Robust Method to Compute Mutual-Spatial Feature for Image Parsing Problem

Thi Ly Vu and Chong Ho Lee
School of Information and Communication Engineering, Inha University, Incheon 402-751
Email: vuthily.tin2@gmail.com, chlee@inha.ac.kr

***Abstract*—The paper presents new method to improve computational performance by introducing the mutual spatial feature in order to make strong visual cue in image parsing problem based on non-parametric model. This feature models the spatial context and mutual information in our previous study [1] to enhance accuracy and performance of image parsing problem in calculating the probability of co-occurrence objects. The experimental results based on Matlab programming language using SIFTFlow and Barcelona datasets showed that the mutual-spatial feature is promising to refine image parsing problem.**

***Index Terms*—image parsing, mutual-spatial, Matlab, SIFTFlow**

I. Introduction

Image parsing is one of challenging problems in computer vision. It is of focus in many recent works [1], [2], [3], [4], [5], [6], [7]. Firstly as noted in [2], this issue is an incredible confusing of visual words, which means one region can be matched with another region from hundreds of different labels. Secondly, the scenes are random, the objects are assumed to appear randomly leading to huge object distribution. Thirdly, due to the limit of the number of object labels in a parsing model, thus it is hard to build a completely plausible model, since the number of objects in the real scenes is actually unlimited. Recently, several works based on nontable parametric models for image parsing are presented in literature [1], [2], [3], [4], [5], [6], [7]. Our system is inspired from the method introduced by Tighe and Lazebnik for image parsing using scalable nonparametric model in region levels [3]. Their system shows pioneer result using the K nearest neighbor method which is then become the basis for other researchers for improvement. There are several directions to improve the performance of MRF model in [3]. Among them, the probability of object co-occurrence is commonly utilized [2], [4]. Eigen *et al.* [5] improve MRF baseline model for image parsing problem by learning the weight of neighboring objects and Myeong*et al.* [3] build the graph based on context model representing object relationship. Recently, Joseph Tighe *et al.* [6] continue developing their work from [1] by learning Per-Exemplar Detectorsfeature. However, most of these works are time-consuming for learning the object relationship. In our work, the probability of object co-occurrence is computed by accumulating features in the dataset which does not have to train for parameters. In other hand, the spatial context and mutual information are both considered to provide the avenue of approach of object relationship. For example, the probability of "Car" above "Building" that can be occurred in other worksis avoided.

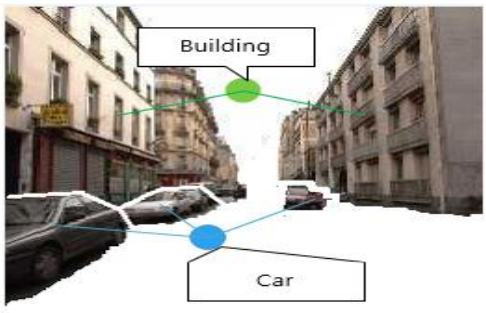

Figure 1. The construction of the mutual-spatial feature in our method. Each image only considers two objects that co-occurs at the same scene and all the other objects is ignored. The location of each object is computed by converge location of regions belonging to that object. Comparing between object locations gives the probability of the mutual-spatial feature. There are only two objects considered in one picture ("Building" and "Car"). Comparing location of "Car" regions and "Building" regions provides the probability of "Car" and "Building" in the location relation.

To improve the accuracy of the image parsing problem based on non-parametric model, the probability of object co-occurrence is commonly utilized. In image parsing area, the non-parametric model is the technique that the inference of a label in an image does not rely on the data belonging to any particular distribution, but it is based on the rank of observation. Therefore, the labels of the input image are understood through knowledge labels from a set of other images which are the most similar with this input image. For example, if one object label is already categorized as “street”, then it has a high probability to believe that the surrounding object labels are likely “car”, “sidewalk”, “building” etc. Hence, taking into account this mutual relationship, the accuracy of the image parsing model can be enhanced.

Our previous work [1] presents a novel approach for improving accuracy of the image parsing problem by using the spatial context and mutual information features in calculating the probability of object co-occurrence. The spatial context and mutual information that capture

Manuscript received October 23, 2013; revised May 13, 2014.

©2014 Engineering and Technology Publishing
doi: 10.12720/joig.2.1.54-58

the relationship of the object location and the frequency of co-occurrence objects in one image, is a strong visual cue. However, that approach calculates the spatial context and mutual information separately on pixel-level. This is simple computation method but considering all pixels in the image seems to be costly. Therefore, in present work, the spatial context and mutual information are combined as mutual-spatial feature based on region level. Our key contribution is that the mutual-spatial feature based on region level is introduced instead of using these two features separately. The spatial context and mutual information are considered simultaneously for every image by calculating the mutual-spatial feature. This combination provides a strong cue and reduces the time in computing the probability of co-occurrence objects. Therefore, this new method not only increases the accuracy rate but also reduces processing time for the image parsing problem.

This paper is organized as follows: Section 1 introduces in general image parsing reviewing previous methods and briefly describes our proposed method. The mutual-spatial feature is detailed in Section 2. Section 3 describes how to incorporate the mutual feature in image parsing problem. The experimental results are shown in Section 4 and Section 5 discusses about our proposed method.

II. Proposed Method

A. Contruct of the Region

In image parsing problem, as mentioned in Section 1, the distribution of object is not uniform, hence we prefers to use nonparametric model that infers a region from the most similar regions in the so-called retrieval set of image which contains K images most similar to the test image. Based on [8] our system also uses four types of global image features: Spatial pyramid, GIST, Tiny image and Color histogram to calculate the distance from each image in the training set to the test image. Then K images corresponding to smallest distances are selected to put into the retrieval set. An informative retrieval set should contain scene images similar to the test image.

As several approaches in image parsing area [1], [2], [5], [9], [10], [11], the labels are assigned in the region level. As mentioned in [2], this reduces computational load for the system. The region is produced by a segmentation algorithm. In this work, the fast graph-based segmentation algorithm [12] is applied for segmenting image into regions; and each region is presented by 20 features as in [3] which are calculated for every region in each image to measure the distance between regions in the test image and regions in the retrieval set. Fig. 1 shows the regions of "Car" and "Building" are the results of the segmentation algorithm. In fact, the "Building" can be segmented into many regions, so that the "Building" segment is summarized of all regions that belong to "Building" object.

B. The Mutual-Spatial Feature

In our previous work [1], the spatial context and mutual information are used to improve accuracy of image parsing problem by enhancing the reliability of the probability between co-occurrence objects. However, the spatial context and mutual information features are computed separately on the pixel level. The spatial context is calculated according to the work of Galleguillos *et al.* [13] which includes relation pairs: above/bellow. The probability of co-occurrence objects for one location relation is computed by accumulating number of pixels in the bolder between two neighboring objects. Therefore, this procedure must be scanned on every pixel of all images in the dataset, which is time-consuming. In addition, similar to other works [8, 10, 14],the mutual information is used as weight function (which can provide more information about the co-occurrence objects) to enhance the calculated feature. However, the separately use of these two features is computational expensive. In order to avoid these weakness points the new method is introduced in the present study.

Fig. 1 represents our new method to calculate the mutual-spatial feature. We assume that there are only two object labels *Li* and *Lj* in each image in the dataset. Each image is segmented into n regions: R1, R2,..., Rn, the coordinate of a region is the coordinate of region center. The location of object label Li in the image is calculated as average locations of all coordinate regions in the image that belong to object label Li.

$$Coor(L_i) = \frac{1}{\mathrm{m}} \sum_{k=1}^{m} \mathrm{Coor}(R_k) \quad (1)$$

The above/below relation is verified by comparing the coordinate of object labels. For example, in the above relation, the probability is computed by (3) considering condition (2):

$$IfCoor(L_i).y > Coor(L_j).y \quad (2)$$

$$ThenPr\ (L_i above L_j) = \frac{(w_i + w_j)}{w * h} \quad (3)$$

where w_i and w_j are the weight of labelsL_i and L_j(e.g. the size of L_i and $L_j$label regions in our work);w and h are the width and height of the image.

This process is performed for every image in the dataset. Then final probability of object label L_i and L_j is multiplied by the probability in every image as described in (4):

$$P(L_i\ above\ L_j) = \prod_{N\ images} Pr(L_i\ above\ L_j) \quad (4)$$

$$E_{mutualspatial} = -\log P(L_i above L_j) \quad (5)$$

In this method, each image is consideredonly for two object labels for calculating the spatial probability. Therefore the spatial context and mutual information are both considered in each computation. Because the probability of object spatial is considered as sum of object regions in one image, so that we only need to calculate "above" relation and do not have to consider "below" relation. The performance is highly improved because the spatial context is considered on region level resulting finding on every pixel of the image and checking that objects are neighbor or not are skipped.

©2014 Engineering and Technology Publishing

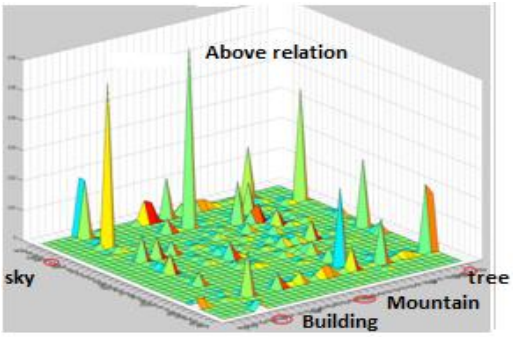

Figure 2. The probability of location relation between two objects. The sky has high probability to occur above building, mountain and tree.

This is the key contribution for our method to speed up computing process while the accuracy of the problem is maintained.

Fig. 2 shows the probability of two objects in "above" relation. This value provides both the mutual information and spatial context of each pair of objects in the dataset. The contribution of our previous work [1] is that the spatial context and mutual information are used to improve the accuracy of the image parsing problem. The present work also uses these two features but in a different direction to speed up the calculation of the probability.

III. Incoporating the Mutual-Spatial Feature in the Image Parsing Problem

A. Retrieval Set

In image parsing problem, as mentioned in Section 1, the distribution of object is not uniform, hence we prefers to use nonparametric model that infers a region from the most similar regions in the so-called retrieval set of image which contains K images most similar to the test image. Based on [3] our system also uses four types of global image features: Spatial pyramid, GIST, Tiny image and Color histogram to calculate the distance from each image in the training set to the test image. Then K images corresponding to smallest distances are selected to put into the retrieval set. An informative retrieval set should contain scene images similar to the test image.

B. Contextual Inference

In order to enforce contextual constraints on image parsing problem, MRF model is preferred to the CRF model because the CRF model is very costly in learning and inference. Therefore, to assign label $l = \{l_1, l_2, \ldots, l_j\}$ to the set of regions $r = \{r_1, r_2, \ldots, r_i\}$ the per-class likelihood score of regions and probability of every co-occurrence object in retrieval set are put into the fully connected MRF model[15]. Similar to [1], [2], [3], [4], [5], [6], [7], the image labeling is formulated by minimizing of standard MRF energy function defined based on labels l:

$$J(l) = \sum_{j=1:n;\, i=1:m} \mu(l_j, r_i) + \lambda E_{smooth} \quad (6)$$

where n and m are the number of object labels in the retrieval set and regions in the test image; $\mu(l_j, r_i)$ presents the negative logarithm of per-class likelihood scores for each region r_i; smoothing term E_{smooth} shows the negative logarithm of the probability between co-occurrence object labels in the dataset.

In our system, in order to increase the plausibility of inference, E_{smooth} is defined by (6):

$$E_{smooth} = E_{mutualspatial} + E_{co-occurence} \quad (7)$$

where $E_{mutualspatial}$ energy from (5) contained the information about probability of object co-occurrencelabels including the spatial context and mutual information. $E_{co-occurence}$ is calculated as the negative logarithm of accumulating the number of object co-occurrence in the dataset.

IV. Experiment

For evaluating our proposed method, several experiments on the Barcelona and SIFTFlow datasets [16] are conducted. The Barcelona dataset contains 14871 training images and 279 testing images in 170 labels. The SIFTFlow dataset includes 2488 train images and 200 test images in 33 labels. The proposed method is performed in MATLAB oni5-core 3.5 GHz Intel(R), 8GB RAM environment.

TABLE I. Per-Pixel Accuracy Rate

No.	SIFTFlow data set	
	Recent works	***(%)***
1	Liu [7]	74.75
2	J.Tighe[3]	76.90
3	DEigen [4]	77.10
4	H.Myeong[2]	77.14
5	J.Tighe[5]	77.00
6	J.Tighe[6]	78.60
7	Our previous work[1]	78.19
8	Our experiment	78.20

To evaluate our novel approach and provide strong comparison, we determine per-class recognition rate and compare with baseline MRF models using the SIFTFlow dataset. The incorporating spatial relationship and mutual information in our previous study [5] is replaced by new method with the best smoothing value. The per-pixel accuracy rate (Table I) indicates the effectiveness of using the spatial context and mutual information on image parsing problem. As shown in Table I, our system achieves an overall per-pixels accuracy rate of 78.20% while the baseline per-pixel rate is 77.19% [5]. The work of Tighe *et al.* [6] has showed better result than ours. However, they use Per-Exemplar Detectors feature which takes much more time and memory for training, resulting in an expensive model.

The spatial context and mutual information applied in the new method presented in Section 2 speeds up the computation by four times compared to our previous

©2014 Engineering and Technology Publishing

work in calculating the mutual-spatial probability of co-occurrence objects.

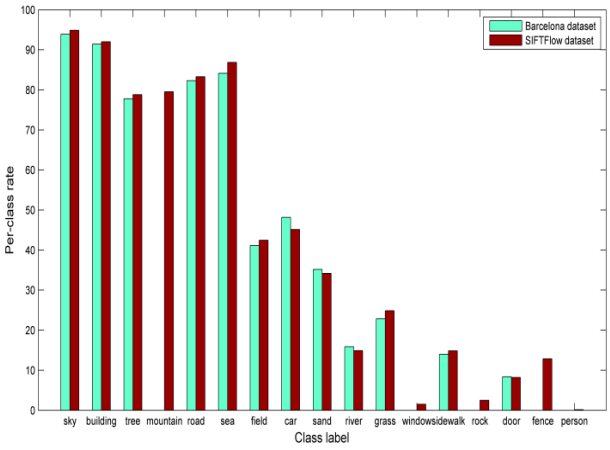

Figure 3. The comparison per-class recognition rate between SIFTFlow and Barcelona dataset. Note that class labels has 0% accuracy are not shown.

Our method can be applied in various dataset because it only requires some simple computations regarding the probability of co-occurrence objects and the mutual-spatial information in pair of objects instead of time-consuming training of the parameters. Therefore, our method can be suitable for various dataset. As shown in Fig. 3, the per-class rate of our model for Barcelona and SIFTFlow dataset are compared with common objects in both datasets.

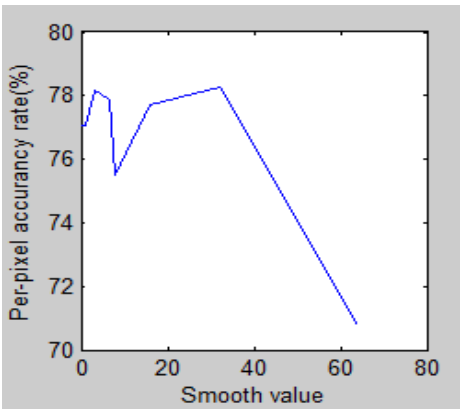

Figure 4. The dependency of smooth value in MRF model. The good smooth values are 3.2 and 32. We also use this for our system and get more accurate for our model.

The effectiveness of smoothing value in (6) with Per-pixel accuracy rate in MRF framework is shown in Fig. 4. In our method, the best smoothing values are 3.2 and 32.

As shown in Fig. 5, our method is more accurate for common objects inreal-time environment with high appearance frequency such as "Street", "Building", "Car", "Tree", "Sky", etc. However, it still has error with some object rarely occurred such as “Sand”. The error happens due to “Sand” has a very low occurrence frequency in the dataset, therefore there is not enough information to infer this object. This is also the limitation of our model, we will gather more information of the object to increase accuracy of per-pixel rate in the future work.

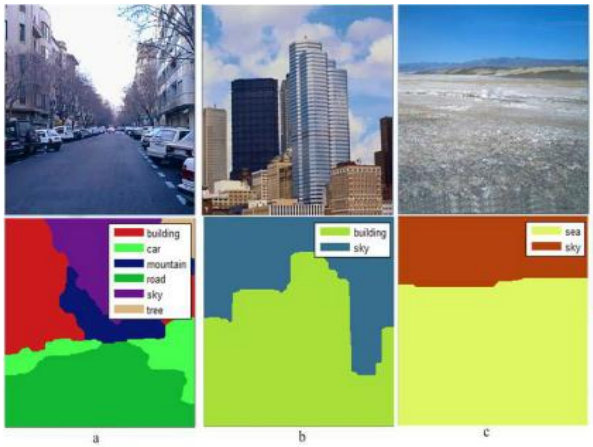

Figure 5. Result images from our method. Figure (a, b) show the accurate result when apply our model. Figure (c) is one wrong case. Sand is confused with Sea object in the sense with Sky object because the frequency of Sand in the dataset is too small.

V. Conclusion

This paper has presented animproved approach to our previous work for image parsing inspired by [3]. The proposed approach does not require time-consuming training except basic computation of some statistics such as label co-occurrence probability. In addition, the accuracy of image parsing is improved by incorporating the spatial context and mutual information into MRF framework.

The spatial context and mutual information that capture the relationship of object location and frequency of co-occurrence objects in an image is clearly a strong visual cue which provides more avenues for improving categorization accuracy. The experimental results showthat the processing time in the new method is faster than our previous work in calculating the mutual-spatial probability.

The key contribution of this work is introducing a new method to simultaneously model the spatial context and mutual information in our previous work [1] in the computation step. The new method not only improves the accuracy of per-pixel rate but also reduces the processing time in calculating the probability of the mutual-spatial feature.

References

[1] T. L. Vu, S. W. Choi, and C. H. Lee, “Improving accuracy for image parsing using spatial context and mutual information,” in *Proc. 20th International Conference of Neural Information Processing*, Daegu, Korea, vol. 8228, 2013, pp. 176-183.

[2] H. Myeong, J. Y. Chang, and K. M. Lee, “Learning object relationships via graph-based context model,” in *Prof. 2012 IEEE Conf. on Computer Vision and Pattern Recognition*, New York, USA, 2012, pp. 2727-2734.

[3] J. Tighe and S. Lazebnik, “Superparsing: Scalable nonparametric image parsing with super-pixels,” in *Proc. 11th European Conf. on Computer Vision*, Berlin, Germany, 2010, pp. 352-365.

[4] D. Eigen and R. Fergus, “Nonparametric image parsing using adaptive neighbor sets,” in *Prof. 2012 IEEE Conf. on Computer Vision and Pattern Recognition*, New York, USA, 2012, pp. 2799-2806.

[5] J. Tighe and S. Lazebnik, “Understanding scenes on many levels,” in *Proc. International Conf. on Computer Vision*, Barcelona, Spain, 2011, pp. 335-342.

©2014 Engineering and Technology Publishing

[6] J. Tighe and S. Lazebnik, "Finding things: Image parsing with regions and per-exemplar detectors," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Portland, OR, 2013, pp. 3001-3008.
[7] J. Tighe and S. Lazebnik, "Superparsing: Scalable nonparametric image parsing with super-pixels," *International Journal of Computer Vision*, pp. 329-349, 2013.
[8] B. C. Russell, A. Torralba, R. Fergus, and W. T. Freeman, "Object recognition by scene alignment," in *Proc. 21st Anual Conf. on Neural Information Processing System*, British Columbia, Canada, 2007.
[9] X. He, R. S. Zemel, and M. A Carreira-Perpinan, "Multiscale conditional random fields for image labeling," in *Proc. 4th IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, Ont. Canada, pp. 695-703.
[10] J. Weeds and D. Weir, "Co-occurrence retrieval: A flexible framework for lexical sistributional similarity," *Journal of Computational Linguistic*, vol. 31, pp. 439-475, December 2005.
[11] S. Kumar and M. Hebert, "A hierarchical field framework for unified context-based classificatio," in *Proc. 10th IEEE International Conf. on Computer Vision-Volume 2*, Washington, DC, USA, pp. 1284-1291.
[12] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient graph-based image segmentation," *Journal of International Journal of Computer Vision*, vol. 59, pp. 167-181, 2004.
[13] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. the IEEE Conf. on Computer Vision and Pattern Recognition*, Anchorage, AK, USA, 2008, pp. 1-8.
[14] W. K. Church and P. Hanks, "Words association norms, mutual information and lexicography," in *Proc. 27th Annual Conf. of the Association for Computational Linguistics*, PA, USA, 1989, pp. 76-83.
[15] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky, "Exploiting hierarchical context on a large database of object categories," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, USA, 2010, pp.129-136.
[16] Dataset for image parsing experiments. [Online]. Available: http://www.cs.unc.edu

Thi LyVu completed her B.S from Le Quy Don Technical University, Vietnam in Department of Information Technology, in 2011. She is currently pursuing the MS degree in the Information Intelligent Processing System Lab. of the Graduate School of Information & Communication Engineering at Inha University, Korea. Her study includes pattern recognition, image processing, and Intelligent processing system. Her most recent publication is: Improving Accuracy for Image Parsing Using Spatial Context and Mutual Information (Daegu, Korea: ICONIP, 2013). She currently studies about image understanding, and image processing.

Chong Ho Lee received his M.S degree in Electrical Engineering from Seoul National University, Korea and Ph.D degree from Iowa State University, USA in Department of Computer Engineering. His research areas are Artificial Neural Networks, Intelligent Systems. He is currently a Professor in School of Information & Communication Engineering at Inha University, Incheon, Korea. He joined in Dynamic Partial Reconfigurable FIR filter design, LNCS 3839, Mar. 2006, Design of CSVM Processor for Intelligence Expression, Journal of Electrical and Electronic Material, feb.2007 and DNA-inspired CVD Diagnostic Hardware Architecture, Journal of KIEE, Feb. 2008. His publications is recently: MARIOBOT marionette robot that interact with an audience (Boston, Massachusetts, USA: ACM, 2012), Interactive display robot: projector robot with natural user interface (Tokyo, Japan: IEEE Press, 2013), Improving Accuracy for Image Parsing Using Spatial Context and Mutual Information (Daegu, Korea: ICONIP, 2013).

©2014 Engineering and Technology Publishing