UTILIZING STRUCTURAL CONTEXT FOR REGION CLASSIFICATION

Zhiyong Wang
School of Information Technologies
The University of Sydney, Australia
zhiyong@it.usyd.edu.au

David D. Feng
School of Information Technologies
The University of Sydney, Australia
and
Department of Electronic and Information Engineering
Hong Kong Polytechnic University
feng@it.usyd.edu.au

Abstract In this paper, we propose to take structural context of image regions into account for region classification through a structural neural network. Firstly, a tree structure of each region is formed to characterize the relationship among the region and its neighbours. Such structures integrate both visual attributes of regions and their structural contexts. Then the structural representations are learned through a Back-propagation Through Structure (BPTS) training algorithm. Comprehensive experimental results demonstrate that our proposed approach has a great potential in region classification.

Keywords: Region Classification, Structural Context, Neural Networks

1. Introduction

While an ever increasing number of digital images play a more and more important role in improving the quality of daily life, users are also confronted with the difficulties in accessing specific images. Content-based image retrieval (CBIR) has been proposed and investigated to allow users to access images in terms of their true content, due to the great demand posed by the drastic growth of digital visual content (Smeulders et al., 2000). However, it is also realized that the semantic gap between low level visual features (e.g. color, shape, and texture) and semantic contents (e.g. objects and events) is

the biggest obstacle of the successful applications of image access (e.g. retrieval, filtering, and summarization) in terms of semantic contents. Automatic or semi-automatic image content understanding is a key to build intelligent image management systems. Image regions, which are meaningful primitives of images, contribute to semantic content of images significantly. In addition, region semantics can be utilized to derive high level semantic concepts. Therefore, it will be ideal to classify individual region into one of the semantic classes.

Various Pattern recognition approaches have been widely employed for region classification. In general, there are two key issues, feature extraction and classifier, involved in region classification. For example, based on visual features (e.g. color, texture, shape, size, and centroid), Campbell *et al.* proposed to classify image regions into semantic classes (e.g. sky, vegetation, and road) by using a three-layer neural network (Campbell et al., 1997). However, the performance of traditional region classification has been seriously limited due to segmentation noise and ambiguity of visual features (e.g. cloud vs. snow). On the other hand, contextual information of regions can be utilized to further improve the performance of region classification, since it is certain that the presence of some concepts or contents can provide important information for identifying other concepts or contents.

There are generally two types of contexts, conceptual context (i.e. global context) and content context, in region classification. Conceptual context is useful for modeling semantics at image level and can be utilized to increase the confidence of assigning certain labels to certain regions as well as the confidence of excluding some labels in terms of a given image theme. For example, it is much less possible to assign *grass* to a green region if an image has been identified as *indoors*. Conceptual context is generally obtained through image classification. For example, Vailaya *et al.* proposed a Bayesian classification approach to classify vacation images hierarchically (e.g. City vs. Landscape, Mountain vs. Coast)(Vailaya et al., 2001). Recently, conceptual context can also be derived through a set of words, since more and more images are accompanied with abundant annotations (e.g. web images). Therefore, many approaches consider extracting conceptual context as a problem of associating *a bag of words* with images by exploiting the co-occurrence of two modalities, visual attributes and labels, of images. The co-occurrence of those two modalities was first investigated by Mori *et al.* (Mori et al., 1999). It is assumed that a region corresponds to a label if they co-occur in images frequently. In (Barnard et al., 2003a), a translation model is proposed to translate a vocabulary of blobs to a vocabulary of terms based on the joint probability of images and terms, and a probabilistic model was established to classify each region into one of the terms. However, such classification is only a by-product of

Figure 1. An illustration of an adjacency structure of region 1.

the model, since not all the textual labels correspond to a particular region or object.

Content context, which represents the context of individual regions, can also be employed to enhance region classification and even identify objects. In (Singhal et al., 2003), content context was represented as the spatial relationship (e.g. above and below) between regions. However, structural context based on spatial adjacency, which is seldom investigated, is also important in region annotation. For example, a white region can be labelled as *cloud* with higher confidence if it is surrounded by *sky* regions. In this paper, we propose to characterize such structural context existing among regions by forming an adjacency graph. In such a graph, each node representing a region receive two inputs, its visual features and structural context (i.e. connections among its neighbours). Therefore, both attributes and context are integrated seamlessly. As shown in our previous study (Wang et al., 2002)(Wang et al., 2004), this graph representation is also effective and efficient in characterizing image content with only a small number of features.

Neural networks have been proposed to process structural data and the back-propagation through structure (BPTS) algorithm can be employed to learn the tree-structure representation(Frasconi et al., 1998). Such an algorithm has been successfully utilized for scene classification (Wang et al., 2004). Therefore, in this paper, we employ such learning algorithm to perform the task of region classification.

2. Representation of Structural Context

It is noticed that human beings perceive the real world in a structure way so that both entities and their relationship can contribute to their content representation. For example, being told that a region is surrounded by "sea", we may think of "beach", island, and "ship". Therefore, the more structural context is available, the more accurate the classification will be. As a result, a formal representation needs to be formed to characterize such structural context for each region. As shown in Figure1, the neighbour regions of Region 1 form its

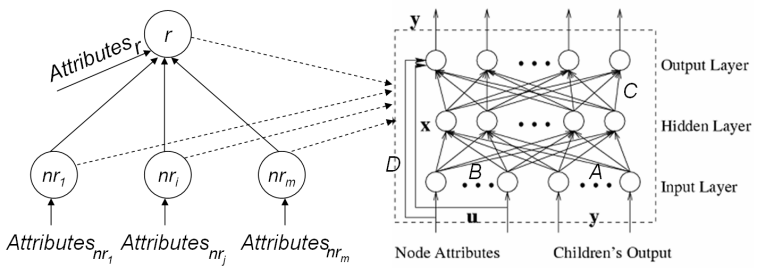

Figure 2. An illustration of a tree-structure encoding network with a single hidden layer.

structural context through a graph. Such structure representation can be noted as a graph $G = \{V, E\}$, where V and E indicate the set of nodes (i.e. regions) and edges (i.e. structural context among regions), respectively.

To process the graph representation, we need to figure out what structure information is and how to model it for each region class. In general, any relationship among regions can be abstracted as structural information, such as spatial relationship and visual similarity. In (Chang et al., 2004), it was proposed to explicitly utilize graph matching methods based on the similarity assigned to each edge and graph isomorphism. As explained in the next section, we employ a structural neural network model to process such structural representation adaptively.

3. Back-propagation Through Structure (BPTS)

Connectionist models have been successfully employed to solve learning tasks characterized by relatively poor representations in data structure such as static pattern or sequence. Most structured information presented in real world, however, can hardly be represented by simple sequences. Although many early approaches based on syntactic pattern recognition were developed to learn structured information, devising a proper grammar is often a very difficult task because domain knowledge is incomplete or insufficient. On the contrary, the graph representation varies in the size of input units and can organize data flexibly. An encoding process of a tree structure is shown in Figure 2. Each node represents a neural network on the right of Figure 2 and all the nodes share the same set of parameters. Neural networks for processing data structures have been proposed by Sperduti (Sperduti and Starita, 1997). It has been shown that they can be used to process data structures using an algorithm namely back-propagation through structure(BPTS). The algorithm extends the time unfolding carried out by back-propagation through time(BPTT) in the case of sequences. A general framework of adaptive processing of data structures was introduced by Tsoi (Tsoi, 1998) and Frasconi *et al.* (Frasconi

et al., 1998). Considering a generalized formulation of graph encoding shown in Figure 2, we have

$$\mathbf{x} = \mathbf{F}_n(\mathbf{A}q^{-1}\mathbf{y} + \mathbf{B}\mathbf{u}) \tag{1}$$

$$\mathbf{y} = \mathbf{F}_p(\mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{u}) \tag{2}$$

where $\mathbf{x}$, $\mathbf{u}$ and $\mathbf{y}$ are respectively the n dimensional output vector of the n hidden layer neurons, the m dimensional inputs to the neurons, and the p dimensional outputs of the neurons. q^{-1} is merely a notation to indicate that the input to the node is taken from its children. The $\mathbf{A}$ matrix is defined as follows:

$$\mathbf{A} = [\mathbf{A}^1\mathbf{A}^2 ... \ \mathbf{A}^c] \tag{3}$$

where c is the maximal out degree of the graph. $\mathbf{A}^i, i = 1, 2, ..., c$ is an $n \times p$ matrix, and is formed from the vector a^i_j, $j = 1, 2, ..., n$. $\mathbf{A}$ is a $c \times (n \times p)$ matrix. And $\mathbf{B}$, $\mathbf{C}$, and $\mathbf{D}$ are respectively matrices of dimensions $n \times m$, $p \times n$ and $p \times m$. $\mathbf{F}_n(.)$ is an n dimensional vector given as follows:

$$\mathbf{F}_n(\alpha) = [f(\alpha)\ f(\alpha)\ ...\ f(\alpha)]^T \tag{4}$$

where $f(.)$ is the nonlinear function such as a Sigmoidal function.

Note that we have assumed only one hidden layer in the formulation, because a single hidden layer with sufficient number of neurons is a universal approximator (Scarselli and Tsoi, 1998).

The training process is to estimate the parameters $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$ from a set of input/output samples by minimizing the cost criterion:

$$J = \frac{1}{2}\sum_{i=1} N_T ||\mathbf{d}_i - \mathbf{y}_i||^2 \tag{5}$$

where $\mathbf{y}_i$ denotes the output of the root of the i-th sample, $\mathbf{d}_i$ denotes the desired output of the i-th sample, and N_T is the number of the samples. The derivation of the training algorithm minimizing the cost criterion (5) will follow a fashion similar to gradient learning by computing the partial derivation of the cost J with respect to $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ and $\mathbf{D}$.

4. Experiments and Discussions

The image database used in our experiments has 304 images taken by ourselves, half of which is used for training, the other half for testing. A sample of each category is shown in Figure 3. All the images are segmented by using EdgeFlow technique (Ma and Manjunath, 2000) since the segmentation can be finely tuned by specifying different scales σ of Gaussian functions. By setting σ to 4, 3064 training regions and 2091 test regions are obtained. These regions are manually labelled with a set of terms. The region classes with less than 20

Figure 3. Samples of the image database.

instances have been removed. Finally, we identified 13 region classes, *auditorium, building, field, flower, grass, ground, people, sand, sky, stone, tree, wall, water*. Each region is characterized with 7-dimension features including the number of colors, percentage of the three most dominant color, average pixel values, standard deviation of pixel values, and region size.

Neighbour regions are not equally important in modeling spatial context, which should be taken into account for structure context. For example, sky region is more informative than building region in classifying a region as mountain. Furthermore, due to the error-prone segmentation, some neighbors are not true neighbors. In order to select important and representative neighbor regions, the length of the boundary between a neighbor region and the target region is considered to investigate the impact of different neighbor regions. In our experiments, the top $M, M = 0, ...5$ regions with the longest boundary length other than the biggest region size will be studied. While M is set to 0, the experiment is the baseline.

Three experimental tasks have been conducted to evaluate the performance of our proposed approach. At first, our approach is benchmarked with neural network methods by using multi-layer perceptrons. Then, the impacts of different segmentation and different visual features are investigated.

Performance Against Multi-layer Perceptron

We compare the proposed approach with the classical pattern classification approach, multi-layer peceptron (MLP). In order to make the comparison fair, we also consider neighbor information by concatenating feature vectors of neighbor regions into a higher dimension feature vector in the MLP method. That is, the feature vector is in $(N + 1) \times d$-dimension, if N neighbor regions are taken into account and each region is represented with a d-dimension feature vector. In this evaluation, regions are segmented by setting σ to 4 and represented with 7-dimension features, and neighbor regions are selected in the descending order of the length of the boundary adjacent to the target re-

Table 1. Comparison between the proposed approach and the MLP method on the test set

	MLP Approach		Proposed Approach	
	Accuracy (%)	#Hidden neurons	Accuracy (%)	#Hidden neurons
0-neighbor	54.15	15	N/A	N/A
1-neighbor	54.53	20	62.36	20
2-neighbor	60.46	20	61.08	10
3-neighbor	61.65	10	60.46	10
4-neighbor	63.25	10	56.08	10
5-neighbor	55.46	15	57.50	15

gion. An MLP with single hidden layer is adopted in this evaluation, since it can be a universal approximator provided with a sufficient number of hidden neurons (Scarselli and Tsoi, 1998). In order to tune the performance of the MLP method, we vary the number of hidden neurons from 5 to 20 and choose the best performance in each case.

As shown in Table 1, our proposed approach clearly outperforms the MLP method while not many neighbour regions (e.g. 1 or 2 neighbour regions) are utilized. In particular, the performance increases 14% while one neighbour region is utilized. It is also noticed that utilizing more neighbor regions is not always helpful, because the performance of both our proposed approach and the MLP method decreases while 5 neighbour regions have been utilized. For example, the performance of 5-neighbor (55.46%) is not as good as that of 2-neighbor (60.46%) for the MLP method. Such experimental results coincide with our assumption that not all the neighbour regions equally contribute to the classification task. More neighbour regions may add noise into the training session and demands higher learning capacity from classifiers. It is noticed that the most significant performance improvement happens while only one neighbor is taken into account. Therefore, it is essential to identify the most informative neighbor regions more effectively, other than simply using the boundary length, to further improve the performance.

Table 1 also shows that MLP methods achieve higher accuracy exceptionally while 4 neighbors are considered. The reason may be that our current database favors the MLP method for such a particular case. For our proposed approach, the classifier learns both structural information and region attributes, which requires more representative training data. Further research on this issue will be conducted.

Impact of Different Segmentation

Segmentation under different conditions generally introduces variations in region extraction and spatial context. As shown in Figure 4, images will be over-segmented at small scales and less over-segmented at great scales. In or-

Figure 4. Segmentation samples at different scales. (a) and (b) σ = 4; (c) and (d) σ=12;

Table 2. The number of regions of images segmented at different scales

	Training Set	Test Set	# of Region Classes
σ=4	3064	2901	13
σ=12	1972	1807	13

Table 3. Classification accuracy (%) of different segmentation

	1-neighbor	2-neighbor	3-neighbor	4-neighbor	5-neighbor
σ=4	62.36	61.08	60.46	56.08	57.50
σ=12	61.21	61.10	60.43	60.10	58.99

der to evaluate the impact of different segmentation, images are segmented by setting σ to 4 and 12 since these settings can generate a reasonable number of homogeneous regions for our image set. The number of training regions, test regions, and the number of region classes are listed in Table 2 for different segmentation scales, respectively. Obviously, segmentation at scale 4 generates more regions than at scale 12. As shown in Table 3, both segmentations can achieve similar performance. There are also two differences between them. First, performance of a larger scale (i.e. σ=12) decreases slighly. It may be that over-segmentation is reduced while segmentation scale increases. Hence, each segmented region is less homogeneous, which demands efficient content representation through visual feature extraction. As a result, we also investigated the impact of using different visual features. Second, the proposed approach is more robust at a larger scale. As can bee seen in Table 3, the classification accuracy of the segmentation at scale 12 is more around 60%. It may be that less over-segmentation introduces less variation for neighbor structures and makes learning slightly easier. Hence, additional experiments will be conducted to explore these discoveries.

Impact of Different Features

Besides the 7-dimension features, five more features including averages of R, G, B components and region centroid (x, y) are used to evaluate the impact

Table 4. Impact of different feature sets at segmentation scale 4

Dimension	1-neighbor	2-neighbor	3-neighbor	4-neighbor	5-neighbor
7	62.36	61.08	60.46	56.08	57.50
12	75.11	75.35	71.49	73.22	70.15

Table 5. Impact of different feature sets at segmentation scale 12

Dimension	1-neighbor	2-neighbor	3-neighbor	4-neighbor	5-neighbor
7	61.21	61.10	60.43	60.10	58.99
12	73.82	73.60	72.16	71.94	69.29

of different feature sets. As shown in Tables 4 and 5 where the best performance of each case is listed, much better performance has been achieved while the new 12-dimension features are adopted. Compared with the 7-dimension features, the 12-dimension features present more helpful information (e.g. region centroid) and benefit region classification, although both of them are quite simple. It can be expected that more representative feature sets will further improve the performance of our proposed approach. As indicated in (Barnard et al., 2003b), color and texture are the most representative features for scenery images, we need to include more texture features such as oriented energy coefficients in our future study.

5. Conclusion and Future Work

A novel region classification approach is present in this paper. Such an approach integrates structural context of image regions and the unique and powerful learning capacity of the BPTS learning algorithm. Comprehensive experiments have been conducted to evaluate our proposed approach. Experimental results demonstrate that our proposed approach can gain significant improvement even when only one neighbour region is utilized. In addition, our proposed approach is robust to the selection of neighbour regions, if suitable segmentation can be obtained.

It is also observed that segmentation and visual features do affect the performance of the proposed approach. For example, more neighbour regions do not always contribute to better classification accuracy, since structural variation also increases the requirement of learning capacity. Therefore, it is worthwhile to investigate how to identify more salient neighbour regions more efficiently based on large scale image databases. Since segmentation, salient regions, and visual features are closely related and interact with each other, it is also essential to balance them to achieve optimal classification performance. Another extension to our current work is to discover the second order structure rather than the adjacency structure exploited here.

Acknowledgement

This research is supported by the ARC and UGC grants.

References

Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., and Jordan, M. (2003a). Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135.

Barnard, K., Duygulu, P., Guru, R., Gabbur, P., and Forsyth, D. (2003b). The effects of segmentation and feature choice in a translation model of object recognition. In *The IEEE International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 675–682. Wisconsin, USA.

Campbell, N. W., Mackeown, W. P. J., Thomas, B. T., and Troscianko, T. (1997). Interpreting image databases by region classification. *Pattern Recognition*, 30(4):555–563.

Chang, R.-F., Chen, C.-J., and Liao, C.-H. (2004). Region-based image retrieval using edge-flow segmentation and regioin adjacency graph. In *The IEEE International Conference on Multimedia and Expo (ICME2004)*, volume 1, pages 1883–1886. Taiwan.

Frasconi, P., Gori, M., and Sperduir, A. (1998). A general framework for adaptive processing of data structures. *IEEE Trans. on Neural Networks*, 9(5):768 – 786.

Ma, W.-Y. and Manjunath, B. S. (2000). EdgeFlow: a technique for boundary detection and image segmentation. *IEEE Trans. on Image Processing*, 9(8):1375–1388.

Mori, Y., Takahashi, H., and Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *The First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM99)*. Florida, USA.

Scarselli, F. and Tsoi, A. C. (1998). Universal approximation using feedforward neural networks: a survey of some existing methods, and some new results. *Neural Networks*, 11(1):15–38.

Singhal, A., Luo, J., and Zhang, W. (2003). Probabilistic spatial context models for scene content understanding. In *The IEEE International Conference on Computer Vision and Pattern Recognition*, volume 1, pages 235–241. Wisconsin, USA.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., and Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380.

Sperduti, A. and Starita, A. (1997). Supervised neural networks for the classification of structures. *IEEE Trans. on Neural Networks*, 8(3):714–735.

Tsoi, A. C. (1998). Adaptive processing of data structures:an expository overview and comments. Technique report, Faculty of Informatics, University of Wollongong, Australia.

Vailaya, A., Figueiredo, M. A. T., Jain, A. K., and jiang Zhang, H. (2001). Image classification for content-based indexing. *IEEE Trans. on Image Processing*, 10(1):117–130.

Wang, Z., Feng, D., and Chi, Z. (2004). Comparison of image partition methods for adaptive image categorization based on structural image representation. In *The 8th International Conference on Control, Automation, Robotics, and Vision*, pages 676–680. Kunming, China.

Wang, Z., Hargenbuchner, M., Tsoi, A. C., Cho, S. Y., and Chi, Z. (2002). Image classification with structured self-organizing map. In *IEEE International Joint Conference on Neural Networks (IJCNN2002)*. Hawaii, USA.