CRF-Based Simultaneous Segmentation and Classification of High-Resolution Satellite Images

Weihong Cui, Guofeng Wang, Chenyi Feng, Yiwei Zheng and Jonathan Li

Abstract Scale selection and uncertainty of image segmentation is still an intractable problem which influences the image classification results directly. To solve this problem, we adopt a CRF (Conditional Random Field)-based method to do segmentation and classification simultaneously. In this method, using probabilistic graphical model, we construct a three-level potential function which includes the pixels, the objects, and the link among the pixels and the objects to model their relations. We transform it to an optimization problem and use the graph cut algorithm to get the optimal solution. This method can refine the segmentation while getting good classification result. We do some experiments on the GF-1 high spatial resolution satellite images. The experiment results show that it is an effective way to improve the classification accuracy, avoid the boring segmentation scale and parameters selection and will highly improve the efficiency of image interpretation.

Keywords Object-oriented · High resolution image · Segmentation · Classification · CRF

W. Cui (✉)
School of Remote Sensing and Information Engineering,
Wuhan University, Wuhan, China
e-mail: whcui@whu.edu.cn; w28cui@uwaterloo.ca

W. Cui
Collaborative Innovation Center for Geospatial Technology, Wuhan, China

W. Cui · J. Li
Mobile Mapping Lab, University of Waterloo, Waterloo, Canada
e-mail: junli@uwaterloo.ca

G. Wang
China Highway Engineering Consulting Corporation, Beijing, China

C. Feng · Y. Zheng
Xi'an University of Science and Technology, Xi'an, China
e-mail: 491016316@qq.com

Y. Zheng
e-mail: 276014260@qq.com

S. Pirasteh and J. Li (eds.), *Global Changes and Natural Disaster Management: Geo-information Technologies*, DOI 10.1007/978-3-319-51844-2_3

1 Introduction

LULC (Land Use and Land Cover) information extraction is a key step of national geographic state surveying and monitoring. To get more detailed information of the LULC, about ten first level classes, 46 second level classes, and some third level classes are proposed by NASG of China. Abundant features are offered by high spatial resolution images, at the same time, large intra-class variations and low interclass variation exist in them. The situation makes even more challenge on high spatial resolution images processes and uses. The familiar eCongnition software has offered a series of object-oriented image classification tool, while because of the uncertainty of segmentation which makes it is very difficult to be used in the surveying project. So, in real applications, the operators should use high-resolution remote sensing images such as GF-1 or UAV images and do heavy manual interpretation work. Many facts proved that the segmentation problem is still very intractable. Researchers have proposed lots of image segmentation methods, such as MST (Felzenszwalb and Huttenlocher 2004), mean shift (Comaniciu and Meer 2002), watershed (Beucher and Meyer 1992), graph cut (Boykov and Jolly 2001), etc. Each of these segmentation methods has to set the segmentation parameters which may lead to over-segmentation or under-segmentation. How to improve the segmentation accuracy? Can we combine the segmentation with classification together to get the optimized segmentation boundaries while improving the classification accuracy? Machine learning and computer vision technology make it possible. Accumulation of manual interpretation results can provide massive training samples which are very useful resource for learning the segmentation and classification model. Multi-scale features and context features are very important clues to recognize the objects. In this paper, we adopt Conditional Random Field (CRF), which can model these features and relationship of among objects to realize the simultaneous segmentation and classification.

2 Related Work

In computer vision and object recognition field, we could divide its processes into two types based on its outputs. One is called object detection, which gives the center position and a rectangle box of the target. For example, face recognition (Déniz et al. 2011), human detection (Zhu et al. 2006), part-based model (Felzenszwalb et al. 2009), sparselet model (Song et al. 2012), bag-of-features model (Lazebnik et al. 2006; Yang et al. 2009), sparse coding (Yang et al. 2009; Gao et al. 2010; Jia et al. 2012; Jiang et al. 2012), deep learning (Krizhevsky et al. 2012; Donahue et al. 2013), and so on. These methods can get the center positions and bounding boxes of the targets in the image. The other one is called semantic image segmentation, which is a process of simultaneous segmentation and recognition of an input image into regions and their associated categorical labels. An effective way to achieve this goal is to assign a label to each pixel of the input image and set some structural constraints on the output label space. Graph is a great tool to model the relations among different objects. With the MRF and MAP (Maximum Posterior Probability) developments, MRF and CRF which use probability graph model to represent this type of problems has been used widely. They model the label problem based on pixel features (Lafferty et al. 2001; Blake et al. 2004; Shotton et al. 2006; Larlus and Jurie 2008; Toyoda and Hasegawa 2008; Gould et al. 2008), region features (Yang et al. 2007; Kohli and Torr 2009; Fulkerson et al. 2009; Tighe and Lazebnik 2010; Yang and Forstner 2011b), multi-level regions (Russell et al. 2009; Schnitzspan et al. 2009; Yang et al. 2010; Kohli et al. 2013; Ladicky et al. 2014). In which, the image classification problem was expressed as a potential function and was transformed to an energy optimal problem, and they can get the boundary of the targets. Yang and Forstner (2011a, b), Zhong and Wang (2007), and Montoya-Zegarra et al. (2015) use CRF to model spatial and hierarchical structures to label and classify images of man-made scenes, such as buildings, roads, etc., and demonstrate the effective of CRF method in extracting man-made targets from high-resolution remote sensing images. We also could find that the common point of these two types of target recognition methods are both represent and make use of many kinds of features, such as SIFT (Scale Invariant Feature Transform), HOG (Histogram of Oriented Gradients), LBP (Local Binary Patterns), etc.

In remote sensing image applications, we usually need to get the accurate boundary of the targets. With this request, we take the second type method and the features used in the first type into consideration to get the object class label and boundary. In this paper, to make use of the spectral, texture and context features, we construct a three-level potential function which includes pixel level, segment level, and up-down-layer level with reference to Yoyoda and Hasegawa (2008), Kohli and Torr (2009), Ladicky et al. (2014), and use graph cut (Boykov et al. 2001; Boykov and Jolly 2001) to find the optimal solution. The experiments on GF-1 satellite image with 2 m spatial resolution showed that the proposed method is an efficient way to improve the segmentation and classification accuracy.

3 CRF-Based Image Classification Method

Features in different scales are very helpful information for image interpretation. High-resolution remote sensing images provided more detailed textures. To utilize the different scales and different kinds of features, we select four kinds of popular features and three scales' segmentation results to describe the classes. Figure 1 shows the workflow. We calculate the SIFT, LBP, Texton, Color SIFT features of each image. To reduce the computing complexity and to realize sparse representation, we apply the k-means clustering on four kinds of features that were got in the trained images, respectively, to get the visual words. Then construct the pixel and different scales segments potentials. Here, we select the mean shift image

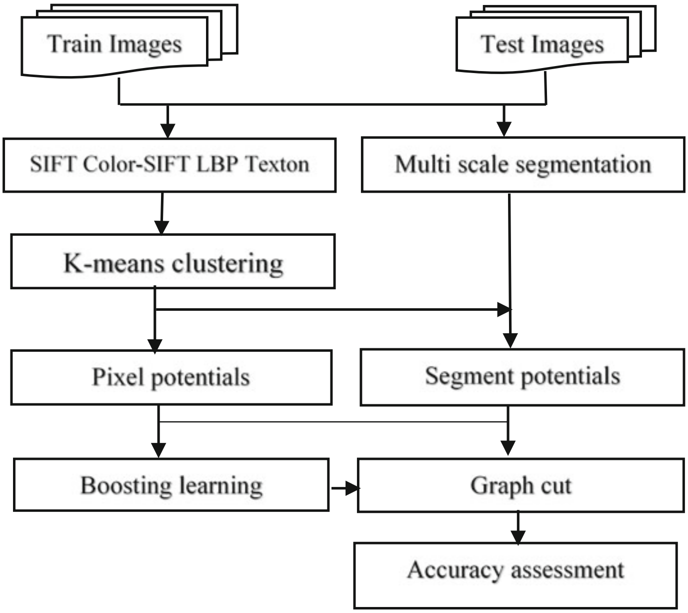


Fig. 1 Workflow of our method

segmentation method. Through boosting learning we got the classifiers. Finally, combine the pixel level, segment level, and the different level potentials together and use graph cut optimize algorithm to find the finest solution which means the fittest class of each pixel.

3.1 Features Calculation

3.1.1 SIFT Descriptor

In computer vision, SIFT descriptor are calculated after getting the key points and are used to realize point match generally. As the SIFT descriptor has the capability of describing the spatial distribution of a window, it has been used in many target detection research (Lazebnik et al. 2006; Yang et al. 2009; Jia et al. 2012), which just labeled the rectangle region where the object may exist.

This feature is derived from a 4 × 4 gradient window by using a histogram of 4 × 4 samples per window in 8 directions. The gradients are then Gaussian weighted around the center. This leads to a 128-dimensional feature vector. It reflects the distribution of gradients' direction. Before using this feature, we should normalize its value.

3.1.2 Color-SIFT Descriptor

Color is an important component for districting objects. Color invariant descriptors are proposed to increase illumination invariance and discriminative power. There are many different methods to obtain color descriptors, Van De Sande et al. (2010) compared the invariance properties and the distinctiveness of color descriptors. In this paper, we choose RGB-SIFT descriptor to describe the color invariant. For the RGB-SIFT descriptor, SIFT descriptors are computed for every RGB channel independently.

3.1.3 LBP Feature

An LBP is a local descriptor that captures the appearance of an image in a small neighborhood around a pixel. Due to its discriminative power and computational simplicity, LBP texture operator has become a popular approach in various applications. An LBP is a string of bits, with one bit for each of the pixels in the neighborhood. Each bit is got by thresholding the neighborhood of each pixel with the value of the center pixel. Here we select a 3×3 neighborhood, it is a string of 8 bits and has 256 possible LBPs.

3.1.4 Texton Feature

The term texton was proposed by Julesz (1981) first. It is described as "the putative units of pre-attentive human texture perception" (Julesz 1981). Leung and Malik (2001) use this term to describe vector quantized responses of a linear filter bank. Textons have been proven effective in categorizing materials as well as generic object classes. In this paper, we select three Gaussians, four Laplacian of Gaussians (LoG) and four first-order derivatives of Gaussians to build the filter bank. The three Gaussian kernels (with $\sigma = 0.1, 0.2, 0.4$) are applied to each CIE L, a, b channel, the four LoGs (with $\sigma = 0.1, 0.2, 0.4, 0.8$) were applied to the L channel only, and the four derivatives of Gaussians were divided into the two x- and y-aligned sets, each with two different values of σ ($\sigma = 0.2, 0.4$). Derivatives of Gaussians were also applied to the L channel only. Thus produced 17 final filter responses. Therefore, each pixel in each image has associated a 17-dimensional feature vector (Julesz 1981).

3.2 Mean Shift Image Segmentation

Mean shift is a robust feature space analysis approach and can delineate arbitrarily shaped clusters in it. Mean shift-based image segmentation has been widely used in many kinds of image, including high-resolution remote sensing images. The segmentation is actually a merging process performed on a region that is produced by the mean shift filtering. It considers both spatial domain and spectral domain while merging. For both domains, the Euclidean metric is used. Because the Euclidean distance in RGB color space does not correlate well to perceived difference in color by people, we use the LUV color space which better models the perceived difference in color in this space Euclidean distance. The use of the mean shift segmentation algorithm requires the selection of the bandwidth parameter h = (hr, hs), which determines the resolution of the mode detection by controlling the size of the kernel. To get a different scale segment result, we select 3 group of parameters which are $h1$ = (3.5, 3.5), $h2$ = (5.5, 3.5), $h3$ = (3.5, 5.5).

3.3 Potential Function

Image classification is a problem of assigning an object category label to each pixel in a given image. MRFs are the most popular models to incorporate local contextual constraints in labeling problems. Let Ii be the label of the ith site of the image set S, and Ni be the neighboring sites of site i. The label set $L\,(=\{l_i\}_{i\in S})$ is said to be a MRF on S w.r.t. a neighborhood N iff the following condition is satisfied

$$P\big(l_i|l_{S-\{i\}}\big) = P(l_i|l_{Ni}) \tag{1}$$

Let l be a realization of L, then $P(l)$ has an explicit formulation (Gibbs distribution):

$$P(l) = \frac{1}{Z}\exp(-\frac{1}{T}E(1)) \tag{2}$$

$$E(l) = \sum_{c\in C} V_C(l) = \sum_{\{i\}\in C_1} V_1(l_i) + \sum_{\{i,i'\}\in C_2} V_2(l_i, l_{i'}) + \cdots \tag{3}$$

where $E(l)$ is the energy function, Z is a normalizing factor, called the partition function, T is a constant, Clique $C_k = \{\{i, i', i'', \cdots\}|i, i', i'', \ldots$ are neighbors to one another$\}$. $V_{\mathrm{C}}(l)$ is the potential function, which represent a priori knowledge of interactions between labels of neighboring sites. Maximizing a posterior probability is equivalent to minimizing the posterior energy:

$$L^* = \arg\min_L E(L|X) \tag{4}$$

Let G = (S, E) be a graph, then (X, L) is said to be a CRF if, when conditioned on X, the random variables l_i obey the Markov property with respect to the graph:

$$P\left(l_i|X, l_{S-\{i\}}\right) = P(l_i|X, l_{N_i}) \tag{5}$$

where $S\{i\}$ is the set of all sites in the graph except the site i, N_i is the set of neighbors of the site i in G. We can find that CRF can directly infer posterior $P(L|X)$. In CRF, the potentials are functions of all the observation data as well as that of the labels. The CRF allows us to incorporate shape, color, texture, layout, and edge cues in a single unified model using a conditional potential. CRF model can be used to learn the conditional distribution over the class labeling given an image. Some kinds of the CRF have been proposed, for example, the image pixels (Gould et al. 2008; Toyoda and Hasegawa 2008), patches (Yang et al. 2007; Fulkerson et al. 2009; Kohli and Torr 2009; Tighe and Lazebnik 2010), or a hierarchy of regions (Russell et al. 2009; Yang et al. 2010; Kohli et al. 2013). We use a CRF model (Kohli and Torr 2009; Russell et al. 2009; Ladicky et al. 2014) to learn the conditional distribution over the class labeling given an image. We define the conditional probability of the class labels L given an image X

$$E(X) = \sum_{i \in V} \theta_v \varphi_i(x_i) + \sum_{(i,j) \in \varepsilon} \theta_\varepsilon \varphi_i\left(x_i, x_j\right) + \sum_{c \in S} \theta_s \varphi_c(X_c) \tag{6}$$

where V is a set of the image pixels, ε is the set of edges in an 8-connected grid structure; S is a set of image segments, $\varphi_i(x_i)$, $\varphi_{ij}\left(x_i, x_j\right)$ and $\varphi_c(X_c)$ are the potentials defined on them, θ_v, θ_ε and θ_s are the model parameters, and i and j index pixels in the image, which correspond to nodes in the graph. In this paper, we defined three potentials which are unary potential, pairwise potential, and region potential. We will describe these potentials as follows.

3.3.1 Unary Potential

The unary potential allows for local and global evidence aggregation, each potential models the evidence from considering a specific image feature. Usually, it is computed from the color of the pixel and the appearance model for each object. However, color alone is not a very discriminative feature and fails to produce accurate segmentations and classification. This problem can be overcome by using sophisticated potential functions based on color, texture, location, and shape priors. The unary potential used by us can be written as

$$\varphi_i(x_i) = \theta_s \varphi_s(x_i) + \theta_l \varphi_l(x_i) + \theta_t \varphi_t(x_i) + \theta_{cs} \varphi_{cs}(x_i) \tag{7}$$

where θ_s, θ_l, θ_t and θ_{cs} are parameters weighting the potentials obtained from SIFT, LBP, texton, and color SIFT respectively.

3.3.2 Pairwise Potential

The pairwise potentials have the form of a contrast sensitive Potts model (Kohli and Torr 2009).

$$\varphi_{ij}(x_i, x_j) = \begin{cases} 0 & \text{if } x_i = x_j \\ g(i,j) & \text{otherwise} \end{cases} \tag{8}$$

where the function $g(i,j)$ is an edge feature based on the difference in colors of neighboring pixels (Song et al. 2012). It is typically defined as

$$g(i,j) = \theta_p + \theta_v \exp(-\theta_\beta ||I_i - I_j||^2) \tag{9}$$

where *Ii* and *Ij* are the color vectors of pixel *i* and *j* respectively. θ_p, θ_v and θ_β are model parameters whose values are learned using training data.

3.3.3 Region Consistency Potential

The region consistency potential is modeled by the robust Pn Potts model (Kohli and Torr 2009). It supports all pixels belonging to a segment taking the same label and allows some variables in the segment to take different labels and reflect the consistency of segments. It is very useful in obtaining object segmentations with fine boundaries. We refer the reader to Kohli and Torr (2009) for more details. It takes the form of

$$\varphi_c(X_c) = \begin{cases} N_i(X_c)\frac{1}{Q}\gamma_{\max} & \text{if } N_i(X_c) \le Q \\ |c|^{\theta_\alpha}(\theta_p^h + \theta_v^h G(c)) & \text{otherwise} \end{cases} \tag{10}$$

where $N_i(X_c) = \min_k (|c| - n_k(X_c))$, which denotes the number of variables in the clique c not taking the dominant label. $\gamma_{\max} = |c|^{\theta_\alpha}(\theta_p^h + \theta_v^h G(c))$, and Q is the truncation parameter which controls the rigidity of the higher order clique potential. $G(c)$ is used to evaluate the consistency of all constituent pixels of a segment, the variance of the response of a unitary was used, that is

$$G(c) = \exp\left(-\theta_\beta^h \frac{||\sum_{i\in c}(f(i) - \mu)^2||}{|c|}\right) \tag{11}$$

where $\mu = \frac{\sum_{i\in c} f(i)}{|c|}$ and $f()$ is a function being used to evaluate the quality of a segment. This enhanced potential function gives rise to a cost that is a linear truncated function of the number of inconsistent variables (Kohli and Torr 2009).

In this paper, we use boosting algorithm to train the three part of the energy function. The boosting algorithm helps us select features and get a strong classifier.

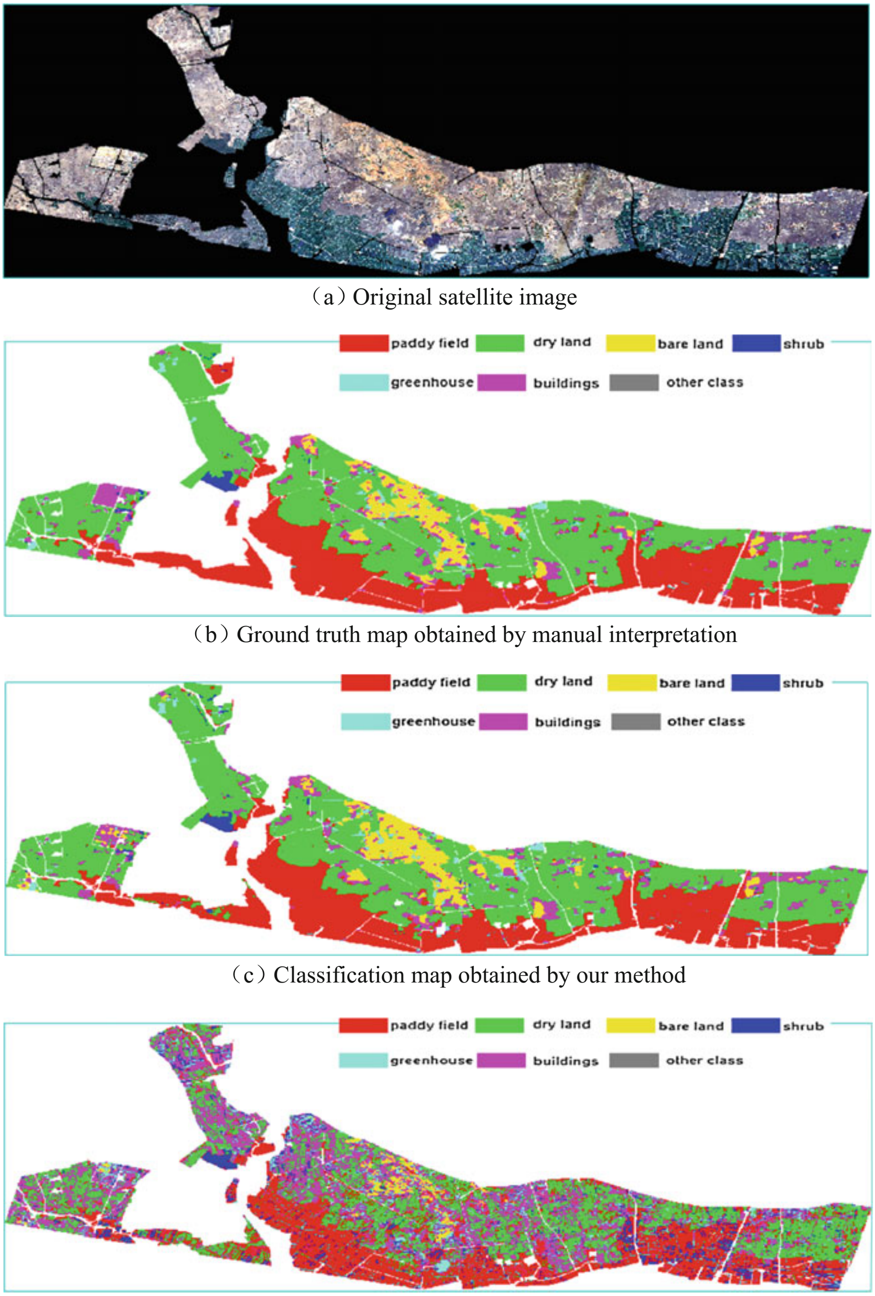
3.4 Graph Cut

Given the CRF model and its learned parameters, we wish to find the most probable labeling l, i.e., the labeling that minimize the energy function of (6). The graph cut-based α-expansion and $\alpha\beta$-swap is an effective way to solve energy minimization problem (Boykov et al. 2001). It transforms the energy minimizing problem to min-cut of graph problem. It has been successfully used to minimize energy functions composed of pairwise potential functions.

As a kind of move making algorithms, the first step of graph cut is initializing nodes, edges and build up the graph. In this paper, in accordance with the constituents of the energy function, the nodes and edges consist of the pixel, pairwise, and three scales segments-level components. The three levels' potentials are calculated and used to initialize the edge weight. The initial label image is set according to the minimum cost of each class. Then, it computes optimal alpha-expansion moves for labels in some order, accepting the moves only if they increase the objective function. The algorithm's output is a strong local maximum, which means the solution of minimum energy was found.

4 Results and Discussion

In our experiments, we select high-resolution satellite images to verify our method's availability. The satellite image was the merged image of the multispectral image and panchromatic image of 2 m spatial resolution of "GF-1." The test area is located in the southeast of Liaoning province in China. The image was manually interpreted, which include paddy field, dry land, bare land, shrub land, buildings, and greenhouse. Some area is labeled to "other class." The image has 20000 × 8000 pixels. We split this large image into lots of 128 * 128 small images which are then divided into the train and test group and take the 30% of the small images as train images. To compare the result with the other object-oriented image classification method, we select the wide used software eCognition to compare. In eCognition classification process, through many trials, we selected the scale parameter 100, calculated the mean value and standard variation of each band, the secondary angle moment of GLCM feature, and use the Nearest Neighbor algorithm to realize the object-oriented image classification. Figure 2 shows the origin image (a), ground truth (b), classification result of this paper's method (c), and classification result of eCognition. Table 1 gives the image classification accuracies of these two methods. From them, we could find the accuracy of our method is higher than the traditional object-oriented image classification. At the same time, compared with the traditional method, the phenomena of salt-and-pepper was greatly improved.



（a）Original satellite image

（b）Ground truth map obtained by manual interpretation

（c）Classification map obtained by our method

（d）Classification map obtained by eCognition

Fig. 2 Original satellite image, ground truth, and classification results

Table 1 Comparison of classification accuracy

	Paddy field	Dry land	Bare land	Shrub	Building	Green house	Others	Overall accuracy	Average accuracy	Kappa
eCognitioan	0.7943	0.8256	0.7534	0.7456	0.8304	0.6021	0.8912	0.8076	0.7775	0.6712
Our method	0.9303	0.9420	0.8014	0.8392	0.9045	0.7328	0.9366	0.9108	0.8695	0.8564

5 Conclusion

From the experiment procedures and results, we could find that this method has three main valuable aspects to be used in image classification. First, it does not need manual segmentation which could avoid the boring segmentation scale selection problem. Second, it can make good use of the existing classification results to train the classifier. Third, it can realize self-selecting features and improve the segmentation and classification results. It may be a way of large of samples-based image interpretation. At the same time, we must admit that this method is still a time-consuming process. How to use it to the task of large-scale geographical conditions general survey and monitoring is still a problem to be solved.

Acknowledgements The study was partially supported by the High-resolution Comprehensive Traffic Remote Sensing Application program under Grant No. 07-Y30B10-9001-14/16, the National Natural Science Foundation of China under Grant No. 41101410 and Foundation of Key Laboratory for National Geographic State Monitoring of National Administration of Survey, Mapping and Geoinformation under Grant No. 2014NGCM.

References

Beucher S, Meyer F (1992) The morphological approach to segmentation: the watershed transformation. Mathematical morphology in image processing. Marcel Dekker, New York, pp 433–481

Blake A, Rother C, Brown M, Perez P, Torr P (2004) Interactive image segmentation using an adaptive GMMRF model. Computer vision-ECCV 2004. Springer, Berlin, pp 428–441

Boykov YY, Jolly MP (2001) Interactive graph cuts for optimal boundary and region segmentation of objects in ND images. In: Proceedings of eighth IEEE international conference on computer vision, 2001, ICCV 2001, vol 1. IEEE, New York, pp 105–112

Boykov Y, Veksler O, Zabih R (2001) Fast approximate energy minimization via graph cuts. IEEE Trans Pattern Anal Mach Intell 23(11):1222–1239

Comaniciu D, Meer P (2002) Mean shift: a robust approach toward feature space analysis. IEEE Trans Pattern Anal Mach Intell 24(5):603–619

Déniz O, Bueno G, Salido J, De la Torre F (2011) Face recognition using histograms of oriented gradients. Pattern Recogn Lett 32(12):1598–1603

Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, Darrell T (2013) Decaf: a deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531

Felzenszwalb PF, Huttenlocher DP (2004) Efficient graph-based image segmentation. Int J Comput Vision 59(2):167–181

Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2009) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645

Fulkerson B, Vedaldi A, Soatto S (2009) Class segmentation and object localization with superpixel neighborhoods. In: International conference on computer vision, vol 9, pp 670–677

Gao S, Tsang IWH, Chia LT (2010) Kernel sparse representation for image classification and face recognition. Computer Vision–ECCV 2010. Springer, Berlin, pp 1–14

Gould S, Rodgers J, Cohen D, Elidan G, Koller D (2008) Multi-class segmentation with relative location prior. Int J Comput Vision 80(3):300–316

Jia Y, Huang C, Darrell T (2012) Beyond spatial pyramids: Receptive field learning for pooled image features. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, New York, pp 3370–3377

Jiang Z, Zhang G, Davis LS (2012) Submodular dictionary learning for sparse coding. In: 2012 IEEE conference on computer vision and pattern recognition (CVPR), IEEE, New York, pp 3418–3425

Julesz B (1981) Textons, the elements of texture perception, and their interactions. Nature 290 (5802):91–97

Kohli P, Torr PH (2009) Robust higher order potentials for enforcing label consistency. Int J Comput Vision 82(3):302–324

Kohli P, Osokin A, Jegelka S (2013) A principled deep random field model for image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1971–1978

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagine classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105

Ladicky L, Russell C, Kohli P, Torr PH (2014) Associative hierarchical random fields. IEEE Trans Pattern Anal Mach Intell 36(6):1056–1077

Lafferty J, McCallum A, Pereira FC (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of the eighteenth international conference on machine learning (ICML '01), San Francisco, CA, USA, pp 282–289

Larlus D, Jurie F (2008) Combining appearance models and markov random fields for category level object segmentation. In: IEEE conference on computer vision and pattern recognition, 2008, CVPR 2008. IEEE, New York, pp 1–7

Lazebnik S, Schmid C, Ponce J (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE computer society conference on computer vision and pattern recognition, vol 2. IEEE, New York, pp 2169–2178

Leung T, Malik J (2001) Representing and recognizing the visual appearance of materials using three-dimensional textons. Int J Comput Vision 43(1):29–44

Montoya-Zegarra JA, Wegner JD, Ladický L, Schindler K (2015) Semantic segmentation of aerial images in urban areas with class-specific higher-order cliques. ISPRS Ann Photogramm Remote Sens Spat Inf Sci 2(3):127–133

Russell C, Kohli P, Torr PH (2009) Associative hierarchical crafts for object class image segmentation. In: 2009 IEEE 12th international conference on computer vision. IEEE, New York, pp 739–746

Schnitzspan P, Fritz M, Roth S, Schiele B (2009) Discriminative structure learning of hierarchical representations for object detection. In: IEEE conference on computer vision and pattern recognition, pp 2238–2245

Shotton J, Winn J, Rother C, Criminisi A (2006) Textonboost: joint appearance, shape and context modeling for multi-class object recognition and segmentation. Computer vision–ECCV 2006. Springer, Berlin, pp 1–15

Song H, Zickler S, Althoff T, Girshick R, Fritz M, Geyer C, Felzenszwalb P, Darrell T (2012) Sparselet models for efficient multiclass object detection. In: European conference on computer vision, pp 802–815

Tighe J, Lazebnik S (2010) Superparsing: scalable nonparametric image parsing with superpixels. Computer vision–ECCV 2010. Springer, Berlin, pp 352–365

Toyoda T, Hasegawa O (2008) Random field model for integration of local information and global information. IEEE Trans Pattern Anal Mach Intell 30(8):1483–1489

Van De Sande KE, Gevers T, Snoek CG (2010) Evaluating color descriptors for object and scene recognition. IEEE Trans Pattern Anal Mach Intell 32(9):1582–1596

Yang M, Forstner W (2011b) Regionwise classification of building facade images. In: Photogrammetric image analysis, LNCS 6952, Springer, Berlin, pp 209–220

Yang MY, Förstner W (2011a) A hierarchical conditional random field model for labeling and classifying images of man-made scenes. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE, New York, pp 196–203

Yang L, Meer P, Foran DJ (2007) Multiple class segmentation using a unified framework over mean-shift patches. In: IEEE conference on computer vision and pattern recognition, 2007, CVPR'07. IEEE, New York, pp 1–8

Yang J, Yu K, Gong Y, Huang T (2009) Linear spatial pyramid matching using sparse coding for image classification. In: IEEE conference on computer vision and pattern recognition (CVPR), 2009. IEEE, New York, pp 1794–1801

Yang M, Forstner W, Drauschke M (2010) Hierarchical conditional random field for multi-class image classification. In: International conference on computer vision theory and applications, pp 464–469

Zhong P, Wang R (2007) A multiple conditional random fields ensemble model for urban area detection in remote sensing optical images. IEEE Trans Geosci Remote Sens 45(12): 3978–3988

Zhu Q, Yeh MC, Cheng KT, Avidan S (2006) Fast human detection using a cascade of histograms of oriented gradients. In: 2006 IEEE computer society conference on computer vision and pattern recognition, vol 2. IEEE, New York, pp 1491–1498