Image Labeling using Integration of Local and Global Features

Takuto Omiya and Kazuhiro Hotta
Meijo University, 1-501 Shiogamaguchi, Tenpaku-ku, Nagoya 468-8502, Japan

Keywords: Image Labeling, Integration of Local and Global Features, Bag-of-Words, RootSIFT.

Abstract: In this paper, we carry out image labeling based on probabilistic integration of local and global features. Many conventional methods put label to each pixel or region using the features extracted from local regions and local contextual relationships between neighboring regions. However, labeling results tend to depend on a local viewpoint. To overcome this problem, we propose the image labeling method using not only local features but also global features. We compute posterior probability of local and global features independently, and they are integrated by the product. To compute probability of global region (entire image), Bag-of-Words is used. On the other hand, local co-occurrence between color and texture features is used to compute local probability. In the experiments using MSRC21 dataset, labeling accuracy is much improved by using global viewpoint.

1 INTRODUCTION

Image labeling is one of the most challenging and important problems in computer vision. The goal is to assign label in pre-defined set of classes e.g. sky, car, road, etc. to every pixel in the image. Image labeling is one of the most crucial steps toward image understanding and has a variety of applications such as image retrieval and image classification. The most fundamental approach for image labeling put labels to each region (pixel, image patches) using the local features (color, texture, etc.) extracted from each region (Barnard and Forsyth, 2001). However, this approach has some problems in which labels in an object tend to be inconsistent, because this approach puts labels to each region independently and the results tend to depend on local viewpoint.

To overcome these problems, some approaches have been proposed recently. Popular approaches use information not only local features but also local contextual relationships between regions. In those methods, Conditional Random Field (CRF) (Lafferty et al., 2001) model is used. Shotton et al. (2006) used a CRF model to jointly model the appearance, shape and context information of different semantic categories. Gould et al. (2008) used a CRF-based model to integrate the relative location prior of different categories by using appearance-based image features. Tu (2008) proposed an approach for learning a contextual model named auto-context without CRF model. The common problem in these approaches is that recognition results tend to get in a local minimum. We consider that the reason of this problem is lack of global viewpoint. Since only local and local relationship information are used, it puts mislabels to regions which are classified easily by global viewpoint.

Figure 1: Global information helps to recognize the regions which are difficult to recognize from only local information.

In this paper, we propose the image labelling method which introduces the global viewpoint. The effectiveness of global information is shown in Figure 1. It is difficult to recognize images in top row but we can recognize red square regions easily by using entire image. This shows that global information much helps to recognize the regions which are difficult to recognize from only local

Omiya T. and Hotta K. (2013).
Image Labeling using Integration of Local and Global Features.
In *Proceedings of the 2nd International Conference on Pattern Recognition Applications and Methods*, pages 613-618
DOI: 10.5220/0004334606130618

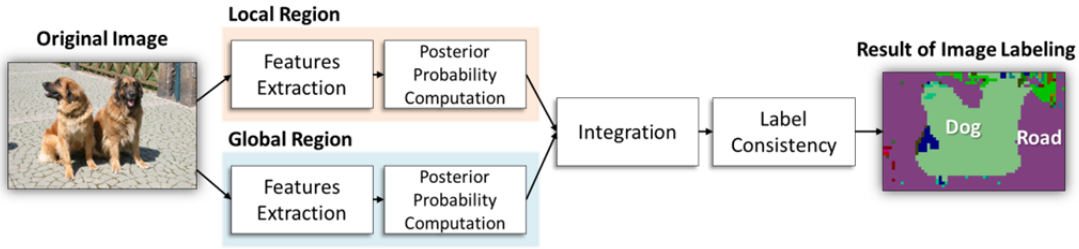

Figure 2: The framework of proposed method.

information. In the proposed method, we estimate class label of each local region in the image, and class label of entire image. Then we integrate the probability for each class label obtained from local and global viewpoints. In local region, we compute the posterior probability of each class label by Support Vector Machine (SVM) (Vapnik, 1995) of local features. We use the local co-occurrence feature between texture and color. In global region, we use Bag-of-Words (BoW) (Csurka et al., 2004) model which is a widely used for image recognition (Nowak et al., 2006). We also use SVM and compute posterior probability of entire image. By taking the product of posterior probability of local and global features, information of local and global are integrated. After that, in order to improve the accuracy, we carry out label consistency process. Concretely, we take the product of the probability between certain region and neighboring regions, and we put the class label with to the region the highest probability.

In the conventional method using the global viewpoint, Galleguillos et al. (2008) used BoW as global features and incorporate into CRF model. The difference from the proposed method is to perform image segmentation before integration of local and global, In addition, by using co-occurrences and relative location of image regions after integration of local and global labels reused. Ladicky et al. (2010) used object class co-occurrence in an image as global viewpoint, and incorporate it into CRF model. It is different from our approach integrating local and global information by product of probabilities.

Experiments are carried out using the MSRC21 dataset (Shotton et al., 2006) with 21 object classes. We confirmed that the accuracy is improved by introducing the global viewpoint. Class average accuracy was 72.5% and pixel-wise accuracy was 76.2%. This is much higher than the accuracy using only local features in which class average accuracy was 48.6% and pixel-wise accuracy was 63.0%.

The remainder of the paper is organized as follows. In section 2, we describe the details of the proposed approach. Section 3 shows experimental results for MSRC21 dataset. Conclude and future works are described in section 4.

2 PROPOSED METHOD

In this section, we describe the proposed method which integrates local and global features. The framework of the proposed method is shown in Figure 2. Our method consists of the extraction of local and global features, computation of the posterior probability of the local and global region, integration of local and global information and label consistency in the neighboring region.

We describe each stage in the following sections.

2.1 Extraction of Local Features

We estimate the class label of each local region based on color and texture features. Color features are effective to identify object classes which have characteristic in color (e.g. sky, grass, etc.). On the other hand, color features are sensitive to changes in brightness. In addition, in some object classes, color variation is large such as red cars and white cars, and color features are not effective. Therefore, we also use local texture features. Texture features are robust to changes in brightness and are not affected by color variation. In this paper, we exploit more robust local feature. That is the local co-occurrence of color and texture features, and is defined by using color and texture histograms.

We use HSV color space to make color histogram because RGB value is sensitive to changes in brightness. In order to create HSV histogram, we quantize HSV color space by discrete intervals; 18 discrete values for hue and 3 values for saturation and value (brightness) which is used in (Smith and Chang, 1996). As a result, HSV histogram becomes 162 (18×3×3) dimensions. Hue is quantized finer than other elements because hue is the most important element to express color.

We use Local Binary Pattern (LBP) (Ojala, 2002) as texture histogram. LBP is also robust to changes in brightness. Since LBP is extracted from grayscale images, it is robust to color changes. LBP is defined as an ordered set of binary comparisons of pixel intensities between the center pixel and its 8 surrounding pixels. The decimal form of LBP code is expressed as

$$\mathrm{LBP}(x_c, y_c) = \sum_{n=0}^{7} s(i_n - i_c)2^n, \quad (1)$$

where i_c corresponds to intensity value of the center pixel at position (x_c, y_c), i_n to the intensity values of the 8 surrounding pixels, and function $s(x)$ is defined as

$$\mathrm{s}(x) = \begin{cases} 1, x \geq 0 \\ 0, x < 0. \end{cases} \quad (2)$$

Each bin in LBP histogram corresponds to the LBP code, and the dimension of LBP histogram is 256.

We describe how to define the local co-occurrence feature. The feature is represented by HSV histogram (162 bins) and LBP histogram (256 bins). Figure 3 shows how to compute the local co-occurrence feature. We extract the values of HSV and LBP at each pixel and vote one to corresponding bin in two dimensional spaces. Thus, this feature can represent local co-occurrence of color and texture features. Final dimension of this feature is 41,472 (256×162).

Here we define local region to make the local co-occurrence histogram. Since the class label is assign to each local region, the size of a local region should not be large. In this paper, the size of a local region is set to 5×5 pixels. However, if we make co-occurrence histogram in 5×5 pixels, it becomes too sparse. Thus, local co-occurrence histogram is made from surrounding 15×15 pixels of the center local region with 5×5 pixels.

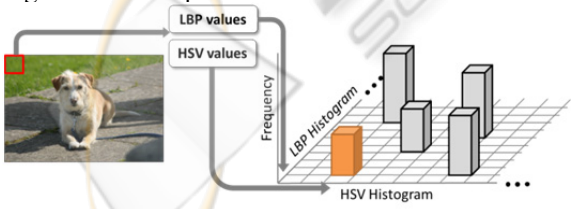

Figure 3: The local co-occurrence histogram of HSV and LBP features.

2.2 Extraction of Global Features

In this section, we describe the global features extracted from the entire image. We use Bag-of-Words (BoW) as global features. SIFT features are extracted from the entire image, and they are divided into several clusters by K-means. The cluster center vectors are used as visual words. The entire image is described by the histogram of frequency of visual words.

We use RootSIFT (Arandjelovic and Zisserman, 2012) instead of standard SIFT (Lowe, 1999) because RootSIFT outperformed standard SIFT. RootSIFT is obtained by the simple transformation of SIFT, and it is an element wise square root of the L1 normalized SIFT vectors.

We extract RootSIFT by grid sampling whose effectiveness is reported in image recognition (Fei-Fei and Perona, 2005). RootSIFT is extracted at the interval of 8 pixels with several scales (8, 12, 16 and 20 pixels). In experiments, the number of visual words is set to 1000 empirically.

2.3 Computation of Posterior Probability

We compute the posterior probability of local and global features. Since image labeling is the multi-class classification problem, we use one-against-one SVM. In this paper, both of local and global features are represented by histograms. Therefore, we use χ^2 kernel and histogram intersection kernel whose effectiveness was been reported (Zhang et al., 2007); (Chapelle et al., 1999). In global features, we use χ^2 kernel is defined as

$$K_{\chi^2}(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}), \quad (3)$$

where γ is the hyper-parameter which is determined on the basis of cross-validation. χ^2 kernel gives high accuracy but its computational cost is high and parameter dependence is high.

In local features, the number of dimensions is high and the number of local features is large. Therefore, we use histogram intersection kernel whose computational cost is lower than χ^2 kernel. In addition, the accuracy is comparable to χ^2 kernel. Histogram intersection kernel is defined as

$$K_{hi}(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \min(x_i, y_i). \quad (4)$$

We compute the posterior probability of local and global regions independently. To compute the posterior probability, LibSVM (Chang and Lin, 2001) is used. The probability of global region is represented as

$$p_{Global_i} = \mathrm{P}(C_i | \boldsymbol{x}_{Global}), \quad (5)$$

where C_i corresponds to the ith object class and $\boldsymbol{x}_{Global}$ is global feature vector.

The probability of local region j in an image is represented as

$$p_{Local_i}^{\ j} = \mathrm{P}(C_i|\boldsymbol{x}_{Local_j}), \quad (6)$$

where j corresponds to the jth local region in an image. $\boldsymbol{x}_{Local_j}$ is local feature vector of the local region j.

2.4 Integration of Local and Global Information

After computing posterior probabilities of local and global features, we integrate them the product of the probability as

$$p_{Integration_i}^{\ j} = p_{Local_i}^{\ j}\, p_{Global_i}. \quad (7)$$

$p_{Integration_i}^{\ j}$ expresses the probability of the ith class for local region j in an image.

2.5 Label Consistency

There is correlation between each region and its neighboring regions. Thus, we use label consistency to improve accuracy after integrating of local and global information. The region N for label consistency is defined by certain local region and its 8 neighboring local regions. Label consistency is obtained by the product of $p_{Integration_i}^{\ j}$ between the center and its 8 neighboring regions. The label l in the local region j is defined as

$$l_j = \underset{i}{\mathrm{argmax}} \prod_{j \in N} p_{Integration_i}^{\ j}. \quad (8)$$

This process helps to put labels more smoothness.

3 EXPERIMENTS

This section shows the experimental results. First, the image dataset used in experiment and evaluation method are explained in section 3.1. Next, evaluation results are shown in section 3.2.

3.1 Image Dataset and Evaluation Method

We evaluate the proposed method using the MSRC21 dataset (Shotton et al., 2006). This dataset consists of 591 color images whose size is approximately 320×213 pixels. The ground truth (correct labeling) with 21 object classes are given. The dataset is already divided into training and test set (276 training and 256 test images), and we also use them.

Image labeling performance is evaluated by two accuracies in conventional methods (Shotton et al., 2006). The first one is class average accuracy which is the average of accuracy of each class. The second one is the pixel-wise accuracy which is the accuracy rate in terms of all pixels. We also evaluated our method by two measures.

3.2 Evaluation Results

First, we evaluate all different steps in the proposed method as shown in the top 3 rows of Table 1. When only local features are used in object classes which are small change in appearance such as sky and grass, high accuracy is obtained. On the other hand, accuracies of object classes with large change in appearance are low.

However, by introducing global features, the accuracies of all classes except for building and tree are much improved. Class average accuracy is improved 15.6% and pixel-wise accuracy is improved 8.5%. In particular, accuracies of object classes with large change in appearance such as chair, boat and sign are greatly improved by integration global information. The result demonstrates the effectiveness global viewpoint.

After that, we add label consistency process to our method. The process improves the accuracy of most classes except for grass, road and sky. Class average accuracy is improved 8.2% and pixel-wise accuracy is improved 4.7%. These results demonstrate that each step of our method is effective.

Next we compare the results to conventional methods (Tu, 2008); (Galleguillos et al., 2008); (Ladicky et al., 2010) as shown in the bottom 3 rows of Table 1. Our method is getting close to auto-context model (Tu, 2008). Pixel-wise accuracy is worse slightly but class average accuracy of our method is better than auto-context. Tu (2008) introduced the auto-context model to use contextual information. While they did not use the global viewpoint. In particular, accuracies of object classes (flower, bird and boat) of the proposed method are higher than auto-context model. By using global features and label consistency, our method improved the accuracy of the object classes with large changes in the appearance.

Galleguillos et al. (2008) used global viewpoint in CRF. In addition, object co-occurrence and spatial

Table 1: Results on the MSRC21 dataset. For each class, the pixel-wise accuracy is provided. Average represents class average accuracy. Pixel-wise represents pixel-wise accuracy. The top three rows show different steps of proposed method. (a) shows results when only local features. (b) shows results when integration local and global features. (c) shows when integration local and global features and label consistency. The bottom three rows show results of conventional methods. (d) shows results of Tu (2008). (e) shows results of Galleguillos et al. (2008). (f) shows results of Ladicky et al. (2010).

	building	grass	tree	cow	Sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	Average	Pixel-wise
(a)	52	92	71	50	53	90	56	56	58	43	65	46	18	13	61	12	70	45	26	31	11	48.6	63.0
(b)	45	94	70	74	75	91	77	63	75	65	89	63	49	34	72	45	79	64	50	40	37	64.3	71.5
(c)	55	92	74	87	85	89	88	63	80	78	92	76	65	47	78	62	77	70	65	52	47	72.5	76.2
(d)	69	96	87	78	80	95	83	67	84	70	79	47	61	30	80	45	78	68	52	67	27	68.7	77.7
(e)	91	95	80	41	55	97	73	95	81	57	60	65	54	52	56	42	96	42	46	77	81	68.4	n/a
(f)	82	95	88	73	88	100	83	92	88	87	88	96	96	27	85	37	93	49	80	65	20	76.8	87

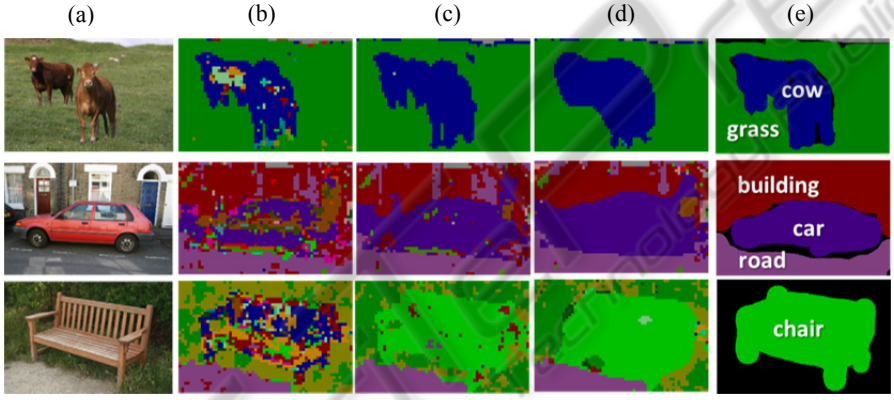

Figure 4: Example results on the MSRC21 dataset. (a) Original images, (b) Results of local features, (c) Results of local + global features, (d) Results of local + global features + Label consistency, (e) Ground truth.

context (above, below, inside and around) are also introduced. Our method outperformed the method in class average accuracy. More that pixel-wise accuracy of the method is not published. Since Galleguillos et al. introduced the spatial context, the accuracies of object classes with strong characteristic spatially such as sky, road and water are higher than our method.

Ladicky et al. (2010) incorporated object co-occurrence statistics as global viewpoint into CRF. Object co-occurrence statistics provide which classes appear in the same image together. The accuracy of this method is high. However, in chair, boat, cat and bird classes, our method is much higher than it. To recognize the object classes with large change in the appearance, global viewpoint of our method is more superior.

Qualitative results of the proposed method are shown in Figure 4. When only local features are used, class labels are scattered. The label of each region has been identified independently, labeling results fall into a local minimum. By introducing global features, the dispersion of the label is improved. In addition, global features provide improvement to recognize object classes which can not be recognized by only local features. In addition, label consistency process provides the more consistent labeling results. These results show the effectiveness of our method.

4 CONCLUSIONS

In this paper, we proposed image labeling method that integrates the labels obtained from local and global viewpoints. We demonstrated the effectiveness of using global viewpoint by experiments. Only local viewpoint can not recognize

objects with complex structure, and labeling results fall into a local minimum. Experimental results show that global viewpoint overcome these problems. In addition, label consistency process provides more smooth labeling results.

The main problem of the proposed method is that global feature can not handle multiple classes and represent the position of the objects. This is because Bag-of-Words method classifies only one object in an image. For example, when the global features are extracted from the image contained car and building, we obtain probability of each class not both classes. Current global feature can not recognize building and car simultaneously, and position of each object is not obtained. We want to introduce the new global feature which can recognize multiple classes and position of the objects. That is a subject for future works.

ACKNOWLEDGEMENTS

This work was supported by KAKENHI No. 24700178.

REFERENCES

Arandjelovic, R. Zisserman, A. (2012). Three things everyone should know to improve object retrieval. *Proc. Computer Vision and Pattern Recognition, pp. 2911-2918.*

Barnard, K. and Forsyth, D. (2001). Learning the semantics of words and pictures. *Proc. International Conference on Computer Vision. vol. 2, pp. 408-415.*

Chang, C. and Lin, J. (2001). LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chapelle, O., Haffner, P. and Vapnik, V. (1999). Support vector machines for histogram-based image classification. *Neural Networks, vol. 10, pp. 1055-1064.*

Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C. (2004). Visual categorization with bags of keypoints. *Proc. ECCV Workshop on Statistical Learning in Computer Vision, pp. 59-74.*

Fei-Fei, L. and Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. *Proc. Computer Vision and Pattern Recognition, vol. 2, pp. 524-531.*

Galleguillos, C., Rabinovich, A. and Belongie, S. (2008). Object categorization using co-occurrence, location and appearance. *Proc. Computer Vision and Pattern Recognition, pp. 1-8.*

Gould, S., Rodgers, J., Cohen, D., Elidan, G. and Koller, D. (2008). Multi-class segmentation with relative location prior. *International Journal of Computer Vision, vol. 80, pp. 300-316.*

Ladicky, L., Russell, C., Kohli, P. and Torr, P. (2010). Graph cut based inference with co-occurrence statistics. *Proc. European Conference on Computer Vision. pp. 239-253.*

Lafferty, J., McCallum, A. and Pereira, F. (2001). Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proc. International Conference on Machine Learning, pp. 282-289.*

Lowe, D. (1999). Object recognition from local scale-ivariant features. *Proc. International Conference on Computer Vision, vol. 2, pp. 1150-1157.*

Nowak, E., Jurie, F. and Triggs, B. (2006). Sampling strategies for bag-of-features image classification. *Proc. European Conference on Computer Vision, pp. 490-503.*

Ojala, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, vol 24, pp. 971-987.*

Shotton, J., Winn, J., Rother, C. and Criminisi, A. (2006). Textonboost: joint appearance, shape and context modelling for multi-class object recognition and segmentation. *Proc. European Conference on Computer Vision, pp. 1-15.*

Smith, J. and Chang, S. (1996). Tools and techniques for color image retrieval. *Symposium on Electronic Imaging: Science and Technoloy-Strage and Retrieval for Image and Video Database IV, pp. 426-437.*

Tu, Zhuowen. (2008). Auto-context and its application to high-level vision tasks. *Proc Computer Vision and Pattern Recognition, pp. 1-8.*

Vapnik, V. (1995). *The nature of statistical learning theory*, Springer-verlag New York. New York.

Zhang, J., Marzaklek, M., Lazebnik, S. and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision, vol. 73, pp. 213-238.*