SCI INSTITUTE
TECHNICAL REPORT

Multi-scale Series Contextual Model for Image Parsing

Mojtaba Seyedhosseini[*,†], *António R. C. Paiva*[*], *Tolga Tasdizen*[*,†]
[*]SCI Institute, University of Utah
[†]Dept. of Electrical and Computer Eng., University of Utah

UUSCI-2011-004

Scientific Computing and Imaging Institute
University of Utah
Salt Lake City, UT 84112 USA

March 3, 2011

Abstract:

Contextual information plays an important role in solving high-level vision problems and has been used widely in the field. However, using contextual information in an effective way remains a difficult problem. To address this challenge, we propose a novel framework which utilizes context information in a multi-scale structure for learning discriminative models. We apply a series of linear filters to the context image consecutively to create a scale space representation. The main idea is to take the advantage of the context image at different scales instead of a single scale giving the classifier access to a larger contextual area. Moreover, finest scale context information can be noisy while a scale space structure is more robust against noise, so our proposed method improves robustness as well as accuracy. In this framework, the improvements in accuracy between consecutive classifiers in a series architecture are larger and convergence is faster. Our strategy is general and independent of the classifier type. In other words, it has the potential to be used in any context based framework. We demonstrate performance of the algorithm on two challenging visual recognition tasks: image parsing and texture segmentation. With nearly same computational complexity our model outperforms the state of the art algorithms.

THE UNIVERSITY OF UTAH

Multi-scale Series Contextual Model for Image Parsing

Mojtaba Seyedhosseini[1,2], António R. C. Paiva[1] and Tolga Tasdizen[1,2]

[1] *Scientific Computing and Imaging Institute,*

[2] Dept. of Electrical and Computer Eng.,

University of Utah, Salt Lake City, UT 84112

email: {mseyed,arpaiva,tolga}@sci.utah.edu

Abstract

Contextual information plays an important role in solving high-level vision problems and has been used widely in the field. However, using contextual information in an effective way remains a difficult problem. To address this challenge, we propose a novel framework which utilizes context information in a multi-scale structure for learning discriminative models. We apply a series of linear filters to the context image consecutively to create a scale space representation. The main idea is to take the advantage of the context image at different scales instead of a single scale giving the classifier access to a larger contextual area. Moreover, finest scale context information can be noisy while a scale space structure is more robust against noise, so our proposed method improves robustness as well as accuracy. In this framework, the improvements in accuracy between consecutive classifiers in a series architecture are larger and convergence is faster. Our strategy is general and independent of the classifier type. In other words, it has the potential to be used in any context based framework. We demonstrate performance of the algorithm on two challenging visual recognition tasks: image parsing and texture segmentation. With nearly same computational complexity our model outperforms the state of the art algorithms.

1. Introduction

Shape contexts are extremely rich descriptors [2] that have been used widely for solving high-level vision problems. Contextual information is interpreted as intra-object configurations and inter-object relationships [19]. These attributes play an important role in scene understanding [5, 16, 14]. For example, the existence of a keyboard in an image suggests that there is very likely a mouse near it [17]. To be precise, contextual information refers to the probability image map which can be used as prior information together with the original image information to solve the maximum aposterior (MAP) pixel classification problem.

There have been many methods that employ context for solving vision problems. Markov random fields (MRF) [6] is one of the earliest and most widespread approaches. Lafferty *et al*. [10] showed that better results for discrimination problems can be obtained by modeling the conditional probability of labels given an observation sequence directly. This non-generative approach is called the conditional random field (CRF). He *et al*. [7] generalized the CRF approach for the image labeling problem by learning features at different scales of the image. Torralba *et al*. [17] introduced boosted random field (BRF) which uses boosting to learn the graph structure of CRFs. Jain *et al*. [8] showed MRF and CRF algorithms perform about the same as simple thresholding in image restoration for binary-like images. They proposed a new version of convolutional neural network [11] strategy for restoring membranes in electron microscopic (EM) images. Compared to previous methods, convolutional networks take advantage of context information from larger regions, but need many hidden layers. Their model has over 34,000 free parameters, which can be problematic in the training step. Tu and Bai [19] proposed the auto-context algorithm which integrates the original image features together with the contextual information by learning a series of classifiers. Similar to CRF, auto-context targets the posterior distribution directly without splitting it to likelihood and prior distributions. The advantage of auto-context over convolutional networks is its easier training due to treating each classifier in the series one at a time in sequential order. Although they used probabilistic boosting tree as classifier (PBT), auto-context is not restricted to any particular classifier and different type of classifiers can be used. Jurrus *et al*. [9] employed artificial neural network (ANN) in a series classifier structure which learns a set of convolutional filters from the data instead of applying large filter banks to the input image.

Even though these approaches indeed improve the accuracy of the achieved segmentation by using context information together with the input image information, we propose that they don't utilize the context information in an effective way. This can be explained by the limitation of the features computed from the context information. Auto-context [19] uses boosting to sparsely sample some context locations from the candidate location pool. In [9], Jurrus *et al*. also utilize context locations which are selected by a stencil and use them as input to a neural network. The performance of the classifier can be improved by using context from a large neighborhood; however, it is not practical to sample every pixel in a very large context area because of computational complexity and the overfitting problem. Therefore, we develop a multi-scale strategy to take the advantage of context from a larger area while keeping the computational complexity tractable and avoiding overfitting. To address this problem, we apply a series of linear averaging filters to the context image consecutively to generate a scale space representation of the context and thus the classifier can have as input a small neighborhood, i.e. a 5×5 patch, at the original scale as well as the coarser scales. This strategy provides rich context information for the classifier. While scale-space methods are well known, to our knowledge their use to model context in classification problems is novel. Finally, we employ artificial neural networks in a series as in [9]. Our method doesn't depend on any particular classifier type.

Scale space filtering constructs a hierarchical descriptor of an image that provides multiresolutional information [1]. Different linear [1, 13] and nonlinear [15] scale space methods have been introduced. We simply use Gaussian filter as in [1] to generate scale space representation of the context. Again, our method is not restricted to any specific scale space method and different methods can be employed.

Combining scale space representation and contextual information provides a novel segmentation and parsing framework. The main advantage of this framework over other context based methods is providing more information from the context for the classifier in a similar number of features. This extra information from the context helps the later classifiers to correct the mistakes of the early stages and thus improves the overall performance.

2. Problem Formulation

Given a set of training images and corresponding ground truth labels for each pixel, we learn a set of classifiers in sequential order as in [19, 9]. The first classifier is trained only on the input image features. The output of this classifier, the probability image map, is used together with the input image features to train the next stage classifier. The algorithm iterates until the improvement in the performance of the current stage is small compared to the previous stage.

Let $X = (x(i,j))$ be the input image that comes with a ground truth $Y = (y(i,j))$ where $y(i,j) \in \{-1,1\}$ is the class label for pixel (i,j). The training set is $T = \{(X_k, Y_k); k = 1,\ldots,M\}$ where M denotes the number of training images. Given an input image X, the MAP estimation of Y for each pixel is given by:

$$\hat{y}_{MAP}(i,j) = \operatorname{argmax} p(y(i,j)|X) \tag{1}$$

A typical approximation of equation (1) is obtained by using the Markov assumption which decreases the computational complexity:

$$\hat{y}_{MAP}(i,j) = \operatorname{argmax} p(y(i,j)|X_{N(i,j)}) \tag{2}$$

where $N(i,j)$ denotes all the pixels in the neighborhood of pixel (i,j). Instead of using the entire input image, classifier has access to a limited number of neighborhood pixels at each input pixel (i,j).

In auto-context [19] and series-ANN [9], a classifier is trained based on the neighborhood features at each pixel. We call the output image of this classifier $C = (c(i,j))$. The next classifier is trained not only on the neighborhood features of X but also on the neighborhood features of C. The MAP estimation formula for this classifier can be written as:

$$\hat{y}_{MAP}(i,j) = \operatorname{argmax} p(y(i,j)|X_{N(i,j)}, C_{N'(i,j)}) \tag{3}$$

where $N^{'}(i,j)$ is the set of all neighborhood pixels of pixel (i,j) in the context image. Note that N and N' can be different neighborhoods. The same procedure is repeated through the different stages of the series classifier until convergence. It is worth mentioning that equation (3) is closely related to the CRF model; however, multiple models in series are learned which is an important difference from standard

CRF approaches. It has been previously shown that this approach outperforms iterations with the same model [19].

According to equation (3), context provides prior information to solve the MAP problem. Even though the Markov assumption is reasonable and makes the problem tractable, it still results in a significant loss of information from global context. However, it is not practical to sample every pixel in a very large neighborhood area of the context due to computational complexity problem and overfitting.Previous approaches [19, 9] have used a sparse sampling approach to cover large context areas as shown in Figure 2a. However, single pixel contextual information in the finest scale conveys only partial information about its neighborhood pixels in a sparse sampling strategy while each pixel in the coarser scales contains more information about its surrounding area due to averaging filters used. Furthermore, single pixel context is noise prone whereas context at coarser scales is more robust to the averaging effect. In other words, while it is reasonable to sample context at the finest level a few pixels away, sampling context at the finest scale tens to hundreds of pixels away is error prone and presents a non-optimal summary of its local area. We argue that more information can be obtained by creating a scale space representation of the context and allowing the classifier access to samples of small patches at each scale. Conceptually, sampling from scale space representation increases the effective size of the neighborhood while keeping the number of samples small.

3. Multi-scale Contextual Model

Figure 1 illustrates the multi-scale contextual model. Each stage is composed of two layers: a feature pooling layer and a classifier layer.

Classifier: Different types of classifiers can be used in auto-context architecture such as PBT and neural network. The first classifier operates only on the input image while the later stages are trained based on both the input image and context from the previous stage.

Feature Pooling: In conventional auto-context structure, the feature pooling layer simply takes sparsely sampled context as in Figure 2a and combines them with input image features. In our proposed method, the feature pooling layer treats each feature map as an image and creates a scale space representation by applying a series of Gaussian averaging filters. This results in a feature map with lower resolution

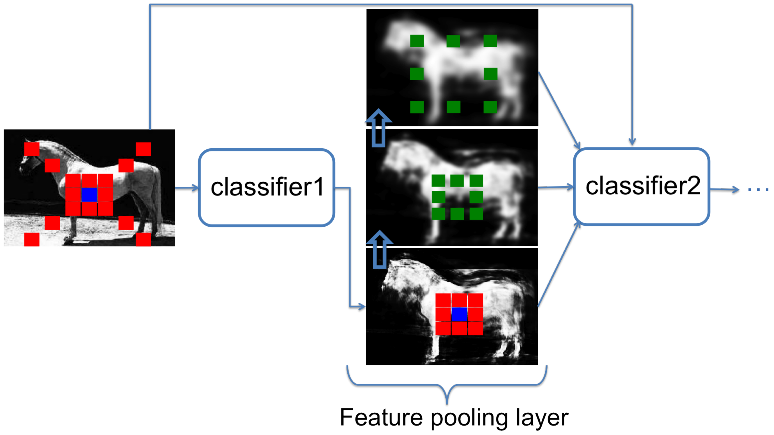

Figure 1. Illustration of the multi-scale contextual model. Each feature map is sampled at different scales (green rectangles). The blue rectangles represent the center pixel and the red rectangles show the selected context locations at original scale.

that is robust against the small variations in the location of features as well as noise. Figure 2 shows our sampling strategy versus single space sampling strategy. In Figure 2b the classifier can have as an input the center 3×3 patch at the original scale and a summary of 8 surrounding 3×3 patches at a coarser scale (The green circles denote the summaries of dashed circles). The green circles in Figure 2b are more informative and less noisy compared to their equivalent red circles in Figure 2a. The summaries become more informative as the number of scales increase. For example, in the first scale the summary is computed over 9 pixels (3×3 neighborhood) while it is computed over 25 pixels (5×5 neighborhood) in the second scale. In practice, we use Gaussian averaging filters to create the summary (green circles). Other methods like maximum pooling can be used instead of Gausian averaging [12]. The number of scales and Gaussian filter size are set according to the characteristics of the particular application.

From a mathematical point of view, equation (3) can be rewritten as:

$$\hat{y}_{MAP}(i,j) = \operatorname{argmax} p(y(i,j)|X_{N(i,j)}, C_{N'_0(i,j)}(0), C_{N'_1(i,j)}(1), \ldots, C_{N'_l(i,j)}(l)) \tag{4}$$

where $C(0), C(1), \ldots, C(l)$ denote the scale space representation of the context and $N'_0(i,j), N'_1(i,j), \ldots, N'_l(i,j)$ are corresponding neighborhood structures. Unlike equation (3) that uses the context in a single scale, equation (4) takes the advantage of multi-scale contextual information. Even though in

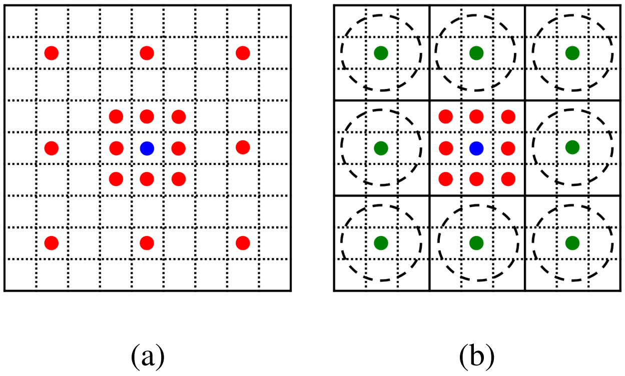

Figure 2. Sampling strategy of context: (a) Sampling at a single scale (b) sampling at multi scale. Green circles illustrate the summary of pixels in dashed circles. We use linear averaging to create the summary.

equation (4), we still use the Markov assumption, the size of the neighborhood is larger and thus we lose less information compared to equation (3).

The series multi-scale contextual model updates the equation (4) iteratively:

$$\hat{y}^{k+1}_{MAP}(i,j) = \operatorname{argmax} p(y(i,j)|X_{N(i,j)}, C^k_{N'_0(i,j)}(0), C^k_{N'_1(i,j)}(1), \ldots, C^k_{N'_l(i,j)}(l)) \tag{5}$$

where $C^k(0), C^k(1), \ldots, C^k(l)$ are the scale space representation of the output of classifier stage k, $k = 1, \ldots, K-1$ and $\hat{y}^{k+1}_{MAP}(i,j)$ denotes the output of the stage $k+1$. In turn, the $k+1$'th classifier output as defined in equation (5) creates the context for the $k+2$'th classifier. The model repeats equation (5) until the performance improvement between two consecutive stages becomes small. Because context is being used more effectively, the number of stages needed in the multi-scale contextual model is less than the auto-context and the series-ANN algorithms.

4. Experimental Results

We illustrate and verify the performance of our proposed method on two vision problems: Horse segmentation and texture segmentation. In these experiments, a stencil is used to sample the input image at each pixel ($X_{N(i,j)}$). We create a scale space representation of the context by applying a series of Gaussian averaging filters of size 5×5, 7×7, 9×9 and 11×11 to the context. The corresponding standard deviations of the Gaussian filters are $\frac{2}{3}$, 1, $\frac{4}{3}$ and $\frac{5}{3}$ respectively. The classifier then gets as input the 5×5 patch at the original resolution ($C_{N'_0(i,j)}(0)$) and 5×5 patches at three/four coarser scales ($C_{N'_l(i,j)}(l)$) depending on the application. For the first experiment we use ANN as the classifier and for the second experiment we compare PBT [18] and ANN as the classifier.

4.1. Horse Segmentation

We used the Weizmann dataset [3] containing 328 gray scale horse images with corresponding foreground/background truth maps. Similar to Tu *et al.* [19], we used half of the images for training and the remaining images were used for testing. In this experiment, we employed ANN in an auto-context architecture as in [9]. Each classifier in the series has one hidden layer with 30 nodes. Input image feature vectors were computed on a 31×31 sparse stencil centered on each pixel. The size of the feature vector is 57. The context features were computed using 5×5 patches at five scales (one at original resolution and four at coarser scales). The average $F - value = \frac{2 \times Precision \times Recall}{Precision + Recall}$ curves for outputs of stages are shown in Figure 3. The average F-value at threshold 0, at stage 5, is 87.3%. This result outperforms the Tu's result which is 84% [19]. It must be emphasized that the improvement from the first stage to the last stage in our method is 25.2% while the improvement in Tu's method is almost 5%. One can notice that we use a simple stencil to generate the input image feature vector instead of applying large filter banks to the input image as in [19] and our first stage F-value (62.1%) is less than Tu's first stage F-value(79%), but, our last stage result F-value is higher. This shows that multi-scale contextual model can compensate for the bad result of the first stage and improves the performance in later stages by using context in an effective manner

Figure 4 shows the average F-value at zero threshold for different stages and different number of scales. As we expected, the performance increases with the number of scales. Figure 5 shows some examples of our test images and their segmentation results at different stages of the multi-scale contextual model. As we can see, our method does a great job in parsing the horses from the background in most of the cases. In the first row, the side effects of the background are removed by the later stages. In the seventh row, the rider is removed gradually as the number of stages increases. Even in the last row which is a difficult case, The multi-scale contextual model does a fair job.

4.2. Texture Segmentation

As we mentioned before, the multi-scale contextual model is not restricted to any particular classifier and different type of classifiers can be used. In this section, we examine the performance of our model

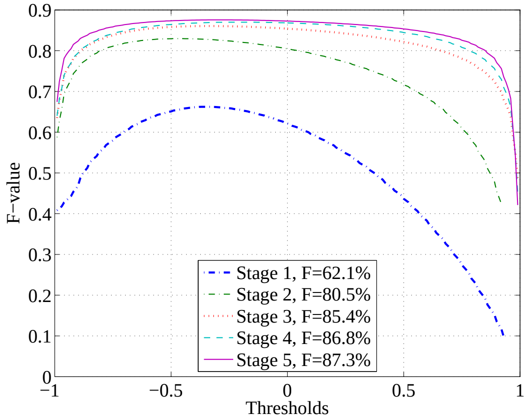

Figure 3. The average F-value curves for test images at different stages of the classifier for horse segmentation experiment.

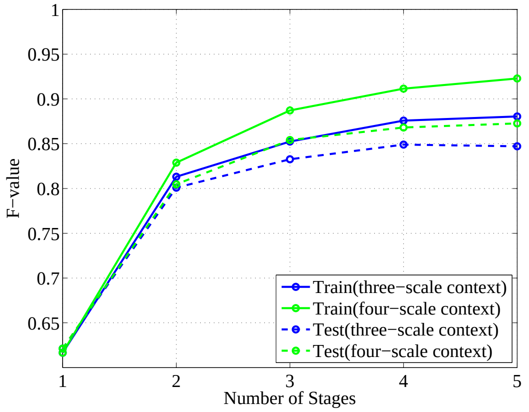

Figure 4. The average F-value at different stages of the multi-scale contextual model for training and testing images (horse segmentation). Using more scales improves the results.

for texture segmentation by selecting PBT [18] as the classifier. We use decision-tree with two nodes as the weak classifier in PBT.

In this dataset we have 20 images which are generated with four different textures for background and five different textures for foreground using textures from the Brodatz database [4]. A star shape is used to generate the mask for the foreground. Ten of these images were used for training and the remaining images were used for testing. We used the filter responses of different Gabor functions and a set of Haar-like features [20, 18] to generate the input image feature vector. The size of this feature vector is 276. By using this large vector the performance of the first stage classifier was improved 10% compared

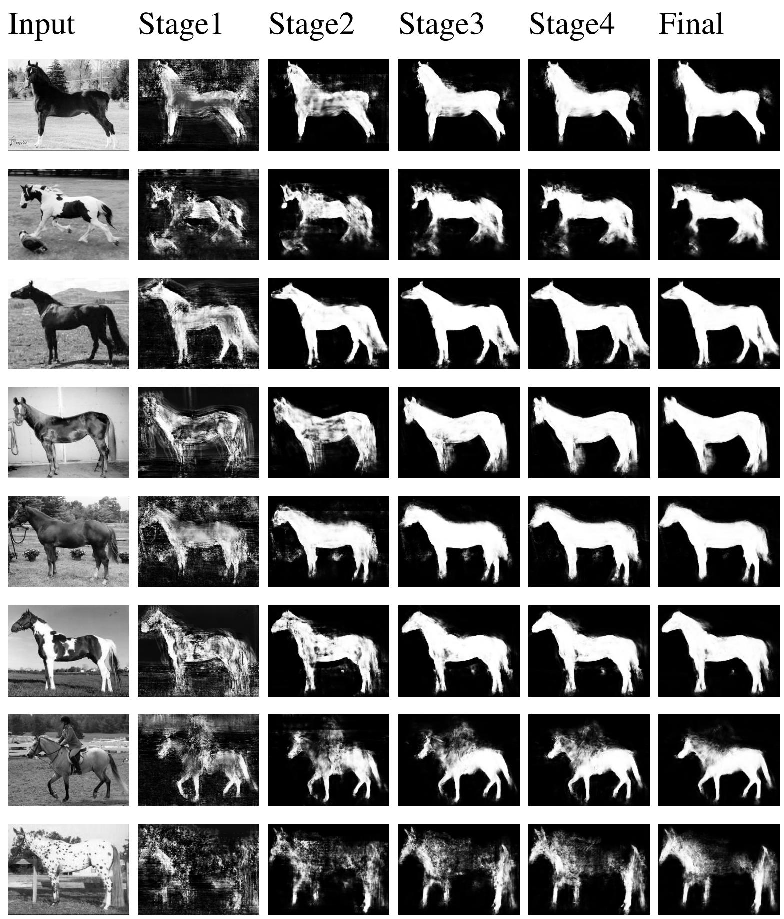

Figure 5. Test results for the horse segmentation experiment. The first column shows the input image and the remaining columns show the output at different stages of ANN based multi-scale contextual model.

to ANN where we used 11×11 stencil to generate the input image feature vector(Figure 7). The context features were computed on the 5×5 patch at three levels (one at original resolution and three at coarser scales). The average F-value curves for different stages are shown in Figure 6.

Figure 7 shows the average F-value at zero threshold for different type of classifiers and different number of scales. The performance of the classifier is improved significantly by using multi-scale context instead of single scale context. Comparing PBT based multi-scale with ANN based multi-scale, we can see that although the performance of the former at the first stage is significantly better than the performance of the latter, the final results after four stages are close. This verifies that while the single scale can't make up the first stage bad results, the multi-scale contextual model can better compensate for the bad results of the first stage. It is worth mentioning that the multi-scale contextual model converges

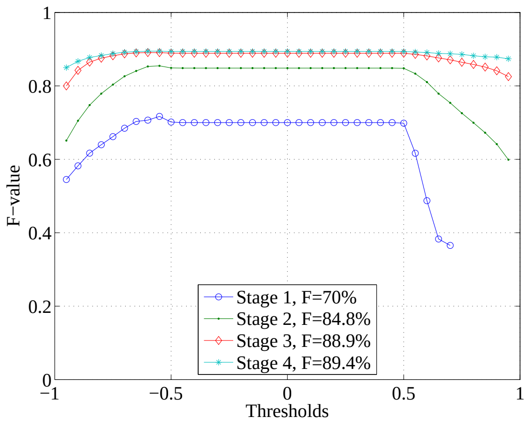

Figure 6. The average F-value curves for test images at different stages of the classifier using PBT for texture segmentation experiment.

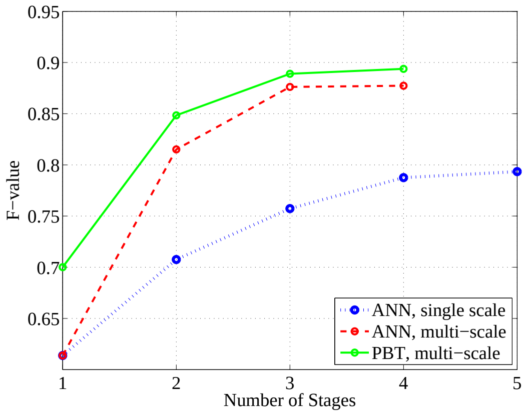

Figure 7. The average F-value at different stages of the multi-scale contextual model for testing images by different classifiers. Multi-scale context compensate the bad results of the first stage in later stages.

faster (four stages in this case) compared to the single scale model (five stages in this case).

Figure 8 shows the segmentation results at different stages of multi-scale contextual model for some test images and the last row shows the worst case. As we can see, even in this extremely challenging case the result is acceptable.

5. Conclusion

This paper introduced an image parsing algorithm using multi-scale contextual model. The main idea of the proposed method is to take the advantage of context image in different scales instead of a single

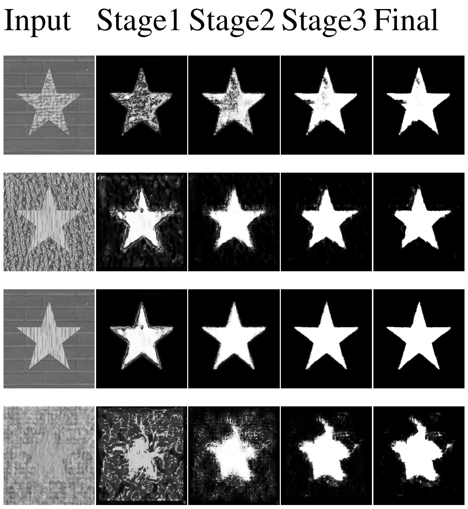

Figure 8. Test results for the texture segmentation experiment. The first column shows the input image and the remaining columns show the output at different stages of PBT based multi-scale contextual model.

scale. Our goal is to provide the classifier with a richer set of information by sampling the context image at different scales. The proposed method is very general and it doesn't depend on any particular classifier or any specific scale space method.

We applied our method to two challenging image segmentation tasks and simulation results indicate that, under nearly identical computational complexity, the proposed method outperforms bottom-up, state of the art algorithms. We used linear averaging filters to generate the scale space representation of the context with a typical depth of 3 or 4, however, a full study of the effect of scale space depth and the advantage of using other linear or nonlinear scale space methods is needed. The multi-scale contextual model can be extended to multi-class classification tasks by using the one-versus-all approach.

Although the multi-scale contextual model improves the classification performance, it has some limitations: (1)This is a supervised model and needs label maps for training images; (2)The computational complexity increases as the number of scales increases. Extensions to unsupervised learning will be investigate in future research. Different dimensionality reduction may be used to overcome the latter disadvantage.

References

[1] J. Babaud, A. P. Witkin, M. Baudin, and R. O. Duda. Uniqueness of the gaussian kernel for scale-space filtering. *IEEE Trans. on PAMI*, 8(1):26–33, 1986. 3

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on PAMI*, 24(4):509–522, 2002. 2

[3] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. *Proc. of CVPRW*, pages 46 –46, 2004. 8

[4] P. Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover, 1966. 9

[5] M. Fink and P. Perona. Mutual boosting for contextual inference. *NIPS*, 2004. 2

[6] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. on PAMI*, 6(6):721–741, 1984. 2

[7] X. He, R. Zemel, and M. Carreira-Perpinan. Multiscale conditional random fields for image labeling. *Proc. of CVPR*, 2:695–702, 2004. 2

[8] V. Jain, J. F. Murray, F. Roth, S. Turaga, V. Zhigulin, K. L. Briggman, M. N. Helmstaedter, W. Denk, and H. S.Seung. Supervised learning of image restoration with convolutional networks. *Proc. of ICCV*, pages 1–8, 2007. 2

[9] E. Jurrus, A. R. C. Paiva, S. Watanabe, J. R. Anderson, B. W. Jones, R. T. Whitaker, E. M. Jorgensen, R. E. Marc, and T. Tasdizen. Detection of neuron membranes in electron microscopy images using a serial neural network architecture. *Medical Image Analysis*, 14(6):770–783, 2010. 2, 3, 4, 5, 8

[10] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proc. of ICML*, pages 282–289, 2001. 2

[11] Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. *Proc. of CVPR*, 2:97 –104, 2004. 2

[12] Y. LeCun, K. Kavukcuoglu, and C. Farabet. Convolutional networks and applications in vision. *Proc. of ISCAS*, pages 253–256, 2010. 6

[13] T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, pages 224–270, 1994. 3

[14] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. *NIPS*, 2003. 2

[15] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. on PAMI*, 12(7):629–639, jul. 1990. 3

[16] A. Singhal, J. Luo, and W. Zhu. Probabilistic spatial context models for scene content understanding. *Proc. of CVPR*, 1:235–241, 2003. 2

[17] A. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. *NIPS*, 2004. 2

[18] Z. Tu. Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. *Proc. of ICCV*, 2:1589–1596, 2005. 8, 9

[19] Z. Tu and X. Bai. Auto-context and its application to high-level vision tasks and 3d brain image segmentation. *IEEE Trans. on PAMI*, 32(10):1744–1757, 2010. 2, 3, 4, 5, 8

[20] P. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004. 9