Scale Invariant Auto-context for Object Segmentation and Labeling

Hongwei Ji*, Jiangping He and Xin Yang
Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240
[e-mail: hongweiji@sjtu.edu.cn]
*Corresponding author: Hongwei Ji

Received March 2, 2014; revised May 25, 2014; revised June 21, 2014; accepted July 7, 2014; published August 29, 2014

Abstract

In complicated environment, context information plays an important role in image segmentation/labeling. The recently proposed auto-context algorithm is one of the effective context-based methods. However, the standard auto-context approach samples the context locations utilizing a fixed radius sequence, which is sensitive to large scale-change of objects. In this paper, we present a scale invariant auto-context (SIAC) algorithm which is an improved version of the auto-context algorithm. In order to achieve scale-invariance, we try to approximate the optimal scale for the image in an iterative way and adopt the corresponding optimal radius sequence for context location sampling, both in training and testing. In each iteration of the proposed SIAC algorithm, we use the current classification map to estimate the image scale, and the corresponding radius sequence is then used for choosing context locations. The algorithm iteratively updates the classification maps, as well as the image scales, until convergence. We demonstrate the SIAC algorithm on several image segmentation/labeling tasks. The results demonstrate improvement over the standard auto-context algorithm when large scale-change of objects exists.

Keywords: Image segmentation, image labeling, context information, auto-context, scale invariance

This research was supported by the National Basic Research Program of China (No. 2010CB732506).

http://dx.doi.org/10.3837/tiis.2014.08.018

1. Introduction

Context and high-level information plays a very important role in image segmentation/labeling [1-9]. Many types of information can be referred to as context: different parts of an object can be context to each other; different objects in a scene can be each other's context. For example, a clearly visible horse's head may suggest the locations of its tail and leg, which are often occluded. A boat might suggest the existence of water [1].

In vision, models like Markov Random Fields (MRFs) [4,5] and Conditional Random Fields (CRFs) [6-9] have been widely used to capture the context information. Though MRFs and CRFs have been successfully applied in many applications, they still have some weaknesses. The main shortcoming is that they use a fixed neighborhood structure with a fairly limited number of connections. This property constrains their modeling capability and only short-range context is used in most cases.

The recently proposed auto-context algorithm [1] integrates image appearances together with the context information by learning a series of classifiers. There are two types of features for the classifier to choose from: (1) image appearance features computed on the local image patches, and (2) context features from a large number of sites on the classification maps. Given a set of training images and their corresponding label maps, the first classifier is learned based on image appearance features. The classification maps created by the learned classifier are then used as context information, along with image appearance features, to train the next classifier. The algorithm iterates to approximate the ground truth until convergence. In testing, the algorithm follows the same procedure by applying the sequence of learned classifiers to compute the classification maps. Compared to MRFs and CRFs, the auto-context algorithm is not limited to a fixed neighborhood structure. Each pixel can obtain support from a large number of neighbors (either short or long range), and the classifiers in different stages may choose different supporting neighbors. In [1], the auto-context algorithm was illustrated on several challenging vision tasks. The results demonstrated improved performance over many existing algorithms using MRFs and CRFs.

Although the auto-context algorithm is a powerful method, it is sensitive to large scale-change of objects. This is mainly because it samples the context locations utilizing a fixed radius sequence. In this paper, we present a scale invariant auto-context (SIAC) algorithm. We attempt to approximate the optimal scale for the image and use the corresponding optimal radius sequence to sample context locations, both in training and testing. At each round of the SIAC algorithm, we use the classification map created by the current trained classifier to estimate the image scale, and the corresponding radius sequence is then used to extract context features, which will be applied to train the next classifier. The algorithm iterates until convergence. Finally, we can obtain the best scale for the image, and the best radius sequence for extracting context features. We demonstrate the SIAC algorithm on several image segmentation/labeling tasks. The results demonstrate improvement over the standard auto-context algorithm when large scale-change of objects exists.

The main contribution of this paper is twofold. First, in order to achieve scale-invariance for the auto-context algorithm, we propose adopting different radius sequences to extract context features for images of different scales. Second, we use an iterative method to estimate and approximate the optimal scale for the image.

The remainder of this paper is structured as follows: Section 2 briefly reviews the standard auto-context algorithm. Section 3 describes the proposed SIAC algorithm in detail. Section 4 shows some comparative experiments on two challenging vision tasks. Section 5 concludes the paper.

2. Auto-context

In this section, we briefly review the standard auto-context algorithm proposed by Tu [1]. The algorithm takes into account the posterior distribution directly and integrates image appearances together with the context information by learning a series of classifiers.

In training, each image X comes with a ground truth Y. Given a set of training images and their corresponding label maps, $\{(Y_j, X_j), j=1..m\}$, where m denotes the number of training images. The algorithm first constructs a training set

$$S=\{(y_{ji}, X_j(N_i)), j=1..m, i=1..n\}, \tag{1}$$

where m is the number of training images, n is the number of pixels in each image, and $X_j(N_i)$ denotes the local image patch centered at pixel i in image X_j. The first classifier is learned based on the image appearance features computed on the local image patches $X_j(N_i)$. For each training image X_j, the classification maps P_j are then computed by the learned classifier. The algorithm then constructs a new training set

$$S'=\{(y_{ji}, (X_j(N_i), P_j(i))), j=1..m, i=1..n\}, \tag{2}$$

where $P_j(i)$ is the classification map centered at pixel i for image j. A new classifier is then trained, not only on the image features extracted from $X_j(N_i)$, but also on the context features extracted from $P_j(i)$. Once a new classifier is learned, the algorithm repeats the same procedure until convergence. Finally, the algorithm outputs a sequence of learned classifiers

$$p^{(t)}(y_i \mid X(N_i), P^{(t-1)}(i)), \tag{3}$$

where $P^{(0)}$ is a uniform distribution, and thus the context features are not selected by the first classifier, i.e., $p^{(1)}(y_i \mid X(N_i), P^{(0)}(i)) = p^{(1)}(y_i \mid X(N_i))$. In testing, the algorithm follows the same procedure by applying the sequence of learned classifiers to compute the classification maps. The auto-context algorithm iteratively updates the classification maps to approximate the marginal distribution $p(y_i \mid X)$. The convergence has been proved in [1].

In the auto-context algorithm, there are two types of features for the classifier to choose from: (1) image appearance features extracted from the local image patches, and (2) context features obtained from a large number of sites on the classification maps. In [1], a set of Haar features was employed as the main image appearance features, and a fixed image patch size 21 $\times$21 was used for their 2D application experiments. The context features are obtained from the classification maps from the previous iterations. For each pixel of interest, 8 rays in 45°

intervals are stretched out from the current pixel and a fixed radius sequence is then used for sparsely sampling the context locations on each ray. The classification probabilities on these locations are used as context features (both individual probabilities and the mean probabilities within a 3×3 window). **Fig. 1** gives an illustration.

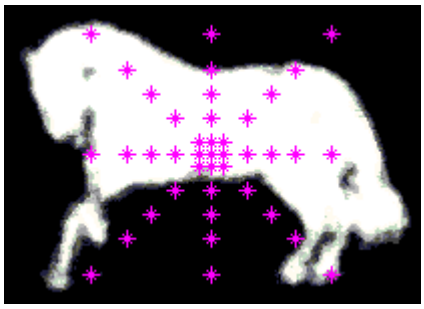


Fig. 1. An illustration of context features.

Regarding the choice of classifier, although the auto-context algorithm is not restricted to any specific choice of classifier, a boosting-based auto-context was adopted in [1], due to the natural feature selection and fusion capability of the boosting algorithms.

The auto-context algorithm makes an attempt to recursively select and fuse context information, as well as appearance, in a unified framework. The first trained classifier is based purely on the local appearance; objects with strong appearance cues are often correctly classified even after the first round. These probabilities then start to influence their neighbors, especially if there are strong correlations between them. In [1], the auto-context algorithm was illustrated on several challenging vision tasks. The results demonstrated improved performance over many existing algorithms using MRFs and CRFs.

3. Scale Invariant Auto-context

3.1 Motivation

Although the auto-context algorithm is a powerful method, it is sensitive to large scale-change of objects. This is mainly because it samples the context locations utilizing a fixed radius sequence, which can cause obvious feature inconsistency when large scale-change of objects exists. **Fig. 2** gives an illustration of feature inconsistency.

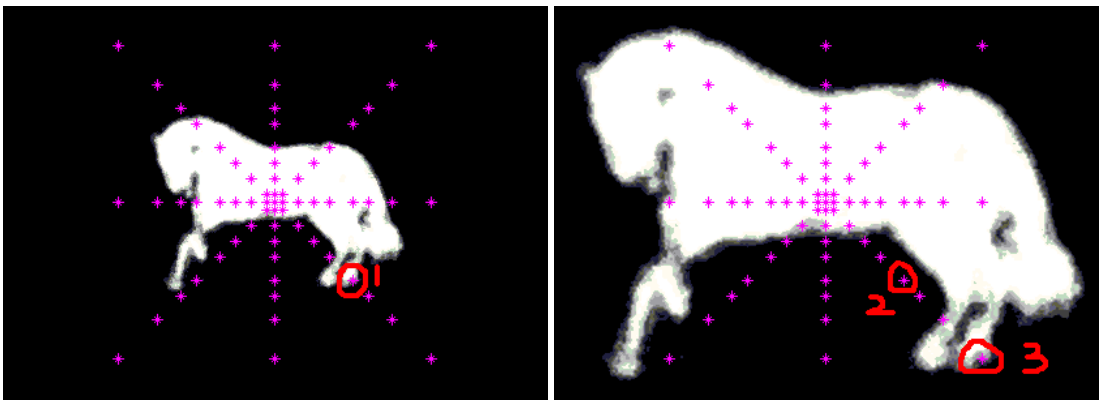


Fig. 2. An illustration of feature inconsistency. For example, by sampling the context locations according to a fixed radius sequence, Feature 1 is falsely matched to Feature 2. Actually, Feature 1 should be matched to Feature 3.

A direct method to tackle such a problem is trying to find the scale of objects beforehand. Then, for images of different scales, radius sequences of different sampling intervals are used for context location sampling. **Fig. 3** gives an illustration.

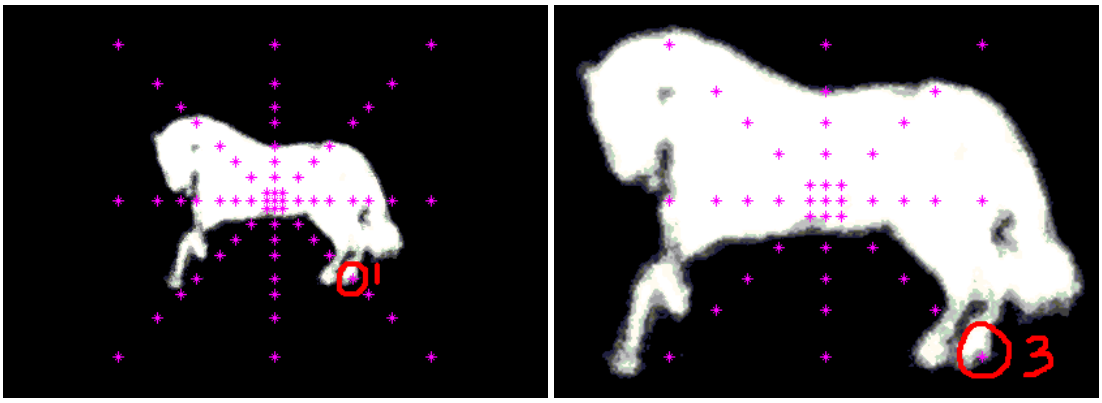


Fig. 3. For example, by adopting radius sequences of different sampling intervals to sample the context locations, Feature 1 is correctly matched to Feature 3.

However, in many cases, it is difficult to acquire the image scale through the image appearance directly without human interference. Notice that if we know the label map for an image, the scale of the image can be easily estimated. Since the auto-context algorithm is an iterative algorithm and produces an intermediate classification map at each round, we can iteratively estimate the image scale through these intermediate classification maps.

3.2 SIAC

In this section, we present a scale invariant auto-context (SIAC) algorithm, which is an improved version of the auto-context algorithm [1]. In order to achieve scale-invariance, we attempt to approximate the optimal scale for the image and use the corresponding optimal radius sequence to sample context locations, both in training and testing.

At each round of the SIAC training process, the classification maps $P_j^{(t)}$ created by the current trained classifier are used to estimate the image scale $a_j^{(t)}$ for each training image X_j, and the corresponding radius sequence $R(a_j^{(t)})$ is then used to extract context features, which will be used to train the next classifier. Here, $a_j^{(t)}$ denotes the estimated scale for image X_j at round t, $R()$ is a function of scale, and thus $R(a_j^{(t)})$ denotes the chosen radius sequence for image X_j at round t. The algorithm iterates until convergence. **Fig. 4** outlines the training procedure of the SIAC algorithm.

Given a set of training images together with their label maps, $\{(Y_j, X_j), j = 1..m\}$: For each image X_j, construct initial classification maps $P_j^{(0)}$ with uniform distribution on all the labels; Set the initial scale $a_j^{(0)} = a_{init}$ for all the images. For $t = 1,...,T$:

- Construct a training set $S_t = \{(y_{ji}, (X_j(N_i), P_j^{(t-1)}(i))), j = 1..m, i = 1..n\}$.
- Extract image appearance features from $X_j(N_i)$.
- Extract context features from $P_j^{(t-1)}(i)$ utilizing the radius sequence $R(a_j^{(t-1)})$.
- Train a classifier on both image and context features.
- Use the trained classifier to yield new classification maps $P_j^{(t)}$ for each training image X_j.
- Use the classification maps $P_j^{(t)}$ to estimate new scale $a_j^{(t)}$ for each training image X_j.

The algorithm outputs a sequence of trained classifiers for $p^{(t)}(y_i \mid X(N_i), P^{(t-1)}(i))$ **,** $t = 1,...,T$ **.**

Fig. 4. The training procedure of the SIAC algorithm.

In testing, the algorithm follows the same procedure by applying the sequence of learned classifiers to compute the classification maps. **Fig. 5** gives an illustration of the testing procedure of the SIAC algorithm.

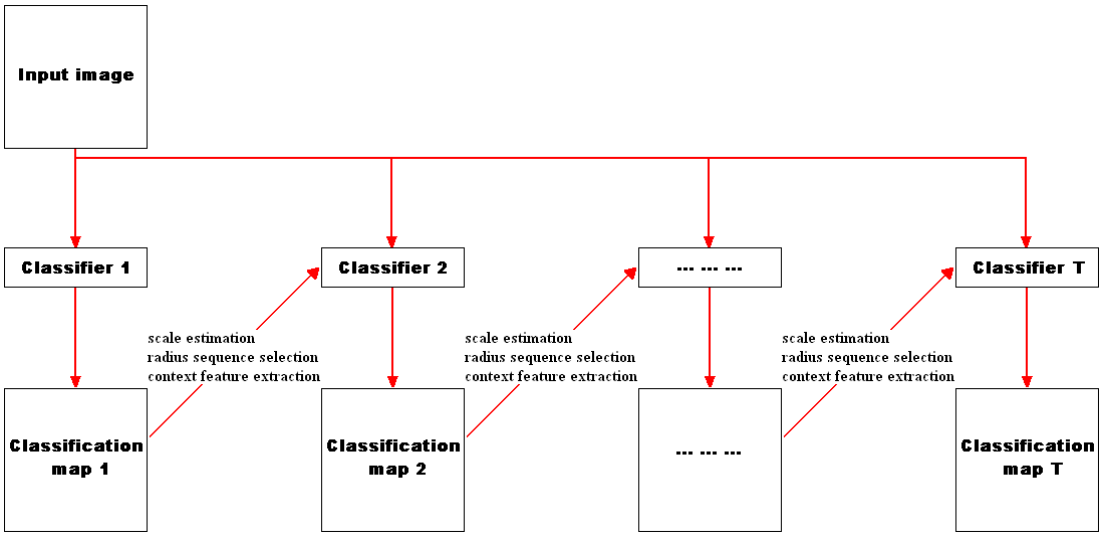


Fig. 5. An illustration of the testing procedure of the SIAC algorithm. The SIAC algorithm iteratively updates the classification maps, as well as the image scales, to approach the ground truth.

3.3 Scale Estimation and Radius Sequence Selection

In this section, we discuss several important implementation issues of the SIAC algorithm.

A. Scale space

Since the standard auto-context algorithm is only sensitive to large scale-change of objects, it

is not necessary to estimate the exact scale. In this paper, we simply consider three types of scales: “small”, “medium”, and “large”, i.e., the scale $a \in \{"small","medium","large"\}$, and we let $a_{init} = "medium"$.

B. Scale estimation

The SIAC algorithm iteratively updates the estimated scale to approximate the optimal scale for the image, both in training and testing. At each round of the algorithm, the intermediate classification map is used to estimate the image scale. Here, the image scale refers to the scale of foreground objects in the image. In this paper, we simply use the total number of foreground pixels to measure the image scale. **Fig. 6** outlines the scale estimation procedure at each round of the SIAC algorithm.

Given the intermediate classification maps for an image:

- Obtain the label map through the intermediate classification probability maps.
- Count the total number, N_F, of pixels labelled as foreground objects using the label map.
- Set the image scale $a = \begin{cases} "large" & if\ N_F \geq 4\theta \\ "small" & if\ N_F \leq \theta/4 \\ "medium" & otherwise \end{cases}$, where θ is the mean number of foreground pixels, which is calculated from the known label maps of all the training images.

Fig. 6. The scale estimation procedure at each round of the SIAC algorithm.

C. Radius sequence selection

At each round of the SIAC algorithm, we choose appropriate radius sequences to extract context features. For images of different scales, we adopt different radius sequences. Specifically, in this paper, we have

$$R("medium") = [0,2,4,6,8,10,12,16,20,24,30,36,42,50,60,70,80,90,100,120,140,160,180,200];$$
$$R("small") = R("medium")/2$$
$$= [0,1,2,3,4,5,6,8,10,12,15,18,21,25,30,35,40,45,50,60,70,80,90,100];$$
$$R("large") = R("medium")\times 2$$
$$= [0,4,8,12,16,20,24,32,40,48,60,72,84,100,120,140,160,180,200,240,280,320,360,400];$$

The procedures of the SIAC algorithm described in Section 3.2 are generic. The settings and the scale measurement described in this section will be applied to all experiments in this paper. However, one can slightly modify these settings to satisfy other different applications.

3.4 Understanding SIAC

The standard auto-context algorithm is sensitive to large scale-change of objects. In order to achieve scale-invariance, we proposed the scale invariant auto-context (SIAC) algorithm. By introducing the steps of scale estimation and radius sequence selection in each iteration, the algorithm makes an attempt to approximate the optimal scale for the image and use the corresponding optimal radius sequence to extract context features. For images of different scales, the algorithm adopts different radius sequences to extract context features, which can decrease the intra-class variation effectively. In theory, the smaller the intra-class variation is,

the better classification accuracy the classifier can achieve. Thus our SIAC can outperform the standard auto-context algorithm when large scale-change of objects exists.

4. Experiments

In this section, we illustrate the SIAC algorithm on two challenging vision tasks: horse segmentation and human body configuration.

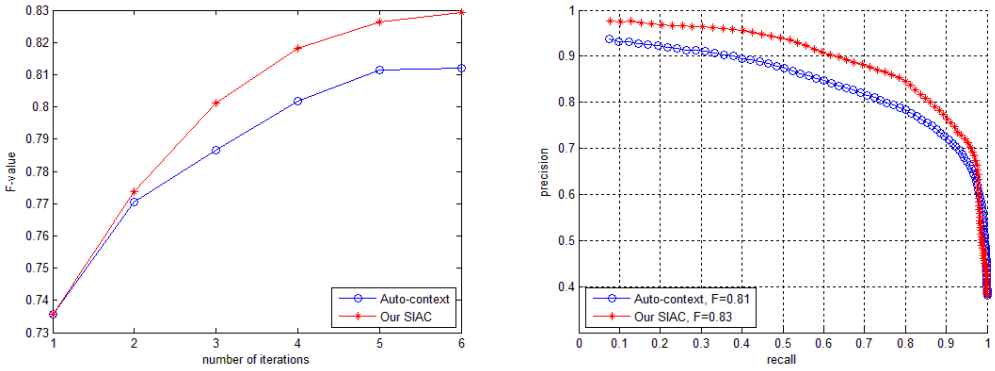
4.1 Horse segmentation

We use the Weizmann dataset consisting of 328 gray scale horse images [10]. The dataset also contains manually annotated label maps. Because the horses in the dataset have almost the same size, we randomly choose the sampling ratio to upsample or downsample all the images (and the corresponding label maps) in the dataset to create a new dataset, in which large scale-change of objects exists. Some images in the new dataset are shown in **Fig. 7**.

Fig. 7. Some images in the new dataset.

We randomly split the new dataset into two parts: half for training and half for testing. In this experiment, we employ Haar features as the image appearance features and AdaBoost [11]

as the basic classifier for both auto-context and our SIAC algorithms. **Fig. 8.a** shows the values of the F-measure [12] at different stages of both auto-context and our SIAC algorithms for horse segmentation and **Fig. 8.b** gives the corresponding overall precision-recall curves. **Fig. 9** shows some segmentation results. As we can see, by introducing the steps of scale estimation and radius sequence selection, our SIAC algorithm outperforms the standard auto-context algorithm when large scale-change of objects exists. **Fig. 10** shows the estimated scale at each iteration of the SIAC algorithm for horse segmentation. The initial estimated scale is "medium" and the SIAC algorithm iteratively updates the estimated scale to approximate the optimal scale for the image.



(a) (b)

Fig. 8. (a) shows the values of the F-measure at different stages of both auto-context and our SIAC algorithms for horse segmentation. (b) gives the corresponding overall precision-recall curves.

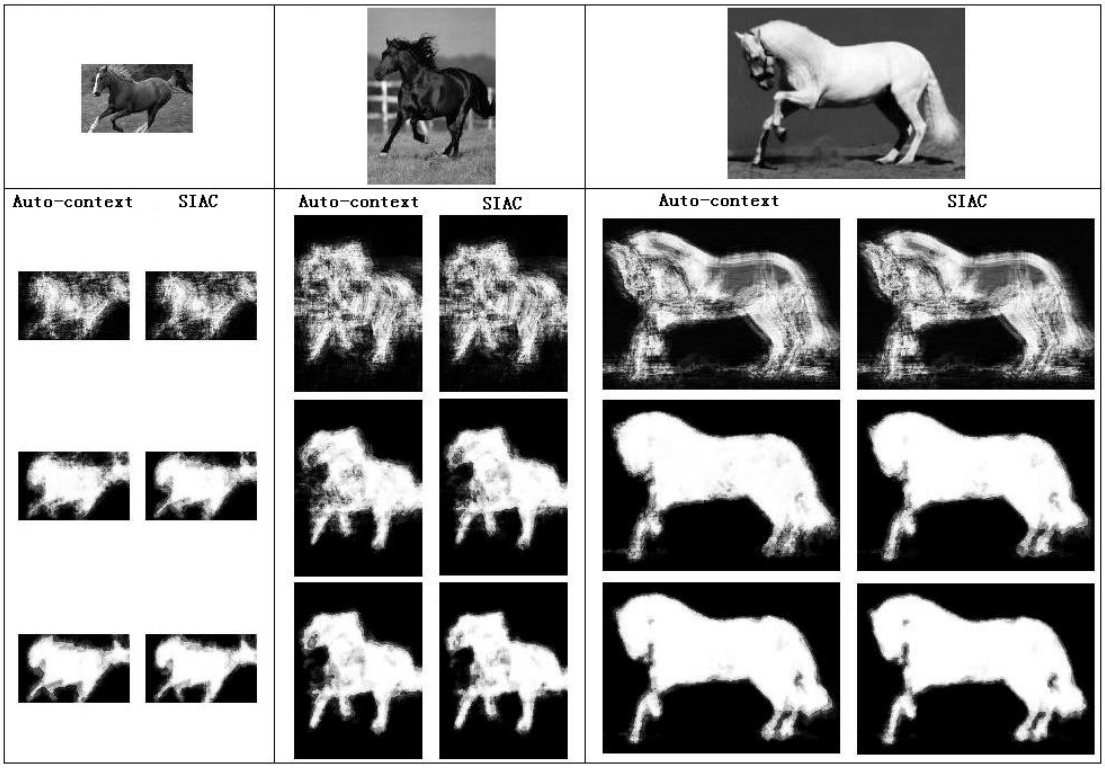


Fig. 9. The first row displays some test images. The second, third and forth row shows the classification maps by the first, third and fifth stage of the auto-context and SIAC algorithms.

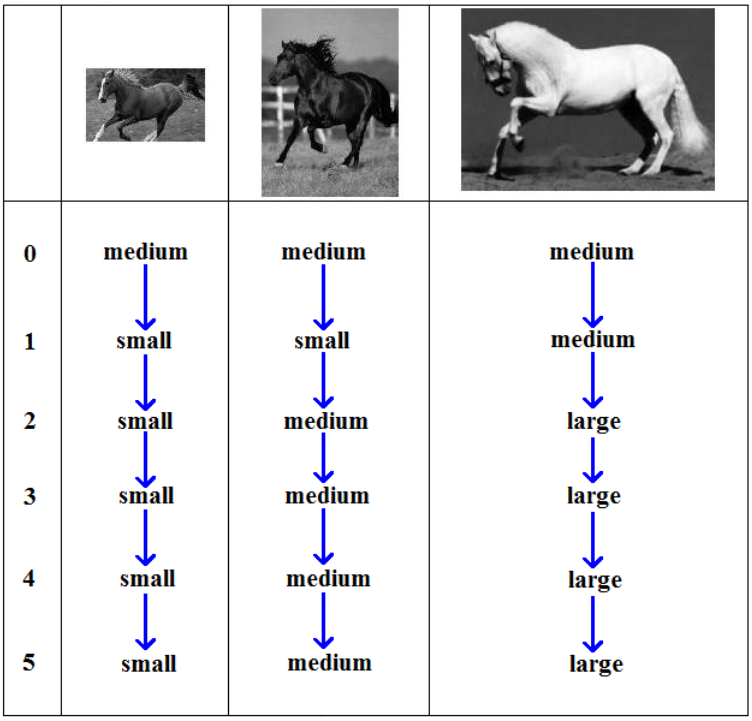


Fig. 10. The estimated scale at each iteration of the SIAC algorithm for horse segmentation.

4.2 Human Body Configuration

To further illustrate the effectiveness of our SIAC algorithm, we apply it on another problem, human body configuration. Each body part is assigned with a label and **Fig. 11** shows the template.

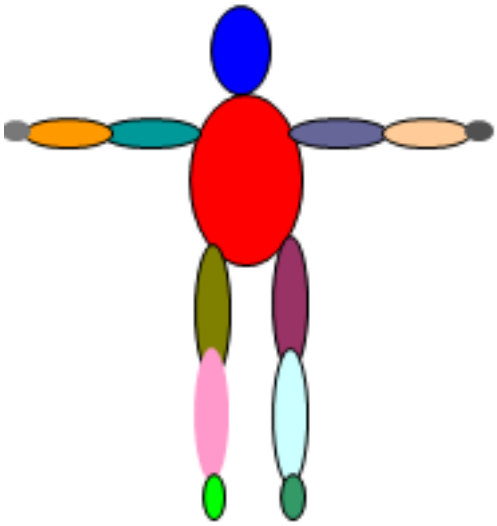


Fig. 11. A human body template, in which body parts are colored into 14 labels.

We collect around 80 images of baseball players and randomly upsample or downsample all the collected images to create a dataset. Similarly, the dataset is split into two parts: half for training and half for testing. In this experiment, we use the same set of features as in the horse segmentation problem, and adopt the one-vs-all strategy [13] to directly combine two-class AdaBoost classifiers into a multi-class classifier. **Fig. 12** shows the estimated scale at each iteration of the SIAC algorithm for human body labeling. The initial estimated scale is "medium" and the SIAC algorithm iteratively updates the estimated scale to approximate the optimal scale for the image. **Fig. 13** shows some labeling results at different stages of the auto-context and SIAC algorithms. In **Fig. 13**, for the baseball player on the left, the standard auto-context algorithm can not recognize the leg, while the proposed SIAC algorithm can label the leg well. For the player in the middle, the proposed SIAC algorithm can label the upper body and the head well, while the standard auto-context algorithm does not work. For the player on the right, the proposed SIAC algorithm can achieve better labeling results of the upper body than the standard auto-context algorithm. As we can see, our SIAC algorithm improves the results over the standard auto-context algorithm. The overall pixel-wise accuracy by 5 stages of SIAC is 78.9% which is better than 75.2% achieved by auto-context.

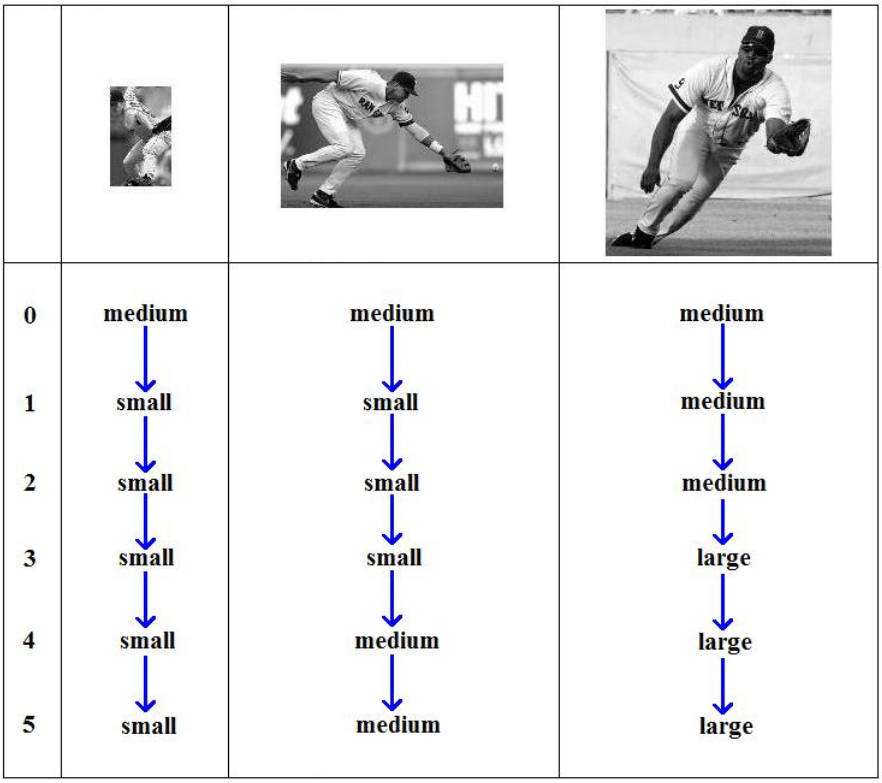


Fig. 12. The estimated scale at each iteration of the SIAC algorithm for human body labeling.

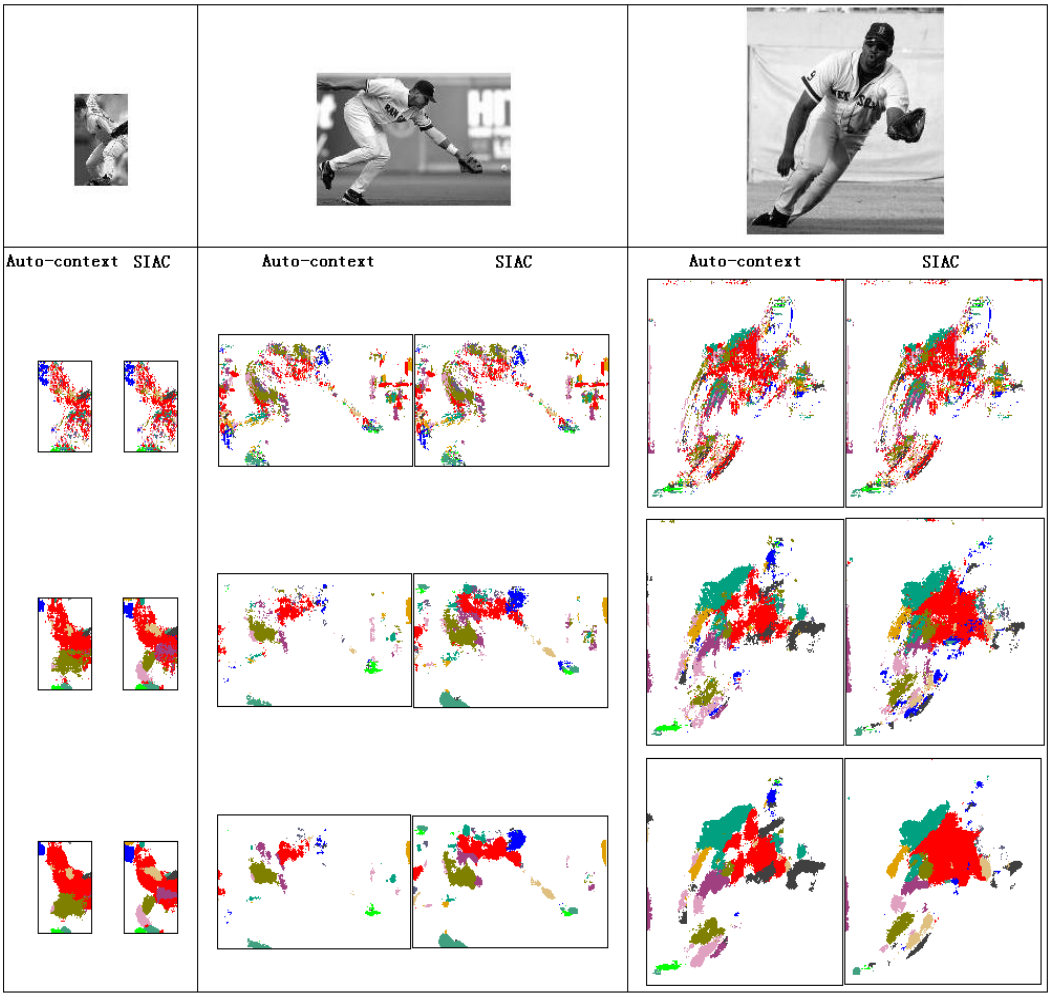


Fig. 13. The first row displays some test images. The second, third and forth row shows the classification maps by the first, third and fifth stage of the auto-context and SIAC algorithms.

5. Conclusions

In this paper, we have presented a scale invariant auto-context (SIAC) algorithm for image segmentation and labeling. By introducing the steps of scale estimation and radius sequence selection in each iteration, the algorithm makes an attempt to approximate the optimal scale for the image and use the corresponding optimal radius sequence to extract context features. We illustrate the SIAC algorithm on two challenging vision tasks. The results demonstrate improvement over the standard auto-context algorithm when large scale-change of objects exists. The future research directions include adopting different patch sizes to extract image features for images of different scales and achieving orientation invariance by orientation estimation.

References

[1] Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. PAMI*, 32:1744-1757, 2010. Article (CrossRef Link).

[2] F. Melgani and S. B. Serpico, "A statistical approach to the fusion of the spectral and spatio-temporal contextual information for the classification of remote sensing images," *Pattern Recognition Letters*, 23(9):1053-1061, July 2002. Article (CrossRef Link).

[3] F. Melgani, "Classification of multitemporal remote-sensing images by a fuzzy fusion of spectral and spatio-temporal contextual information," *International Journal of Pattern Recognition and Artificial Intelligence*, 18(2):143-156, February 2004. Article (CrossRef Link).

[4] S. Geman and D. Geman, "Gibbs distributions and the Bayesian restoration of images," *IEEE Trans. PAMI*, 6:721-741, Nov. 1984. Article (CrossRef Link).

[5] F. Melgani and S. B. Serpico, "A markov random field approach to spatio-temporal contextual image classification," *IEEE Transactions on Geoscience and Remote Sensing*, 41(11):2478-2487, Nov. 2003. Article (CrossRef Link).

[6] S. Kumar and M. Hebert, "Discriminative random fields: a discriminative framework for contextual interaction in classification," in *Proc. of ICCV*, Oct. 2003. Article (CrossRef Link).

[7] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: probabilistic models for segmenting and labeling sequence data," in *Proc. of 10th Int'l Conf. on Machine Learning*, pages 282-289, San Francisco, 2001. Article (CrossRef Link).

[8] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. of CVPR*, 2008. Article (CrossRef Link).

[9] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost: Joint appearance, shape and context modeling for mult-class object recognition and segmentation," in *Proc. of ECCV*, 2006. Article (CrossRef Link).

[10] E. Borenstein, E. Sharon, and S. Ullman, "Combining top-down and bottom-up segmentation," in *Proc. of IEEE workshop on Perc. Org. in Com. Vis.*, June 2004. Article (CrossRef Link).

[11] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. of Comp. and Sys. Sci.*, 55(1):119-139, 1997. Article (CrossRef Link).

[12] X. Ren, C. Fowlkes, and J. Malik, "Cue integration in figure/ground labeling," in *Proc. of NIPS*, 2005. Article (CrossRef Link).

[13] R. Rifkin and A. Klautau, "In defence of one-vs-all classification," *J. Mach. Learn. Res.*, 5:101-141, 2004. Article (CrossRef Link).

Hongwei Ji received the M.S. degree in pattern recognition and intelligent system, from the Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, China, in 2006. He is currently working toward the Ph.D. degree at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. His research interests include medical image analysis, machine learning and pattern recognition.

Jiangping He received the M.S. degree in signal and information processing, from Lanzhou University, Lanzhou, Gansu, China, in 2009. He is currently working toward the Ph.D. degree at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. His research interests include image processing and pattern recognition.

Xin Yang received the M.S. degree in control engineering from Northwestern Polytechnic University, Xi'an, China, in 1982, and the Ph.D. degree in applied sciences from Vrije Universiteit Brussel, Pleinlaan, Elsene, Belgium, in 1995. He is currently a Professor at the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China. His research interests include medical image analysis, visualization, and partial differential equations in image processing.