Automatic foreground extraction via joint CRF and online learning

Wenbin Zou, Kidiyo Kpalma, Joseph Ronsin

Automatic foreground extraction via joint CRF and online learning

W. Zou, K. Kpalma and J. Ronsin

A novel approach is proposed for automatic foreground extraction which aims to segment out all foreground objects from the background in the image. The segmentation problem is formulated as an iterative energy minimisation of the conditional random field (CRF), which can be efficiently optimised by graph-cuts. The energy minimisation is initialised and modulated by a soft location map predicted by a discriminative classifier which is learned on-the-fly from a set of segmented exemplar images. Iteratively minimising the CRF energy leads to optimal segmentation. Experimental results on the Pascal visual object classes (VOC) 2010 segmentation dataset, a widely acknowledged difficult dataset, show that the proposed approach outperforms the state-of-the-art techniques.

Introduction: Foreground extraction is of great practical importance in a variety of applications in image processing and computer vision, such as object recognition, object tracking, content-based image retrieval and image editing. According to the need for human involvement, foreground extraction can be classified into two categories: interactive foreground extraction and the automatic one. The interactive foreground extraction, which requires the human being to roughly separate objects from the background like drawing a bounding box surrounding objects, has been extensively studied, e.g. [1, 2]; however, automatic foreground extraction is not fully investigated, mainly due to the difficulties to automatically localise objects in complex scenes.

Recently, exemplar images, in which foreground objects have been manually segmented from the background, are shown to be helpful cues for automatic foreground extraction. In [3], geometrically similar images are retrieved by comparing global image descriptors GIST [4], and the object location map of the input image is created by fusing segmentation masks of *k*-nearest exemplar images. Instead, in [5], the object location map is modelled via windows likely to contain objects of exemplar images. Such windows are detected by a state-of-the-art object detection algorithm. In this Letter, we term the former approach as ‘global transfer’, and the latter one as ‘window transfer’. The main and common drawback of these two approaches is the lack of ability to localise objects in a complex scene whose background is cluttered. On the one hand, the global transfer [3] suffers from limited robustness of GIST to handle geometric deformations. On the other hand, transferring windows to an image is sensitive to object variations, e.g. position, rotation and scale.

To address the aforementioned issues, this Letter proposes a novel approach which combines online learning with iterative energy minimisation of the conditional random field (CRF). Specifically, the object location map is predicted by a binary classifier which is online-learned from exemplar images whose ‘appearances’ are similar to that of the query image. Then, this location map is exploited to initialise and modulate the energy minimisation which leads to a binary labelling. The optimal segmentation is obtained when the minimisation reaches convergence.

Exemplar-based online prediction: To retrieve a set of exemplar images similar to the query image, we use scale-invariant feature transform (SIFT) [6] as a local feature for image representation. SIFT descriptors are extracted within 16 × 16 patches with a step-size of two pixels. Standard bag-of-features (BOF) is used as an intermediate representation of SIFT descriptors by associating each of them with one feature vector of the visual dictionary learned by *K*-means. The visual words are accumulated on three levels of a spatial pyramid (1 × 1, 2 × 2, 4 × 4) of image. Thus, an image is represented by a BOF vector with the dimensionality $D = (1 + 4 + 16) \times t$, where t is the size of the visual dictionary, set to 400 in our experiments.

With the BOF image representation, *k*-nearest exemplar images, issued from the training dataset, are retrieved for the query image by measuring χ^2 distances of images features. These exemplar images are then used to learn a region-based foreground/background classifier which is exploited to predict an object location map of the input image. For this purpose, the query image and its *k*-nearest exemplar images are segmented into superpixels by using the contour-based segmentation algorithm gPb [7] (globalised probability of boundary). To describe these superpixels, SIFT and self-similarity (SSIM) [8] are used as local feature descriptors. In addition, BOF along with a spatial pyramid (two levels) visual words accumulation on the superpixel are adopted, where the SIFT and SSIM dictionaries are with sizes of 2000 and 800, respectively. Therefore, a superpixel is represented by a feature vector which is the combination of BOF vectors of SIFT and SSIM.

For superpixel classification, a standard support vector machine (SVM) is utilised. Suppose that $\{\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_N\}$ are feature vectors extracted from N superpixels of *k*-nearest exemplar images, and their corresponding class labels are $\{y_1, y_2, \ldots, y_N\}$, $y_n \in \{+1, -1\}$, where $y_n = +1 (y_n = -1)$ indicates the superpixel n mainly belonging to the foreground (background). The primal SVM optimisation problem is

$$\min_{\boldsymbol{w}} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} \ell\left(y_n \boldsymbol{w}^{\mathrm{T}} \boldsymbol{f}_n\right) \tag{1}$$

where $\ell(u) = \max(0,\ 1 - u)$ is the hinge loss function, C is the regularisation constant set to 20 in our experiments, $\boldsymbol{w}$ is the separating hyperplane and can be obtained by solving formulation (1). With the learned $\boldsymbol{w}$, a test superpixel m represented by a feature vector $\boldsymbol{f}_m$ can obtain a classification score

$$s_m = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{f}_m \tag{2}$$

where s_m is typically in the range of [–3, 3], in which the positive value represents the superpixel m more likely to be the foreground and conversely the negative value indicates that it is the background. Therefore, the object location map of the input image can be obtained by computing the classification scores of all its superpixels and assigning these scores to corresponding pixels. To create a probabilistic location map, a sigmoid is fitted to each of these classification scores.

So far, foreground objects have been roughly extracted by thresholding the object location map. Unfortunately, the SVM classification scores are not always reliable as the training exemplars may not be matched well with the input image. In addition, the superpixels are predicted independently, without considering the relationship between neighbouring superpixels, and may result in noisy segmentation. To make the system more robust and obtain coherent segmentation, we propose a joint segmentation model based on the CRF and the object location map predicted by the SVM.

Segmentation with joint CRF and SVM: Given an input image $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ with object location map $\boldsymbol{S} = \{s_1, s_2, \ldots, s_N\}$, predicted by the online-learned SVM classifier, and contour map $\boldsymbol{E} = \{e_1, e_2, \ldots, e_N\}$, created by the gPb segmentation, foreground/background segmentation can be explicitly formulated as a binary labelling problem: finding a labelling set $L = \{l_1, l_2, \ldots, l_N\}$ to represent the segmentation of $\boldsymbol{X}$, where $l_n = 1 (l_n = 0)$ represents pixel $\boldsymbol{x}_n \in \boldsymbol{X}$ belonging to the foreground (background). The segmentation model is defined as energy minimisation of pairwise CRFs

$$\boldsymbol{E}(L) = \sum_{\{n,j\}\in\Upsilon} \Theta_{n,j}\left(l_n,\ l_j\right) + \sum_n \Lambda_n(l_n) \tag{3}$$

where Υ is a set of all pairs of neighbour pixels (four-way connectivity), $\Theta_{n,j}$ is the smoothness term which ensures that the overall labelling is smooth by considering the labels of neighbour pixels, Λ_n is the data term which measures the likelihood degree of the pixel to be labelled as foreground or background.

The smoothness term is defined as

$$\Theta_{n,j}\left(l_n,\ l_j\right) = \begin{cases} 0, & \text{if} \quad l_n = l_j \\ \Psi(n,\ j), & \text{otherwise} \end{cases} \tag{4}$$

where $\Psi(n, j)$ is a function defined based on contour map $\boldsymbol{E}$

$$\Psi(n,\ j) = \frac{\varphi}{\mathrm{dis}(n,\ j)} \exp\left\{-\beta\left(e_n - e_j\right)^2\right\} \tag{5}$$

here dis(·) indicates the spatial Euclidean distance between neighbouring pixels, the constant parameter φ is set to 50, β is defined as

$$\beta = \frac{1}{2\mathrm{mean}\left(\left(e_n - e_j\right)^2\right)} \tag{6}$$

From (5) we can observe that the segmentation boundary is promoted to be aligned with the contour computed by gPb.

The data term is derived from image colour distributions and object location map $\boldsymbol{S}$. It is defined as

$$\Lambda_n(l_n) = -\log(\Phi(l_n)\Omega(\boldsymbol{x}_n|l_n)) \tag{7}$$

where $\Phi(l_n)$ is the location prior of pixel n to be foreground or background, Ω is an appearance model predicting the foreground or background probability.

The location prior of pixel n for the foreground model is defined as

$$\Phi(l_n = 1) = s(n) \tag{8}$$

Similarly, the location prior of pixel n for the background model is defined as

$$\Phi(l_n = 0) = 1 - s(n) \tag{9}$$

Following the interactive foreground extraction algorithm of grabcut [2], the appearance model is characterised by two Gaussian mixture models (GMMs), one is for the foreground and the other one is for the background. Each of them is a full-covariance Gaussian mixture with five components.

The remaining problem is how to obtain the parameters of GMMs. In the interactive foreground extraction, these parameters are learned from the foreground/background pixels separated by a human being. However, in the automatic foreground extraction, human involvement is not allowed. The solution for this problem is to leverage the object location map predicted by the SVM. Those pixels with object location probability larger than a threshold value η are selected for foreground appearance modelling, and the other pixels are for background appearance modelling. The threshold value η is defined as

$$\eta = \min\,(\tau\mathrm{mean}(\boldsymbol{S}),\ \ \varpi\max\,(\boldsymbol{S})) \tag{10}$$

where τ and ϖ are predefined parameters, which are set to 0.8 and 0.6, respectively, in our experiments. With the well-defined smoothness and data terms, foreground extraction is obtained by iteratively minimising the energy function of (3) via efficient graph-cuts [1].

Experimental results: The proposed algorithm is evaluated on the Pascal visual object classes (VOC) 2010 segmentation dataset [9], which contains 1928 images with 20 object classes plus background. As in the previous works [3, 5], the standard training set, containing 964 images, is for training, and the remaining images are used for testing. Note that, the proposed approach is a generic foreground/background segmentation, whereas object categories information is not used. The segmentation performance is measured by an average union (AvU) metric defined as

$$\mathrm{AvU} = \frac{1}{T}\sum_{t=1}^{T}\frac{P_t \bigcap G_t}{P_t \bigcup G_t} \tag{11}$$

where T is the number of test images, P_t is the set of predicted foreground pixels of test image t and G_t is the ground-truth of the foreground.

For performance comparison, we use three baselines: plain grabcut [2] and two state-of-the-art foreground extraction algorithms which are global transfer [3] and window transfer [5]. The grabcut [2] is an interactive segmentation algorithm. To enable automatic segmentation, a centre box occupying 50% area of the image is used to initially separate the foreground pixels from the background pixels for the grabcut.

System performance is first evaluated by varying the number of nearest neighbours k. Fig. 1 summarises the evaluation. For comparison, results computed from the simple thresholding segmentation of the object location map created by the SVM prediction, and the performance of the global transfer [3], which is the most related to our approach, are also presented in Fig. 1. Clearly, even segmentation by only thresholding the object location map (curve B) is comparable with the global transfer [3] (curve C). The proposed full method (curve A), which integrates the object location map to the CRF energy minimisation, improves the performance further.

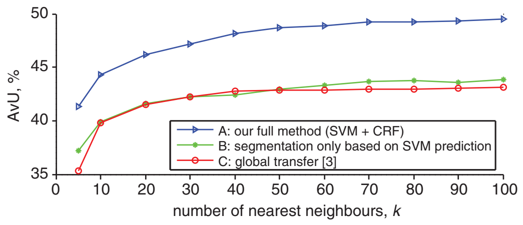

Fig. 1 *Segmentation performance by varying the number of nearest neighbours k*

Curve A shows the performance of our full method; curve B shows the performance of segmentation by thresholding the object location map created by the SVM prediction and curve C presents the performance of global transfer [3]

Table 1 shows the AvU scores of all methods. Among the baselines, window transfer [5] obtains the best result with 47.8% AvU; and global transfer [3], reported with $k = 40$, is ranked second with 42.8% AvU. As can be observed, the proposed approach improves the performance to 49.3% and achieves the best.

Table 1: AvU on Pascal VOC 2010 dataset

Method	Grabcut	Global transfer	Window transfer	Proposal
AvU (%)	30.0	42.8	47.8	**49.3**

Conclusion: We have presented a novel automatic foreground extraction algorithm which combines online learning and iterative CRF energy minimisation. The experimental evaluations demonstrate that the proposed approach consistently outperforms the state-of-the-art methods.

24 June 2013
doi: 10.1049/el.2013.2100
One or more of the Figures in this Letter are available in colour online.

W. Zou, K. Kpalma and J. Ronsin (*IETR, UMR CNRS 6164, INSA de Rennes, Université Européenne de Bretagne, France*)

E-mail: wenbin.zou@insa-rennes.fr

References

1 Boykov, Y., and Funka-Lea, G.: ‘Graph cuts and efficient *n–d* image segmentation’, *Int. J. Comput. Vision*, 2006, **70**, pp. 109–131
2 Rother, C., Kolmogorov, V., and Blake, A.: ‘Grabcut: interactive foreground extraction using iterated graph cuts’, *ACM Trans. Graph.*, 2004, **23**, pp. 309–314
3 Rosenfeld, A., and Weinshall, D.: ‘Extracting foreground masks towards object recognition’, *2011 IEEE Conf. Computer Vision*, Barcelona, Spain, November, 2011, pp. 1371–1378
4 Oliva, A., and Torralba, A.: ‘Modeling the shape of the scene: a holistic representation of the spatial envelope’, *Int. J. Comput. Vis.*, 2001, **42**, (3), pp. 145–175
5 Kuettel, D., and Ferrari, V.: ‘Figure-ground segmentation by transferring window masks’ Proc. IEEE Conf. Computer Vision and Pattern Recognition, Providence, RI, USA, June 2012, pp. 558–565
6 Lowe, D.G.: ‘Object recognition from local scale-invariant features’, *IEEE Int. Conf. Computer Vision*, Kerkyra, Greece, September 1999, Vol. 2, pp. 1150–1157
7 Arbelaez, P., Maire, M., Fowlkes, C., and Malik, J.: ‘Contour detection and hierarchical image segmentation’, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, **33**, (5), pp. 898–916
8 Shechtman, E., and Irani, M.: ‘Matching local self-similarities across images and videos’, Proc. IEEE Conf. Computer Vision and Pattern Recognition, Minneapolis, MN, USA, June 2007, pp. 1–8
9 Everingham, M., Van Gool, L., Williams, C.K., Winn, J., and Zisserman, A.: ‘The pascal visual object classes (VOC) challenge’, *Int. J. Comput. Vis.*, 2010, **88**, (2), pp. 303–338