Joint Semantic Segmentation and Object Detection with Improved Detector Potentials

REN Jin-sheng and JIA Hai-tao

(School of Computer Science and Engineering, University of Electronic Science and Technology of China Chengdu 611731)

Abstract Computer vision algorithms for individual tasks such as object recognition, detection and segmentation have shown impressive results in the recent years. The next challenge is to integrate all these algorithms and address the problem of scene understanding. A new higher order conditional random field (CRF) model is proposed to get semantic segmentation and object detection simultaneously. Specifically, the proposed higher order CRF model consists of low-order potentials and improved detector potentials. To avoid wrong recognition caused by the confidence given by the initial detector, the first-and-second-order pooling and logistic regression are adopted to improve the detector potential. Experimental results show that the proposed model achieves significant improvement over the baseline methods on MSRC 21-class and PASCAL VOC 2007 datasets.

Key words detector potential; first-and-second-order pooling; higher-order CRF model; segmentation

基于改进目标检测能量项的联合语义分割和目标检测

任金胜，贾海涛

(电子科技大学计算机科学与工程学院 成都 611731)

【摘要】提出了一种新颖的高阶CRF模型，能够同时获得语义分割和目标检测结果。该高阶CRF模型由低阶能量项和改进目标检测能量项构成。该模型采用了一二阶合并方法和逻辑斯蒂回归，从而降低了由于初始检测不准确而导致的错误识别率。在MSRC 21和PASCAL VOC 2007两组数据库上进行的实验表明，该方法显著优于传统方法。

关 键 词 目标检测能量项; 一二阶合并; 高阶CRF模型; 语义分割

中图分类号 TP391 文献标志码 A doi:10.3969/j.issn.1001-0548.2014.05.020

Scene understanding has been one of the central goals in computer vision for many decades[1]. Currently, many scholars formulate scene understanding as semantic segmentation, which aims to label each pixel in an image with a class label from a predetermined set, e.g. building, tree, face, body[2-4]. Apparently, semantic segmentation completes image segmentation and object recognition at the same instant, which are the subtasks of scene understanding. However, it fails to distinguish between adjacent instances of objects of the same class [4-5]. On the other hand, object detection approaches[6] provide the number of instances of objects, but do not provide information about background classes, such as grass, sky, and road.

For example, road scene datasets contain classes with specific shapes such as person, car, and bicycle, as well as background classes such as road, sky, and building. Imagine a typical road scene picture, it can be segmented into the object class result for standard CRF approach. The segmentation objects from road scene create the semantic results such as tree or car and so on. Complete scene understanding requires not only the pixel-wise segmentation of an image, but also an identification of object instances of a particular class. Semantic segmentation methods such as discriminative model[2] and hierarchical random field model[7] would label all the cars adjacent to each other as belonging to a large car segment or blob, as illustrated in Fig. 1. Thus, we would not have information about the number of instances of a particular object-car in this case. On the other hand, object detection methods can identify

Received date：2014 – 03 – 03; Record date: 2014 –07 – 08

收稿日期：2014 – 03 –03；修回日期：2014 –07 – 08

Biography：REN Jin-sheng was born in 1981, and his research interests include is neural network and computer vision.

作者简介：任金胜(1981 –)，男，博士生，主要从事神经网络和计算机视觉方面的研究.

the number of objects[6], but cannot be used for background classes.

Recently a few approaches have attempted to combine semantic segmentation with object detection, however they suffer from certain drawbacks [8-9]. The difficulty with these approaches is that the subtask representations can be inconsistent. To overcome the difficulty, Ref.[5] proposed a new CRF model for reasoning about regions, objects, and their attributes such as object class, location, and spatial extent. However, Ladicky directly used the confidence obtained by the initial detector, e.g. the parts based detector proposed in Ref.[6] that would lead to wrong recognition.

Mainly inspired by Ref.[5] work, this paper proposes a new higher order CRF model to get semantic segmentation and object detection simultaneously. Specifically, the proposed higher order CRF model consists of low order potentials and improved detector potential. To avoid the wrong recognition led by the confidence given by the initial detector, first-and-second-order pooling and logistic regression are adopted to improve the detector potential.

1 Description of Higher Order CRF Model

Fig.1 depicts an overview of the joint object detection and semantic segmentation model. Firstly, an object detector is used to complete the initial object detection and obtain the bounding box. Secondly, global shape features are extracted within the bounding box, and local texture features are extracted from the foreground. Then global feature descriptors are obtained by combining these local texture features through the first-and-second-order pooling. The logistic regression classifier with the global feature descriptors improves the detector potential. At last, the improved detector potentials and low-order potentials constitute the higher-order CRF model. The combined segmentation and detection are computed from the CRF model, which is given by formula (1):

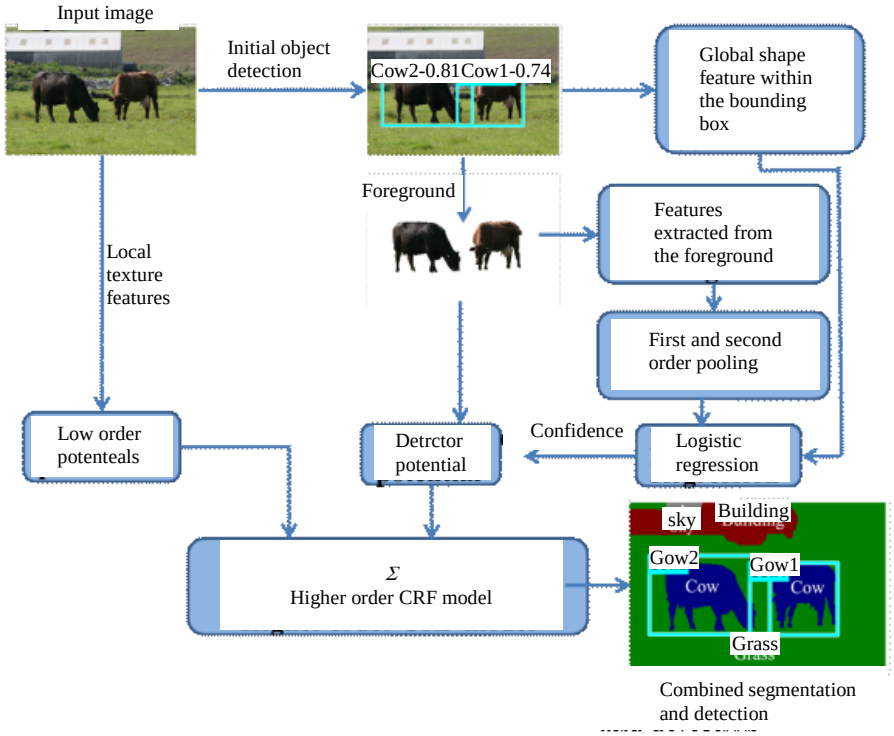


Fig.1 An overview of the proposed algorithm for joint semantic segmentation and object detection

$$E(x)=\underbrace{\sum_{i\in \boldsymbol{V}} E_i(x_i)+\alpha \sum_{(i,j)\in \boldsymbol{\mathcal{E}}} E_{ij}(x_i,x_j)}_{\text{lower-order potentials}}+\underbrace{\gamma \sum_{k\in \boldsymbol{D}} E_{d_k}(x_{d_k})}_{\text{detector potential}} \tag{1}$$

In formula (1), lower-order potentials include unary potential E_i and pairwise potential E_{ij}; E_{d_k} is the detector potential; α and γ are the weights. The two kinds of potentials will be detailed in Section 1.1 and

1.2 separately.

1.1 Lower-order Potentials

Lower-order potentials are usually used to make up the pairwise CRF model as given by formula (2), which is one of the most effective and prevalent models for the image semantic segmentation problem[2].

$$E(x)=\sum_{i\in V}E_i(x_i)+\alpha\sum_{(i,j)\in\mathcal{E}}E_{ij}(x_i,x_j) \qquad (2)$$

In the standard pairwise CRF formulation for the image semantic segmentation problem[2], each pixel is usually represented as a random variable (label). Each of these random variables takes a label from the provided set $\boldsymbol{L}=\{l_1,l_2,\cdots,l_k\}$, which may represent objects such as car, airplane, and bicycle. Let $\boldsymbol{X}=\{X_1,X_2,\cdots,X_N\}$ denote the set of random variables corresponding to the image pixels, which are $N=H\times W$ (H is the Height and W is the width of the image). A clique c is a set of random variables X_c which are conditionally dependent on each other. In the standard pairwise CRF model, the size of the clique c should be 1 or 2. A labelling x refers to any possible assignment of labels to the random variables and takes values from the set $\boldsymbol{L}=L^N$. Fig.2 shows an example of the pairwise CRF model.

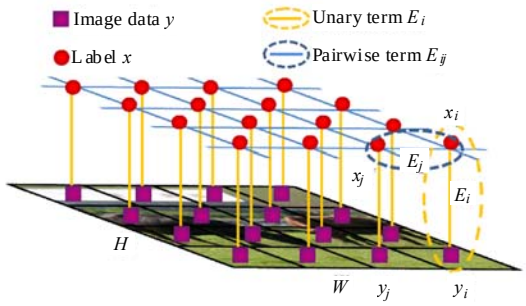


Fig.2 Pairwise CRF model

In the given image of size $H\times W$, y_i represents the image data pixel i (or the features extracted from the small area around pixel i), and x_i represents the random label of pixel i. Unary term E_i and pairwise term E_{ij} are separately defined on the cliques $c_i=\{x_i\}$ and $c_{ij}=\{x_i,x_j\}$. Here are $N=H\times W$ pixels, thus labelling x could be any one of $L^{H\times W}$ combinations.

1) Unary potential

$$E_i(x_i)=-\log P(x_i\mid \boldsymbol{y}) \qquad (3)$$

The unary potential is defined as the negative log of the likelihood of a label being assigned to pixel i, as shown by expression (3) above.

2) Pairwise potential

The pairwise terms E_{ij} of the pairwise CRF take the form of a contrast sensitive Potts model:

$$E_{ij}(x_i,x_j)=\begin{cases}0 & \text{if } x_i=x_j\\ g(i,j) & \text{otherwise}\end{cases} \qquad (4)$$

where the function $g(i, j)$ is an edge feature based on the difference in colors of neighboring pixels[10-11]. It is typically defined as:

$$g(i,j)=\theta_p+\theta_v\exp(-\theta_\beta\left\|I_i-I_j\right\|^2) \qquad (5)$$

where I_i and I_j are the colour vectors of pixel i and j respectively. θ_p, θ_v and θ_β are model parameters whose values are learned by using training data. More details can be found in Ref.[12].

1.2 Detector Potential

The detector potential is defined on the foreground segmentation. The object detection could be obtained by many good detectors, e.g. histogram-based detector proposed in Ref.[13] and parts based detector proposed in Ref.[6]. The method proposed in Ref.[12] is used to extract the foreground in the bounding box, on which the detector potential is defined:

$$E_{d_k}(x_{d_k})=-|x_{d_k}|\max(0,(1-R)\max(0,(C_{d_k}-C_t))) \qquad (6)$$

$$R=N_{d_k}/R_t|x_{d_k}| \qquad (7)$$

where $d_k\in\boldsymbol{L}$ represents the kth detected object region and the corresponding label, $\boldsymbol{x}_{d_k}$ is the set of pixels in the detected object region d_k, E_{d_k} is called a higher-order potential, C_{d_k} is the classifier response and indicates an increased likelihood of the presence of an object at a location, i.e. in the bounding box, C_t is the t hreshold which is generally set 0, and N_{d_k} means the number of pixels whose labels are different from d_k. $R_t\in(0,1]$, and it gets better results when its values are between 0.1 and 0.3. Here R makes sure that in $\boldsymbol{x}_{d_k}$ there could be some pixles of labels different from d_k.

1.3 Inference for Detector Potentials

The x^* which makes the energy $E(x)$ minimum is the semantic segmentation result, as expressed in

formula (8):

$$x^* = \arg\min_x E(x) \tag{8}$$

There are several methods to solve the equation (8), and this paper uses the algorithm proposed by Kohli[14-15]. When Kohli's optimization method is used, some conditions have to be satisfied as shown below:

$$E_d(x_d) = \min(\min_{n\in \boldsymbol{L}}(\gamma_n + R(\gamma_{\max} - \gamma_n)), \gamma_{\max}) \tag{9}$$

Here, we show that our detector potential in equation (6) can be converted into a form solvable by using $\alpha\beta$-swap and α-expansion algorithms, i.e. the equation (6) is equivalent to equation (9).

Define:

$$F = | x_{d_k} | \max(0, (C_{d_k} - C_t)) \tag{10}$$

2 Improved Detector Potential

As described in section 1 the CRF model incorporated with the detector potential can get semantic segmentation and object detection. However, this higher-order term directly uses the response of the object detector which considers only the shape features but appearance features e.g. color or local textures. As Ref.[16]pointed out, better object detectors usually combines local appearance features and global shape features. In consideration of this point, we improved the detector potential shown in section II. As shown in Fig. 1, global shape features are extracted within the bounding box, and local texture features are extracted from the foreground. Then global feature descriptors are obtained by combining these local texture features through the first-and-second-order pooling. The logistic regression classifier with the global feature descriptors gives the final confidence, i.e. C_{d_k} in equation (6).

2.1 Global Feature Descriptors

In order to obtain more robust feature descriptors from the object detect region, we firstly extract the local features including shape and texture features, and then combine them by first-and-second-order pooling.

Histogram of oriented gradient (HOG)[17] is used as the global shape feature. The global shape feature extracted within the kth detected object bounding box is denoted by s_k. Three different local descriptors are employed to capture the local features, including color, Gray-Level SIFT [18], and Opponent SIFT [10]. Color feature at pixel i is denoted as c_i, constituted by RGB, HSV, and LAB color values. Gray-Level SIFT at pixel i is denoted as gs_i, and similarly Opponent SIFT is denoted as os_i. now we have the local feature set $\boldsymbol{Y} = (s_k, \{c_i\}_k, \{\mathrm{gs}_i\}_k, \{\mathrm{os}_i\}_k) = y_1, y_2, \cdots, y_m$. $|\boldsymbol{Y}| = m$ means the dimensions of the feature.

The most common first-order pooling operator is given by formula.

$$G^1_{\mathrm{avg}}(\boldsymbol{Y}) = \sum_{\boldsymbol{y}_i \in \boldsymbol{Y}} \boldsymbol{y}_i / |\boldsymbol{Y}| \tag{11}$$

and the second-order pooling operator to conquer this problem:

$$G^2_{\mathrm{avg}}(\boldsymbol{Y}) = \sum_{\boldsymbol{y}_i \in \boldsymbol{Y}} \boldsymbol{y}_i \boldsymbol{y}_i^{\mathrm{T}} / |\boldsymbol{Y}| \tag{12}$$

In this paper, we further improve Carreira's method and propose the novel first-and-second-order pooling operator expressed by formula (13) and (14).

$$\boldsymbol{\mu}_R = \sum_{\boldsymbol{y}_i \in \boldsymbol{Y}} \boldsymbol{y}_i / |\boldsymbol{Y}| \tag{13}$$

$$\Sigma_R = \sum_{\boldsymbol{y}_i \in \boldsymbol{Y}} (\boldsymbol{y}_i - \boldsymbol{\mu}_R)(\boldsymbol{y}_i - \boldsymbol{\mu}_R)^{\mathrm{T}} / |\boldsymbol{Y}| \tag{14}$$

$\boldsymbol{\mu}_R$ and the upper tridiagonal matrix of Σ_R are combined into the final global feature descriptor. This new global feature descriptor takes into account the average information and correlation between different features, so it is more robust than formulas (11) and (12).

2.2 Logistic Regression Classifier

The logistic classifier is one of the most widely used classifiers. This paper uses it as the classifier since it could be naturally generalized to the multi-class classifier. The standard logistic function is defined as below:

$$f(z) = \mathrm{e}^z / (\mathrm{e}^z + 1) = 1/(1 + \mathrm{e}^{-z}) \tag{15}$$

It is easy to transform into the probabilistic expression:

$$P(\boldsymbol{x} \mid \boldsymbol{y}; \boldsymbol{w}) = 1/1 + \mathrm{e}^{-z(y)}, \text{ where } z(\boldsymbol{y}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{y} \tag{16}$$

Formula (16) is usually applied to the two-class recognition, where y is the input feature vector, x is the target classes (0 or 1), and w is the model parameter. However, this paper deals with multi-class recognition, and the two-class logistic classifier need to be generalized to multi-class classifier. The general multi-class logistic classifier is given by formula (17):

$$P(x=k\mid \boldsymbol{y};\boldsymbol{w})=\exp(\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{y})/\sum_{l=1}^{|\boldsymbol{L}|}\exp(\boldsymbol{w}_l^{\mathrm{T}}\boldsymbol{y}) \tag{17}$$

where $\boldsymbol{L}$ represents the class label set, and $|\boldsymbol{L}|$ represents the number of the classes. Formula (17) is often called a softmax function. According to the softmax function, $|\boldsymbol{L}|$ parameters need to be learned. However, once $|\boldsymbol{L}|-1$ parameters are defined, and then the last parameter is also defined. Thus, formula (17) is over complete. This paper applies another form of multi-class logistic classifier:

$$P(x=k\mid \boldsymbol{y})=\begin{cases}\dfrac{\exp(\boldsymbol{w}_k^{\mathrm{T}}\boldsymbol{y})}{1+\sum_{l=1}^{|\boldsymbol{L}|-1}\exp(\boldsymbol{w}_l^{\mathrm{T}}\boldsymbol{y})} & \text{if } k<|\boldsymbol{L}|\\ \dfrac{1}{1+\sum_{l=1}^{|\boldsymbol{L}|-1}\exp(\boldsymbol{w}_l^{\mathrm{T}}\boldsymbol{y})} & \text{if } k=|\boldsymbol{L}|\end{cases} \tag{18}$$

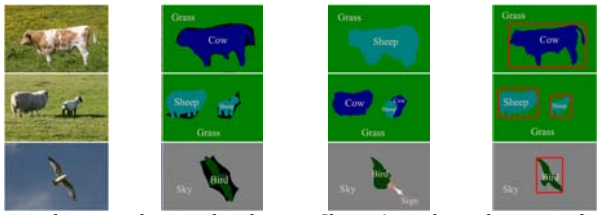
The final class is decided by the formula below:

$$\boldsymbol{x}^{*}=\arg\max_{\boldsymbol{x}} P(\boldsymbol{x}\mid \boldsymbol{y}) \tag{19}$$

3 Simulation Results

We evaluate our model on MSRC 21-class[2] and PASCAL VOC 2007 datasets.

3.1 Qualitative Results



a. original images b. ground truth c. Shotton's results d. mine results

Fig.3 Semantic segmentation on MSRC 21-class dataset

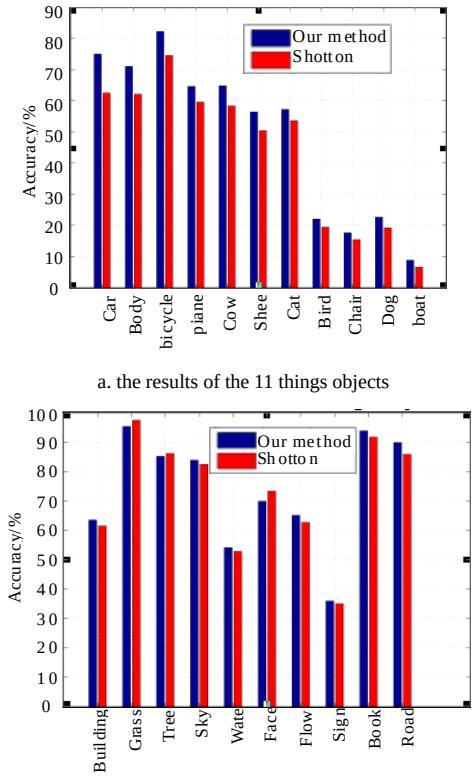
The qualitative results of MSRC 21-class dataset are shown in Fig.3. Fig.3a represents the original images, Fig.3b represents the ground truth images, 3c is the results of Shotton's algorithm[2], and Fig.3d is the results of our method. In these images, different colors mean different object classes. For clarity, textual labels have been superimposed on the resulting segmentations. For example, green means Grass object and blue means Cow object. It is easy to find that the segmentation results are not accurate and make wrong recognition[2]. This is because the pairwise CRF model considers sole information from the pixel level and thus performs not so well[2]. Different from the prior model, our model takes the object detection into account and incorporates the higher-order potential, and thus obtains better semantic segmentation results and computes the number of objects at the same time.

3.2 Quantitative Results

We also give the quantitative results on MSRC 21-class and PASCAL VOC 2007 datasets. Accuracy is used to evaluate the performance, defined as formula (20).

$$\text{Accuracy}=\frac{\text{True Positive}}{\text{True Positive + False Negative}} \tag{20}$$

Fig.4 shows the quantitative results of MSRC 21-class dataset. Our model performs better than Shotton's for these 11 thing objects for which the DPM detector responses (see Fig.4a). However, our model is normal and even worse than Shotton's for the rest 10 stuff objects (see Fig.4b). This phenomenon proves that the detector potential does work well.



a. the results of the 11 things objects

b. the results of the rest 10 stuff objects

Fig.4 The quantitative results of MSRC 21-class dataset

Table 1 gives the quantitative results of PASCAL VOC 2007 dataset. our model provides a small increase in accuracy: 2% than the pairwise model[2] and 1% than the associative model[7].

Table 1 The quantitative results of PASCAL VOC 2007 dataset

	Pairwise CRF	Associative CRF	Our CRF model
Background	83	78	75
Aeroplane	12	14	18
Bicycle	28	27	25
Bird	24	26	25
Boat	2	0	3
Bottle	2	0	1
Bus	25	29	33
Car	8	10	12
Chair	1	0	2
Potted plant	2	2	3
Sheep	15	17	18
Sofa	2	0	3
Train	30	33	38
TV/Monitor	22	22	25
Average	18.2	19.2	20.1

4 Conclusions

This paper proposes a higher-order CRF model with improved detector potentials. Especially, we introduce a novel first-and-second-order pooling operator to get robust feature descriptor of the foreground. The benefits of this approach can be seen in the results, and our approach consistently demonstrates an improvement over the baseline methods. This work increases the expressibility of CRFs, and shows how they can be used to identify object instances and obtain the number of objects.

参 考 文 献

[1] BARROW H G, TENENBAUM J M. Computational vision[J]. Proceedings of the IEEE, 1981, 69(5): 572-595.

[2] SHOTTON J, WINN J, ROTHER C, et al. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context[J]. International Journal of Computer Vision, 2009, 81(1): 2-23.

[3] LADICKÝ L. Global structured models towards scene understanding[D]. Oxford, England: Oxford Brookes University, 2011.

[4] GOULD S. Multiclass pixel labeling with non-local matching constraints[C]//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, United states: IEEE, 2012: 2783-2790.

[5] LADICKY L, STURGESS P, ALAHARI K, et al. What, where and how many? combining object detectors and CRFs[C]//Proceedings of the 11th European Conference on Computer Vision. Heraklion, Crete, Greece: [s.n.], 2010: 424-437.

[6] FELZENSZWALB P F, GIRSHICK R B, MCALLESTER D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[7] LADICKY L U, RUSSELL C, KOHLI P, et al. Associative hierarchical CRFs for object class image segmentation[C]// Proceedings of the 12th IEEE International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009: 739-746.

[8] LARLUS D, JURIE F. Combining appearance models and Markov random fields for category level object segmentation[C]//Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition. Anchorage, AK, United states: IEEE, 2008.

[9] GU C, LIM J J, ARBELAEZ P, et al. Recognition using regions[C]//Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Miami, FL, United states: IEEE, 2009: 1030-1037.

[10] VAN DE SANDE K, GEVERS T, SNOEK C. Evaluating color descriptors for object and scene recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1582-1596.

[11] BOYKOV Y Y, JOLLY M P. Interactive graph cuts for optimal boundary region segmentation of objects in N-D images[C]//Proceedings of the 8th International Conference on Computer Vision. [S.l.]: [s,n,], 2001: 105-112.

[12] ROTHER C, KOLMOGOROV V, BLAKE A. 'GrabCut' - interactive foreground extraction using iterated graph cuts[C]//Proceedings of the ACM SIGGRAPH 2004. Los Angeles, United States: ACM, 2004: 309-314.

[13] VEDALDI A, GULSHAN V, VARMA M, et al. Multiple kernels for object detection[C]//Proceedings of the 12th International Conference on Computer Vision. Kyoto, Japan: [s.n.], 2009: 606-613.

[14] BOYKOV Y, KOLMOGOROV V. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2004, 26(9): 1124-1137.

[15] KOHLI P, LADICKÝ L, TORR P H S. Robust higher order potentials for enforcing label consistency[J]. International Journal of Computer Vision, 2009, 82(3): 302-324.

[16] GOULD S, GAO T, KOLLER D. Region-based segmentation and object detection[C]//Proceedings of the 23rd Annual Conference on Neural Information Processing Systems. Vancouver, BC, Canada: [s.n.], 2009: 655-663.

[17] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection[C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, United states: IEEE, 2005: 886-893.

[18] LOWE D. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

编 辑 税 红