log \frac{\exp\left\{\phi_{ij}\left(\tilde{x}_i, \tilde{x}_j, f_{ij}, \boldsymbol{\alpha}\right)\right\}}{\sum_{x_i, x_j} \exp\left\{\phi_{ij}\left(x_i, x_j, f_{ij}, \boldsymbol{\alpha}\right)\right\}} \\ & + \sum_{(i,o) \in \tilde{C}_2^2} \log \frac{\exp\left\{\phi_{io}\left(\tilde{x}_i, \tilde{x}_o, \mu_i^o, \boldsymbol{\beta}\right)\right\}}{\sum_{x_i, x_o} \exp\left\{\phi_{io}\left(x_i, x_o, \mu_i^o, \boldsymbol{\beta}\right)\right\}} \\ = & J_{\boldsymbol{\lambda}} + J_{\boldsymbol{\alpha}} + J_{\boldsymbol{\beta}} \end{aligned} \tag{11}$$

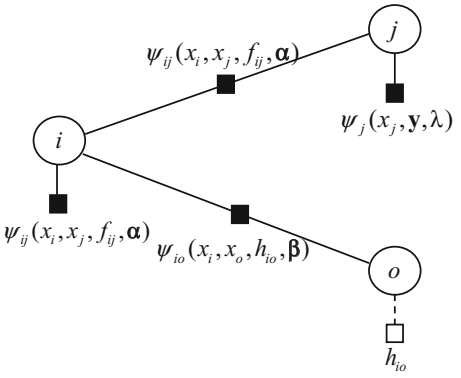


Fig. 4 The factor graph for the MGCCRF model

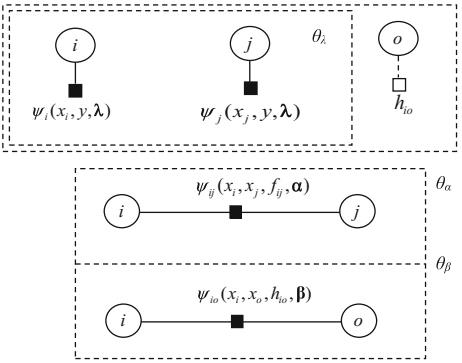


Fig. 5 Piecewise training of the MGCCRF parameters

In Eq. (11), J_{λ}, J_{α}, J_{β} are used to represent its first, second and third term for notation simplicity respectively. The MGCCRF model can be trained by independently training over every type of clique according to the piecewise training framework.

4.1.1 Training for the first term J_{λ}

Due to the normalization condition $\sum_{k=1}^{|L|} p(x_i = k|\tilde{\mathbf{y}}, \boldsymbol{\lambda}) = 1$, the denominator in J_{λ} as the normalization condition just equals to constant 1 and the derivation process is as follows:

$$\begin{aligned}\sum_{x_i} \exp\left\{\phi_i\left(x_i, \tilde{\mathbf{y}}, \boldsymbol{\lambda}\right)\right\} &= \sum_{x_i} \exp\left\{\sum_{k=1}^{|L|} \delta(x_i = k) \log p\left(x_i = k \middle| \tilde{\mathbf{y}}, \boldsymbol{\lambda}\right)\right\} \\ &= \sum_{x_i=1}^{|L|} p\left(x_i = k \middle| \tilde{\mathbf{y}}, \boldsymbol{\lambda}\right) \\ &= 1\end{aligned} \tag{12}$$

where $p(x_i|\tilde{\mathbf{y}}, \boldsymbol{\lambda})$ is modeled by the extended logistic regression classifier in Eq. (3). Now the parameter $\boldsymbol{\lambda}$ is represented as a $|L|$ -dimensional vector, and the objective function for the unary potential is rewritten by the following equation:

$$\begin{aligned}J_{\lambda} &= \sum_{\substack{i \in S \\ j \in \eta_i^1}} \phi_{ij}\left(\tilde{x}_i \middle| \tilde{\mathbf{y}}, \boldsymbol{\lambda}\right) \\ &= \sum_{\substack{i \in S \\ j \in \eta_i^1}} \log p\left(\tilde{x}_i \middle| \tilde{\mathbf{y}}, \boldsymbol{\lambda}\right) \\ &= \sum_{i \in \tilde{C}_1} \left[\sum_{k=1}^{|L|} \delta\left(\tilde{x}_i = k\right) \lambda_k^{\mathrm{T}} \tilde{\mathbf{y}}_i - \log \sum_{k=1}^{|L|} \exp\left(\lambda_k^{\mathrm{T}} \tilde{\mathbf{y}}_i\right)\right]\end{aligned} \tag{13}$$

Equation (13) is exactly the objective function for the unary potential. The intuition of unary piece training is to get parameter $\boldsymbol{\lambda}$ which maximizes J_{λ} in Eq. (11). We use the bound

optimization algorithm with a component-wise update procedure to complete the training, and the objective function J_λ for the unary potential is optimized by the iterative optimization of an even simpler surrogate function f, thus

$$\hat{\boldsymbol{\lambda}} = \underset{\boldsymbol{\lambda}}{\operatorname{argmax}}\ \mathrm{f}\left(\boldsymbol{\lambda}\middle|\boldsymbol{\lambda}^{(t)}\right) \tag{14}$$

where $\hat{\boldsymbol{\lambda}}$ represents the estimation of the parameter $\boldsymbol{\lambda}$, and $\boldsymbol{\lambda}^{(t)}$ is the optimum parameter vector at the tth iteration. The search goal of surrogate function is to meet a certain key condition, namely, $J_{\boldsymbol{\lambda}} - \mathrm{f}(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(t+1)})$ attain its minimum when $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$. This condition can ensure that the iterative procedure monotonically increases the value of the objective function, i.e., $J_{\boldsymbol{\lambda}^{(t+1)}} \geq J_{\boldsymbol{\lambda}^{(t)}}$. Then the key of estimating parameter $\boldsymbol{\lambda}$ is to find a suitable surrogate function using the bound optimization approach. In this paper, the surrogate function in [10] is adapted but without the sparse prior term, i.e.,

$$\mathrm{f}\left(\boldsymbol{\lambda}\middle|\boldsymbol{\lambda}^{(t)}\right) = \boldsymbol{\lambda}^{\mathrm{T}}\left[\mathbf{g}\left(\boldsymbol{\lambda}^{(t)}\right) - \mathbf{B}\boldsymbol{\lambda}^{(t)}\right] + \frac{1}{2}\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}\boldsymbol{\lambda} \tag{15}$$

where $\mathbf{B}$ is the bound of the Hessian matrix $\mathbf{H}(\boldsymbol{\lambda})$, $\mathbf{g}(\boldsymbol{\lambda})$ is the gradient vector of J_λ, i.e., $\mathbf{H}(\boldsymbol{\lambda}) \succeq -\frac{1}{2}\left[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{L}\right] \otimes \sum_{i \in \tilde{C}_1} \tilde{\mathbf{y}}_i (\tilde{\mathbf{y}}_i)^{\mathrm{T}} \equiv \mathbf{B}$ where $\mathbf{1} = [1, 1, \cdots, 1]^{\mathrm{T}}$ is a ($|L|$-1)-dimensional vector, and $\otimes$ denotes the Kronecker matrix product. Then $\mathbf{B}$ is a square matrix of size $d \times$ ($|L|$-1). Gradient vector $\mathbf{g}(\boldsymbol{\lambda})$ of J_λ is obtained as

$$\mathbf{g}(\boldsymbol{\lambda}) = \sum_{i \in \tilde{C}_1} \left(\tilde{\mathbf{x}}_i{}' - \mathbf{p}_i(\boldsymbol{\lambda})\right) \otimes \tilde{\mathbf{y}}_i \tag{16}$$

where $\mathbf{p}_i(\boldsymbol{\lambda}) = [p_i^1(\boldsymbol{\lambda}), \cdots, p_i^{|L|-1}(\boldsymbol{\lambda})]^{\mathrm{T}}$ and $p_i^k(\boldsymbol{\lambda}) = p\left(\tilde{x}_i^k = 1 | \tilde{\mathbf{y}}_i, \boldsymbol{\lambda}\right)$. Here the class label $\tilde{x}_i$ is represented as a "1-of-m" encoding vector $\tilde{x}_i = \left[\tilde{x}_i^1, \cdots, \tilde{x}_i^{|L|-1}\right]^{\mathrm{T}}$ such that $\tilde{x}_i^k = 1$ if $\tilde{x}_i$ corresponds to an example belonging to class k, otherwise $x_{i,k}^m = 0$. $\tilde{\mathbf{x}}_i^{'}$ denotes the vector $\left[\tilde{x}_i^1, \cdots, \tilde{x}_i^{|L|-1}\right]^{\mathrm{T}}$. Maximization of the surrogate function $\mathrm{f}(\boldsymbol{\lambda}|\boldsymbol{\lambda}^{(t)})$ leads to a simple update equation

$$\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}^{(t)} - \mathbf{B}^{-1}\mathbf{g}\left(\boldsymbol{\lambda}^{(t)}\right) \tag{17}$$

Algorithm 1 gives the iterative steps to estimate the parameter $\boldsymbol{\lambda}$ based on the bound optimization approach.

Algorithm 1 Parameter estimation based on the bound optimization approach

Input: the vector of the class label $\tilde{\mathbf{x}}_i'$, the feature vectors $\tilde{\mathbf{y}}_i$, and the MRL model $\mathbf{p}_i(\boldsymbol{\lambda})$, $i \in S$

Output: $\hat{\boldsymbol{\lambda}}$

// Computing the bound of the Hessian matrix $\mathbf{H}(\boldsymbol{\lambda})$.

1: $\mathbf{B} = -\frac{1}{2}[\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{L}] \otimes \sum_{i \in \tilde{C}_1} \tilde{\mathbf{y}}_i (\tilde{\mathbf{y}}_i)^{\mathrm{T}}$

//Computing the gradient vector $\mathbf{g}(\boldsymbol{\lambda})$ of J_λ.

2: **repeat**

3: for each $i \in S$

4: $\mathbf{g}(\boldsymbol{\lambda}) = \sum_{i \in \tilde{C}_1} (\tilde{\mathbf{x}}_i' - \mathbf{p}_i(\boldsymbol{\lambda})) \otimes \tilde{\mathbf{y}}_i$

// Computing the surrogate function instead of the objective function J_λ.

5: $\mathrm{f}(\boldsymbol{\lambda} | \boldsymbol{\lambda}^{(t)}) = \boldsymbol{\lambda}^{\mathrm{T}}[\mathbf{g}(\boldsymbol{\lambda}^{(t)}) - \mathbf{B}\boldsymbol{\lambda}^{(t)}] + \frac{1}{2}\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{B}\boldsymbol{\lambda}$

// Update the parameter $\boldsymbol{\lambda}$.

6: if $\mathrm{f}(\boldsymbol{\lambda} | \boldsymbol{\lambda}^{(t+1)}) \geq \mathrm{f}(\boldsymbol{\lambda} | \boldsymbol{\lambda}^{(t)})$ then

7: $\boldsymbol{\lambda}^{(t+1)} = \boldsymbol{\lambda}^{(t)} - \mathbf{B}^{-1}\mathbf{g}(\boldsymbol{\lambda}^{(t)})$

//Stopping the iteration condition.

7: **Until** convergence

4.1.2 Training for the second term J_α

For the fine-granular pairwise potential, the transferring probability of the labels with homogenous features has the similar structure to the MLR model. As stated above, the denominator in the second term J_α of (11) is just the constant 1. So the objective function of the fine-granular pairwise potential is similar to that of the unary potential, which can be formulated as

$$\begin{aligned} J_\alpha &= \sum_{(i,j) \in \tilde{C}_2^1} \phi_{ij}\left(\tilde{x}_i, \tilde{x}_j, f_{ij}, \alpha\right) \\ &= \sum_{(i,j) \in \tilde{C}_2^1} \log p\left(\left(\tilde{x}_i, \tilde{x}_j\right) \triangleq k \middle| f_{ij}, \alpha\right) \\ &= \sum_{(i,j) \in \tilde{C}_2^1} \left[\sum_{k=1}^{|L|+1} \delta(x_i = k)\delta(x_j = k)\alpha_{kk}^{\mathrm{T}} f_{ij} - \log \sum_{k=1}^{|L|+1} \exp\left(\alpha_{kk}^{\mathrm{T}} f_{ij}\right)\right] \end{aligned} \tag{18}$$

where $p\left((\tilde{x}_i,\tilde{x}_j)\triangleq k\middle|f_{ij},\alpha\right)$ is defined as MLR model in the form of Eq. (6). But we should point out that the term $p\left((\tilde{x}_i,\tilde{x}_j)\triangleq k\middle|f_{ij},\alpha\right)$ in (18) incorporates an additional parameter vector $\alpha_{|L|+1,|L|+1}$ compared with (6). The parameter $\boldsymbol{\alpha}$ is represented as a $(|L|+1)$-dimensional vector, i.e. $\boldsymbol{\alpha}=(\alpha_{11},\alpha_{22},\cdots,\alpha_{|L|,|L|},\alpha_{|L|+1,|L+1|})$, and $\alpha_{|L|+1,|L|+1}$ is fixed as 0 in the whole training procedure. That is used only for the compact denotation of the equation.

Then (18) is exactly the objective of the estimation of the parameter $\boldsymbol{\alpha}$. The obvious conclusion is that the Algorithm 1 can be used to estimate the parameter $\boldsymbol{\alpha}$ for the fine-granular pairwise potential in the same way as estimating the parameter $\boldsymbol{\lambda}$.

4.1.3 Training for the third term J_β

Firstly, we adopt the sparse representation to define h_{io} which describes the spatial co-occurrence context of different classes in the coarse-granular neighborhood. The detailed steps to measure h_{io} are as follows:

Step 1: We compute the classification cost which describes the likelihood of assigning label classes to each pixel. Let r denote arbitrary label belonging to the label set $\boldsymbol{L}$, we calculate the average value $\overline{y}_r$ of the features for label r from all training samples. The classification cost of the pixel is calculated by using the following equation

$$U(x_i=n)=1-\frac{K\left(y_i,\overline{y}_n\right)}{\sum_{r\in\boldsymbol{L}}K\left(y_i,\overline{y}_r\right)} \tag{19}$$

where $K(y_i,\overline{y}_r)$ denotes the intersection kernel between two feature vectors y_i and $\overline{y}_r$. To reduce the computation complexity, we map feature vectors into a high-dimensional space $\varphi(y_i)$ where the inner product approximates the intersection kernel.

$$K(y_i,y_r)\approx <\varphi(y_i),\varphi(y_r)> \tag{20}$$

Step 2: We obtain the initial semantic knowledge of the observed data, which is represented by the pixel classification likelihood maps as follows

$$l(p,n)=\frac{1}{1+\delta(x_p=n)\exp(-U_p(n))} \tag{21}$$

where $U_p(n)$ is the cost of assigning label n to pixel p as in Eq. (19). These classification maps in the label set grant us naturally sparse representation of semantic information without an extra sparse coding step.

Step 3: For arbitrary generalized neighbor site o of the site i, we compute its coarse-granular context descriptor $h_{io}=\{\mu_{io}^n,n=1,2,\cdots,|L|\}$ by max pooling of the classification likelihood maps

$$\mu_{io}^n=\max_{p\in o}l(p,n) \tag{22}$$

Then using the coarse-granular pairwise potential defined in (7), the third term J_β in (11) can be formulated as a closed form

$$J_\beta = \sum_{(i,j)\in\tilde{C}_2^2} \log p\left(\tilde{x}_i, \tilde{x}_o \middle| h_{io}, \boldsymbol{\beta}\right) \tag{23}$$

where

$$p\left(\tilde{x}_i = k, \tilde{x}_o = l \middle| h_{io}, \boldsymbol{\beta}\right) = \frac{\exp\left(\beta_{kl}^{\mathrm{T}} \mu_{io}^{l}\right)\delta(k\neq l)}{\sum_{k=1}^{|L|}\sum_{l=1}^{|L|}\exp\left(\beta_{kl}^{\mathrm{T}}\mu_{io}^{l}\right)} \tag{24}$$

It shows that Eq. (24) acts also as an MLR model with L^2 classes. So Algorithm 1 is also suitable for estimating the parameter $\boldsymbol{\beta}$. It is clearly seen that the computation of parameter $\boldsymbol{\beta}$ does not cause more computational cost when the number of the classes are less than the dimension of the feature vector, because μ_{io}^{l} is a 1-dim parameter vector.

4.2 Inference for labeling an image

Given a new image y, the inference procedure is to find the optimal label configuration x over the image sites. The optimality is evaluated with respect to a particular cost function. The widely-used criterion for inferring labels from the posterior distribution is MPM adopted in this paper. The MPM criterion, which maximizes the expected number of the correctly labeled sites by taking the modes of posterior marginals:

$$x_i^* = \underset{x_i}{\operatorname{argmax}}\, P(x_i|y), \quad \forall i\in S \tag{25}$$

The computation of the MPM requires marginalization over a large number of variables, which is generally NP-hard. At the same time, evaluating $P(x_i|y)$ in our model is intractable due to its dependent structure. To tackle these difficulties, Gibbs sampling is used based on its simplicity and fast convergence. The basic idea in Gibbs sampling is to make a separate probabilistic choice for each of the parameters in the model not to probabilistically pick the next state all at once. A reasonable initial point for the sampling can be obtained by considering the outputs of the MLR classifier.

5 Experiments

The experiments were all run in MATLAB 2011a environment on dual-core CPU (T2390 1.8GHz) and 2 G memory. We first present the datasets used and the extracted features followed by, in section 5.2, the results of automatic image pixel labeling for the proposed model, in which the convergence performance of training method, and the quantitative and qualitative performance of the proposed labeling model are evaluated.

5.1 Data sets and features extraction

MGCCRF model was applied to two natural image datasets. The first dataset is called URS database which is a subset of the database that consists of the natural scenes from a collection of public image datasets: LabelMe, PASCAL, and GC [5]. 500 images were selected , and these images were labeled into 9 classes as 'sky', 'bird', 'water', 'flower', grass', 'face', 'tree', 'body' and 'boat'. The second dataset is a 375-image subset of the LHI database, consisting of 15 types of objects ('sheep', 'car', 'bike', 'airplane', 'horse', 'cow', 'grass', 'tree', 'building', 'sky', 'rhinoceros', 'mountain', 'elephant', 'road' and 'water') [31]. The selection criterion is that each image contains 2–5 semantic classes. The semantic classes for per image are not all background classes, and contains at least one foreground structured object.

Without losing generalization, the images in the experiments were selected by taking into account the following conditions: camera viewpoint, little occlusions, multi-objects, lighting conditions, object pose, deformation, and scale variance. Moreover, the pixels labeled as 'void' were not considered during evaluating our model for the direct comparison of the quantitative result.

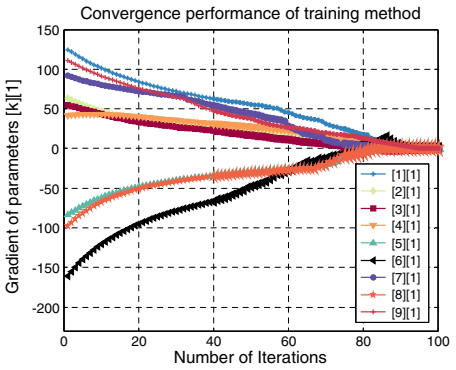
In the experiments, all images were rescaled to a resolution of 240×180 pixels. A set of image feature vectors y_i at each image site i were extracted, including CIELAB color and textures. In this paper, each site corresponds to a single image pixel which is represented by 35 dim image feature vectors. For color information, the RGB values were transformed into CIELAB color space due to its perceptual uniformity. Since Gabor function is similar to the biological role of the human visual system, the texture information was extracted by a filter bank of Gabor wavelets at 8 orientations and 4 scales, which is robust for object shape and category appearance classification. Before training, all pixel feature vectors were normalized to give zero mean and unit variance so that convergence of the training parameters was easier to achieve.

5.2 Performance evaluation

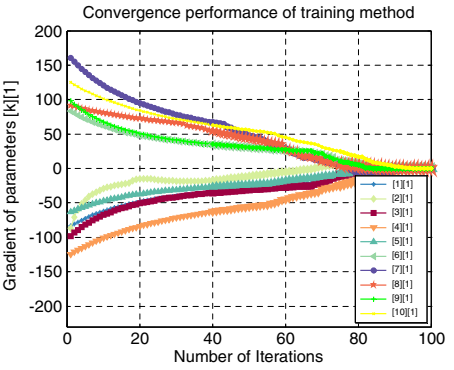
These two datasets use the same split setting. They are split randomly into roughly 50 % for training, 25 % for validation and 25 % for test, while ensuring approximately proportional contributions from each class. This section describes the performance evaluation for our model on the image datasets.

5.2.1 Convergence performance of the training method

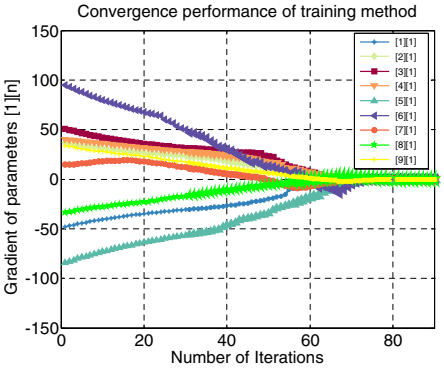
At first, the convergence performance of the training method was evaluated through experiments over the URS dataset. The parameters $\theta = \{\boldsymbol{\lambda}, \boldsymbol{\alpha}, \boldsymbol{\beta}\}$ of MGCCRF were separately learned by the efficient piecewise training method. The convergence property of the training method is illustrated in Fig. 6 through the plots of gradients with change of the number of iterations. The images in the URS dataset include 9 classes of labels, and each site is represented by 35 dim image feature vectors. So there are in total 315 elements ($\{\lambda_{kd}, k = 1, \cdots, 9, d = 1, \cdots, 35\}$), 700 elements ($\{\alpha_{kd}, k = 1, \cdots, 10, d = 1, \cdots, 70\}$) and 81 elements ($\{\beta_{mn}, m = 1, \cdots, 9, n = 1, \cdots, 9\}$) respectively in parameters $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$. It is impossible to demonstrate the gradients corresponding to all the parameters. Without losing generalization, the presented convergence behaviors are only the gradients of parameters $\boldsymbol{\lambda}$, $\boldsymbol{\alpha}$ corresponding to the first dimension of the feature vectors and the gradients of parameters $\boldsymbol{\beta}$ corresponding to the first dimension of the label classes.



(a) The plots of the gradients of parameters $\boldsymbol{\lambda}$



(b) The plots of gradients of parameters $\boldsymbol{\alpha}$



(c) The plots of gradients of parameters $\boldsymbol{\beta}$

Fig. 6 The convergence performance of the training method. **a** The plots of the gradients of parameters $\{\lambda_{k1}, k=1,\cdots,9\}\}$; **b** The plots of the gradients of parameters $\{\alpha_{k1}, k=1,\cdots,10\}$; **c** The plots of the gradients of parameters $\{\beta_{1n}, n=1,\cdots,9\}$

Figure 6 shows the convergence behaviors of the training processes. All the training processes of the parameters show convergences with less than one hundred iterations. Since the estimations of those three parameters in the MGCCRF model are independent, the whole training process can be implemented in a simple parallel way to accelerate the training times. Training our model is extremely fast which benefits from the piecewise training based on the bound optimization algorithm. The parallel operations of the piecewise training reduce the whole training. For the training set in the URS dataset, the whole training procedure implemented by the bound optimization algorithm in a parallel manner took about 86 min.

5.2.2 Quantitative results

Table 1 shows the comparison of the average pixel wise accuracy on the URS 9-class and the LHI 15-class datasets. It displays the quantitative experimental results comparing MGCCRF with 3 other methods which also accomplish the pixel labeling in the image. H-CRF [24] is a hierarchical random field model that allows integration of features computed at different levels of the quantization hierarchy. The approach enables the contextual priors defined over multiple image quantization in the framework to obtain good results. DS [5] is defined in terms of a

Table 1 Classification accuracies of different models

Database	Accuracy			
	MLR classifier	H-CRF [24]	DS [5]	MGCCRF
URS	62.4 %	70.5 %	73.6 %	80.9 %
LHI	53.9 %	66.5 %	71.1 %	81.7 %

unified energy function over scene appearance and structure. It adopts an effective inference technique for optimizing this energy function. The DS framework provides a basis on which many valuable extensions can be layered. Besides multi-class image segmentation, it can be applied to the task of 3D reconstruction.

The evaluation criterion is the same for the four methods based on the average pixel accuracy. The results obtained are comparable to 3 other methods on the two datasets, as shown in Table 1. MLR is a simple global classifier with the multinomial logistic regression whose expression is shown in Eq. (3). It can be seen that the labeling performance of MLR classifier is the worst because the classifier can be easily fooled without the contextual information in the image. The H-CRF and DS perform better but they may mislead the labeling due to the capture of only the local context. It demonstrates that our MGCCRF model generates more accurate labeling than the three other methods. The average pixelwise labeling accuracy of MGCCRF is 80.9 % and 81.7 % on the two datasets respectively, which implies the multi-granular contextual information is effectively captured.

To further measure the performance of our approach, Tables 2 and 3 are used to illustrate the confusion matrix by applying MGCCRF model on the testing data in the two datasets, in which the accuracy values are computed as the percentage of image pixels assigned to the correct class labels. These tables show that the errors in our model are consistent across the classes. The average pixelwise labeling accuracies of the objects are 81.73 % and 80.89 % on the two datasets, which show the advantage of CRF model with the multi-granular context configuration. It is clearly seen that some objects with different surrounds can be correctly recognized as shown in Table 2, e.g. “airplane” vs “car”, “bike” vs “rhinoceros”, “building” vs “cow”, and “sky”, “water” vs “mountain”. This implies that the large range contextual information is efficiently captured to improve the labeling accuracy. It is also discovered that some objects such as “sheep”, “horse” and “cow” have relatively high confusions due to their structure similarities simultaneously surrounded by the similar contexts.

Moreover Table 3 shows similar object recognition behaviors. For example, “flower” vs “water”, “body” vs “bird”, and “face” vs “grass” have lower confusion ratios. Meanwhile, “face” and “flower” are confusing because they have similar context configuration in the hand-labeled ‘ground truth’ images.

5.2.3 Qualitative results

Figures 7 and 8 show the example results of pixel labeling on the two datasets. Figures 7a and 8a are the original images, and each image contains 2 ~ 5 objects. The hand-labeled “ground truth” images are shown in Figs. 7b and 8b. The hand-labeled images suffer from another drawback. A significant number of pixels in these images have not been assigned any label. These unlabeled pixels generally occur at object boundaries and are critical in evaluating the

Table 2 Confusion matrix in percentage for LHI dataset

Class	Car	Sheep	Tree	Horse	Plane	Bike	Cow	Grass	Buil-	Sky	Water	El-nt	Road	Rhin-	Moun-
Car	0.85	0	0.02	0	0.01	0.02	0	0.02	0.03	0	0.01	0	0.04	0	0
Sheep	0.01	0.81	0.01	0.03	0.02	0	0.02	0.03	0	0.02	0	0.03	0.01	0	0.01
Tree	0.01	0.02	0.71	0.03	0	0.01	0	0.06	0.04	0.05	0	0.02	0.01	0.01	0.03
Horse	0	0.02	0.04	0.78	0.01	0	0.02	0.04	0.02	0.02	0	0.03	0.01	0	0.01
Plane	0	0.01	0	0	0.84	0	0	0.06	0.02	0.02	0	0	0.04	0	0.01
Bike	0.02	0.02	0.03	0.01	0	0.78	0.01	0.04	0.05	0.01	0	0	0.03	0	0
Cow	0	0.04	0.01	0.04	0.01	0	0.78	0.08	0	0.02	0	0	0.01	0	0.01
Grass	0	0	0.02	0.01	0	0	0.01	0.96	0	0	0	0	0	0	0
Building	0.02	0	0.04	0	0.04	0.01	0	0.02	0.82	0.03	0	0	0.02	0	0
Sky	0	0	0.02	0	0	0	0	0	0.01	0.95	0	0	0.01	0.01	0
Water	0	0	0.05	0.02	0.01	0	0.01	0.03	0.02	0.13	0.64	0.01	0.06	0.02	0
Elephant	0	0.01	0.04	0.01	0.01	0	0.02	0.04	0	0.01	0.02	0.82	0.02	0	0
Road	0.01	0	0.01	0	0	0.01	0.01	0.02	0.01	0	0	0	0.93		0.01
Rhinoceros	0.03	0	0.02	0.03	0	0	0	0.07	0	0	0.05	0.01	0	0.78	0
Mountain	0	0	0.04	0.01	0.03	0.01	0	0.06	0	0.03	0	0	0.01	0	0.81

Table 3 Confusion matrix in percentage for URS dataset

Class	Sky	Bird	Flower	Grass	Tree	Face	Water	Body	Boat
Sky	0.96	0.01	0	0	0	0	0.03	0	0
Bird	0.05	0.87	0.05	0.02	0	0	0	0	0.01
Flower	0.03	0.04	0.71	0.06	0.05	0.09	0	0	0.02
Grass	0	0.02	0.03	0.93	0.02	0	0	0	0
Tree	0.03	0.04	0.02	0.05	0.82	0	0	0.04	0
Face	0	0.01	0.05	0	0.02	0.88	0	0.04	0
Water	0.04	0.03	0	0.04	0.06	0.02	0.71	0.03	0.07
Body	0.03	0	0.04	0.06	0.05	0.08	0.02	0.67	0.05
Boat	0	0.06	0.05	0.03	0.05	0	0.08	0	0.73

accuracy of an image labeling algorithm. It should be noted that obtaining an accurate and fine segmentation of the objects is important for the image labeling in computer vision.

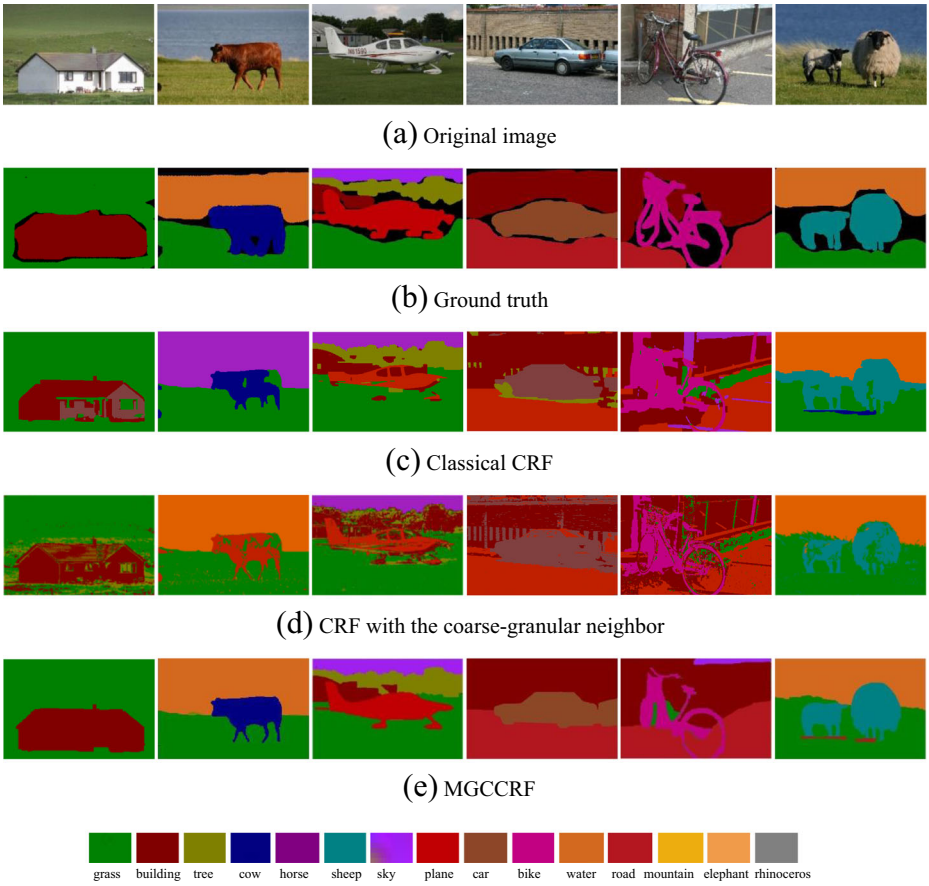


Fig. 7 Labeling results for the LHI database. **a** Original image. **b** Ground truth. **c** Classical CRF. **d** CRF with the coarse-granular neighbor. **e** MGCCRF

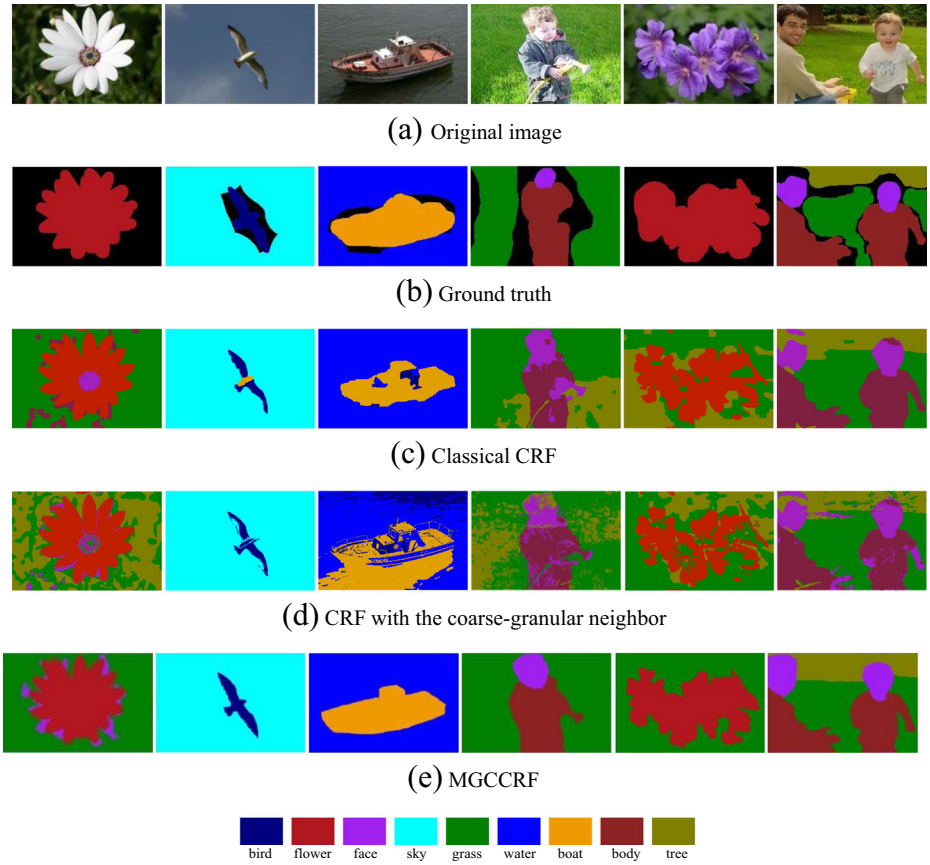


Fig. 8 Labeling results for the URS database. **a** Original image. **b** Ground truth. **c** Classical CRF. **d** CRF with the coarse-granular neighbor. **e** MGCCRF

If the pairwise potential contains only the fine-granular neighbor, the labeling model corresponds to adopt the classical CRF. Figures 7c and 8c show the labeling results with the classical CRF model. It maintains a relatively smooth configuration of the semantic labels. However, it can be mislabeled inside the surface of the objects because the classical CRF model captures only the context in the small scale. For example, "water" is mislabeled as "sky", or "flower" is confused as "face".

When the labeling model contains only the coarse-granular neighbor, the pixel labeling results lacks the statistics of the label consistency between the adjacent pixels. As showed in Figs. 7d and 8d, these labeled results have many discontinuous points in the object surface. Moreover, the boundaries of the objects are not smooth. The discontinuous and sporadic results are significantly different from the human visual perception.

The multi-granular contextual information is integrated into the proposed model whose image labeling results on the two datasets are shown in Figs. 7e and 8e, respectively. In order to get a good estimate of our algorithm accuracy, the multi-granular contextual information is integrated into the CRF model so that the accurate object boundaries are preserved in the fine-granular neighborhood and the wrong predictions from the local classification are corrected by the coarse-granular co-occurrence contextual information.

As shown in Figs. 7 and 8, the qualitative results illustrate the fact that the labeling results obtained from our model always look better than the hand-labeled images. MGCCRF recognizes the unlabeled pixels in the hand-labeled images. Moreover, the labeling results generated by MGCCRF produce not only the smooth label configurations but also the continuous object surface, which is also better than the classical CRF model. Our model can handle the contextual information provided by the multi-granular contexts to generate more reasonable labeling results. Obviously, it is closer to the observations of human visual system.

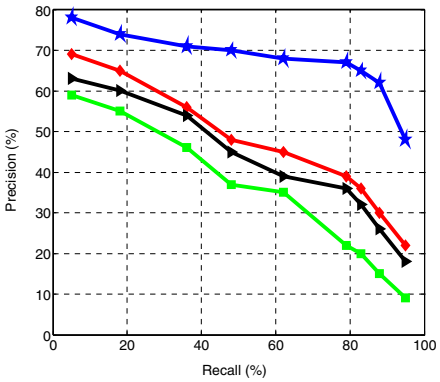
5.3 Improvement for object recognition

Another outstanding property of the MGCCRF model is the improvement for object recognition due to preserving the accurate object boundaries and producing the continuous object surface. In order to further measure the performance of MGCCRF for object recognition, we continue to test the proposed method on the PASCAL VOC 2012 dataset which is accepted as currently one of the most popular object-class benchmarks. This dataset has only ground truth labels for the pixels of individual objects, and the residual regions are labeled with the placeholder "background". We selected 100 images mainly including 5 classes (e.g., "cow", "sheep" , "car", "boat" and "bird") to do the experiments which depict the quantitative results of object recognition. In order to ensure approximately proportional contributions from each class, these images were split randomly into roughly 50 % for training, 25 % for validation and 25 % for test.

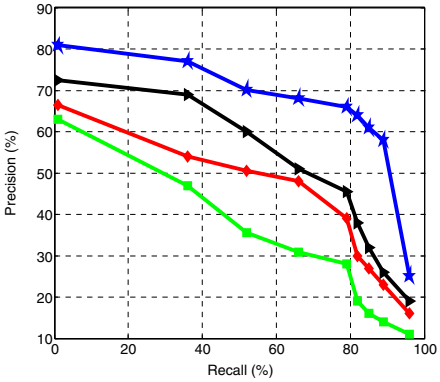
Figure 9 shows the comparable quantitative results of object recognition with 4 methods. Besides MGCCRF, the other methods can also address the task of pixel labeling to accomplish the object recognition. MRSA [4] exploits shape information via masking convolutional features which are applicable for semantic segmentation. Then the framework is generalized for recognition of joint object and stuff by modifying the underlying probabilistic distributions of the training samples. FCN [14] builds fully convolutional networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. The classical CRF model uses the small-scale context information captured by the local interactions of pixels. The panel of Figure 9 shows the precision-recall (PR) curves of object recognition for the 5 classes with MGCCRF and 3 other methods. The precision drops whereas the recall increases in the most common case, thus the results of object recognition are said to be effective if the precision values are higher at the same recall ones. It is clear that our model achieve remarkable improvement to the performance since the object boundaries is accurately preserved and the object surface is continuously labeled by the fine-granular contextual information. Simultaneously, most labeling errors in the fine-granular model can be eliminated by the coarse-granular contextual information.

6 Conclusion

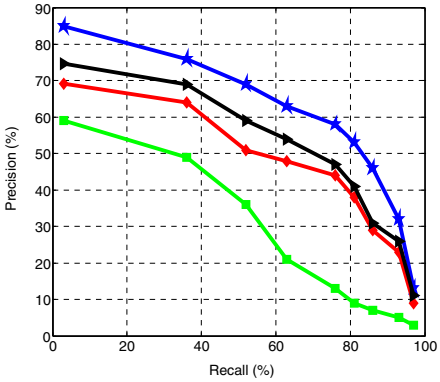
In this paper the MGCCRF model for pixel labeling is proposed which is capable of integrating the multi-granular contextual information into the CRF framework. The model is a combination of multi-granular components, each providing the contextual information of pixel labeling from different sizes of the neighborhood systems. The fine-granular contextual information is useful for preserving the accurate object boundaries and producing the continuous object surface, and the coarse-granular contextual information describes the spatial co-



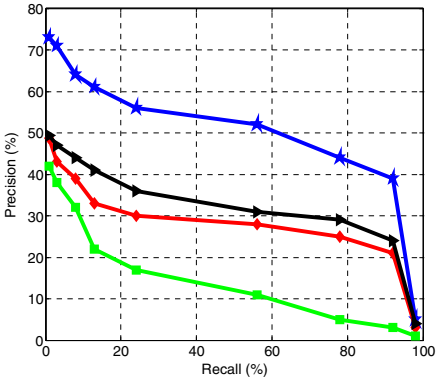
(a) PR curves of "cow" class



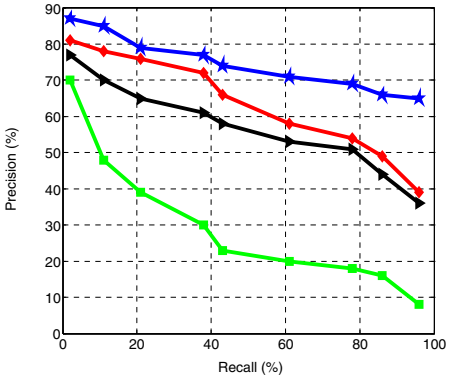
(b) PR curves of "sheep" class



(c) PR curves of "car" class



(d) PR curves of "boat" class



(e) PR curves of "bird" class

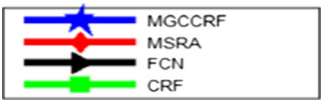


Fig. 9 Comparable quantitative results for object recognition. **a** PR curves of "cow" class. **b** PR curves of "sheep" class. **c** PR curves of "car" class. **d** PR curves of "boat" class. **e** PR curves of "bird" class

occurrence relationship among the semantic classes to improve the performance of object recognition. Experimental results indicate that the MGCCRF model can efficiently capture more context information to enhance the recognition of objects, and has proved to be able to

improve the efficiency and effectiveness of our model in pixel labeling. The experimental results also indicate how to identify the objects with similar structure surrounded by similar context is still an extremely challenging task in computer vision. These results confirm that the feature research should focus on adding the object structures based on the visual system to the proposed method in order to obtain better recognition capability in pixel labeling. Meanwhile, the feature selection algorithm which selects the most discriminative features of each object to maintain a low computational complexity is also the focus of our research.

Acknowledgments This work was supported by Innovation Foundations of Education for Graduate Students of Shanxi Province (No. 2015BY23).

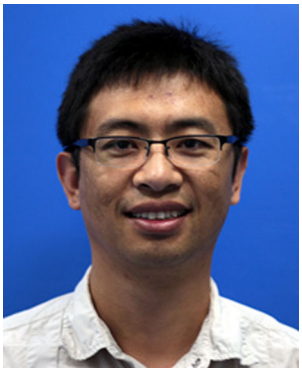
References

1. Bao BK, Li T, Yan S (2012) Hidden-concept driven multilabel image annotation and label ranking. IEEE Trans Multimed 14(1):199–210
2. Chen Y, Bi J, Wang JZ (2006) MILES: multiple-instance learning via embedded instance selection. IEEE Trans Pattern Anal Mach Intell 28(12):1931–1947
3. Chen G, Song Y, Wang F et al (2008) Semi-supervised multi-label learning by solving a sylvester equation. In: SIAM International Conference on Data Mining: 410–419
4. Dai J, He K, Sun J (2015) Convolutional feature masking for joint object and stuff segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 3992–4000
5. Gould S, Fulton R, Koller D (2009) Decomposing a scene into geometric and semantically consistent regions. In: IEEE 12th International Conference on Computer Vision: 1–8
6. He X, Zemel RS, Carreira-Perpiñán MÁ (2004) Multiscale conditional random fields for image labeling. In: IEEE computer society conference on Computer vision and pattern recognition (2): 695–702
7. Heili A, Lopez-Mendez A, Odobez JM (2014) Exploiting long-term connectivity and visual motion in CRF-based multi-person tracking. IEEE Trans Image Process 23(7):3040–3056
8. Huang Q, Han M, Wu B et al (2011) A hierarchical conditional random field model for labeling and segmenting images of street scenes. In: IEEE Conference on Computer Vision and Pattern Recognition: 1953–1960
9. Kae A, Sohn K, Lee H et al (2013) Augmenting CRFs with Boltzmann machine shape priors for image labeling. In: IEEE Conference on Computer Vision and Pattern Recognition: 2019–2026
10. Krishnapuram B, Carin L, Figueiredo MAT et al (2005) Sparse multinomial logistic regression: fast algorithms and generalization bounds. IEEE Trans Pattern Anal Mach Intell 27(6):957–968
11. Kumar S, Hebert M (2003) Man-made structure detection in natural images using a causal multiscale random field. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1): 119–126
12. Kumar S, Hebert M (2005) A hierarchical field framework for unified context-based classification. In: IEEE International Conference on Computer Vision: 1284–1291
13. Liu Y, Jin R, Yang L (2006) Semi-supervised multi-label learning by constrained non-negative matrix factorization. In: Proceedings of the national conference on artificial intelligence 21(1): 421
14. Long J, Shelhamer E, Darrell T (2015) In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition: 3431–3440
15. Mensink T, Verbeek J, Csurka G (2013) Tree-structured CRF models for interactive image labeling. IEEE Trans Pattern Anal Mach Intell 35(2):476–489
16. Nasierding G, Tsoumakas G, Kouzani AZ (2009) Clustering based multi-label classification for image annotation and retrieval. In: IEEE International Conference on Systems, Man and Cybernetics: 4514–4519
17. Neher R, Srivastava A (2005) A Bayesian MRF framework for labeling terrain using hyperspectral imaging. IEEE Trans Geosci Remote Sens 43(6):1363–1374
18. Nguyen TM, Wu QMJ (2012) Gaussian-mixture-model-based spatial neighborhood relationships for pixel labeling problem. IEEE Trans Syst Man Cybern 42(1):193–202

19. Nowozin S, Lampert CH (2011) Structured learning and prediction in computer vision. Foundations and Trends in Computer Graphics and Vision 6(3–4):185–365
20. Poggi G, Scarpa G, Zerubia JB (2005) Supervised segmentation of remote sensing images based on a tree-structured MRF model. IEEE Trans Geosci Remote Sens 43(8):1901–1911
21. Posner I, Cummins M, Newman P (2009) A generative framework for fast urban labeling using spatial and temporal context. Auton Robot 26(2–3):153–170
22. Provost JN, Collet C, Rostaing P et al (2004) Hierarchical Markovian segmentation of multispectral images for the reconstruction of water depth maps. Comput Vis Image Underst 93(2):155–174
23. Roig G, Boix X, De la Torre F et al (2011) Hierarchical crf with product label spaces for parts-based models. In: IEEE International Conference on Automatic Face & Gesture Recognition and Workshops: 657–664
24. Russell C, Kohli P, Torr PHS (2009) Associative hierarchical CRFs for object class image segmentation. In: IEEE 12th International Conference on Computer Vision: 739–746
25. Shotton J, Winn J, Rother C et al (2006) Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In: Proceedings of the European on Computer Vision: 1–15
26. Sutton C, McCallum A (2005) Piecewise training of undirected models. In: Proceedings of UAI: 568–575
27. Sutton C, McCallum A (2007) Piecewise pseudolikelihood for efficient training of conditional random fields. In: Proceedings of the 24th international conference on Machine learning: 863–870
28. Wilson R, Li CT (2003) A class of discrete multiresolution random fields and its application to image segmentation. IEEE Trans Pattern Anal Mach Intell 25(1):42–56
29. Xu X, Frank E (2004) Logistic regression and boosting for labeled bags of instances. In: Proceedings of the Pacific Aisa Conference on Knowledge Discovery and Data Mining: 272–281.
30. Yang MY, Förstner W (2011) A hierarchical conditional random field model for labeling and classifying images of man-made scenes. In: IEEE International Conference on Computer Vision Workshops: 196–203
31. Yao B, Yang X, Zhu SC (2007) Introduction to a large-scale general purpose ground truth database: methodology, annotation tool and benchmarks. Energy Minimization Methods in Computer Vision and Pattern Recognition: 169–183
32. Yu Y, Pedrycz W, Miao D (2013) Neighborhood rough sets based multi-label classification for automatic image annotation. Int J Approx Reason 54(9):1373–1387
33. Zhang ML, Zhou ZH (2007) ML-KNN: a lazy learning approach to multi-label learning. Pattern Recogn 40(7):2038–2048
34. Zhong P, Wang R (2006) Object detection based on combination of conditional random field and markov random field. In: IEEE 18th International Conference on Pattern Recognition: 160–163

Jie Zhao received the B.S. degree and M.S. degree in Electronic information from Taiyuan University of Technology, China, in 2002 and 2006 respectively. She is now a lecturer of Taiyuan University and is pursuing the PhD degree at Taiyuan University of Technology. Her research interests include image processing, artificial intelligence and pattern recognition.

Gang Xie received the B.S. degree and Ph.D. degrees in control theory and control engineering from Taiyuan University of Technology, China, in 1994 and in 2006 respectively. He is currently a professor with the College of Information Engineering at Taiyuan University of Technology. His research interests cover intelligent information processing, image processing and granular computing. He holds five invention patents, and three scientific and technological achievements appraisal. He attained six provincial science and technology award, and has published more than 100 papers.



Jiwan Han received the Ph.D. degrees in computer vision and machine learning from University of Hertfordshire in 2010. He is now an image processing specialist in the National Plant Phenomics Centre of Aberystyth University. His research interests includes computer vision and machine learning.