Estimation of Object Position Based on Color and Shape Contextual Information

Takashi Ishihara, Kazuhiro Hotta, and Haruhisa Takahashi

The University of Electro-Communications,
1-5-1 Chofugaoka Chofu-shi Tokyo, 182-8585, Japan
{ishihara,hotta,takahasi}@ice.uec.ac.jp
http://www.htlab.ice.uec.ac.jp/

Abstract. This paper presents a method to estimate the position of object using contextual information. Although convention methods used only shape contextual information, color contextual information is also effective to describe scenes. Thus we use both shape and color contextual information. To estimate the object position from only contextual information, the Support Vector Regression is used. We choose the Pyramid Match Kernel which measures the similarity between histograms because our contextual information is described as histogram. When one kernel is applied to a feature vector which consists of color and shape, the similarity of each feature is not used effectively. Thus, kernels are applied to color and shape independently, and the weighted sum of the outputs of both kernels is used. We confirm that the proposed method outperforms conventional methods.

Keywords: color context, shape context, object detection, support vector regression, Pyramid Match Kernel, summation kernel.

1 Introduction

It becomes increasingly important to detect specific target from still images or videos as the first step of object tracking and recognition. Although conventional methods detected a target by clipping part of an image, the scenery around and objects it belongs exists a strong relationship with each other. Human detects the object using contextual information obtained from the image as well as information obtained from the object.

Recently, the methods using contextual information to estimate the position of object were proposed [1][2]. Conventional methods were based on only shape contextual information, though the color information is effective to describe scenes. Therefore, we propose the contextual information based on both color and shape information. Color information is robust to background changes and less computational cost in comparison to the shape information.

As a contextual information based on color information, we use color histograms which are computed in subregions of various sizes. We also use Gabor feature which was used in convention methods [1][2] to extract shape information. After extracting a Gabor feature from an image, we construct histograms

P. Foggia, C. Sansone, and M. Vento (Eds.): ICIAP 2009, LNCS 5716, pp. 825–834, 2009.

which are computed in subregions of various sizes, and they are used as shape contexutal information.

Torralba [1] used generative model to estimate the position of object from contextual information. Suzuyama et al. [2] has shown effective use of Support Vector Regression (SVR) instead of using generative model to estimate the position from contextual information, and achieved the high accuracy in comparison to Torralba's method. From these reasons, we select SVR. The generalization ability of SVR depends on the selection of a kernel function. We choose the Pyramid Match Kernel [3] based on the similarity between histograms, because our contextual information is based on histogram.

If we apply one kernel to a feature vector which consists of color and shape contextual information, then it is hard to reflect the similarity of only shape or color. To avoid this problem, we apply a kernel to each color and shape information independently, and the outputs of kernels are integrated by summation [4][5].

In the experiments, we estimate the position of car in an image from only contextual information. To compare the accuracy with the conventional method [2], we use the same image database as the conventional method. Experimental results show that the proposed method which integrates color and shape information outperforms the conventional method based on only shape contextual information. To confirm the effectiveness of the use of Pyramid Match Kernel, we compare the Pyramid Match Kernel and polynomial kernel which was used in conventional method [2]. It turns out that Pyramid Match Kernel achieves the higher accuracy in comparison to polynomial kernel.

This paper is organized as follows. In section 2, we describe color and shape contextual information. Section 3 explains how to estimate the position of object from contextual information. Experimental results are shown in section 4. Finally, conclusion and future works are described in section 5.

2 Contextual Information

To extract shape contextual information, we use Gabor feature which was also used in conventional methods [1][2]. Although conventional methods applied principal component analysis to Gabor feature, the shape contextual information depends heavily on the position of objects. The positions of objects are not stable in the same scene. Thus, the proposed method develops histograms from Gabor feature to be robust to position changes of objects. They are used as shape contextual information. We also use color histogram as color contextual information. This is also robust to position changes of objects. The details of shape contextual information is explained in section 2.1. Section 2.2 explains color contextual information.

2.1 Shape Contextal Information

In this section, we explain how to extract the shape contextual information. First, we explain Gabor feature.

The Gabor filters [8][9] used in this paper are defined as

$$v_{\mathbf{k}}(\mathbf{x}) = \frac{\mathbf{k}^2}{\sigma}\exp(\frac{-\mathbf{k}^2\mathbf{x}^2}{2\sigma^2})\cdot(\exp(i\mathbf{kx}) - \exp(-\frac{\sigma^2}{2})), \tag{1}$$

where $\mathbf{x} = (x, y)^T, \mathbf{k} = k_\nu \exp(i\phi), k_v = k_{max}/f^\nu, \phi = \mu \cdot \pi/4, f = \sqrt{2}, \sigma = \pi$. In the following experiments, Gabor filters of 4 different orientations $\mu = (0, 1, 2, 3)$ with 3 frequency levels ($\nu = 0, 1, 2$) are used. Thus, we obtain 12 output images of Gabor filters from one image. The size of Gabor filters of 3 different frequency levels is set to 9 9, 13 13 and 17 17 pixels respectively.

After extracting Gabor feature, we take the average within unoverlapped local regions of 2×2 pixels to reduce the computational cost. Fig.1, 2 and 3 show the outputs of Gabor filters (the norm of real and imaginary parts) of each scale parameter. Red pixels represent high output and blue pixels represent low output. It turns out that Gabor filters of specific orientation emphasize the specific edges. Since Gabor filter of small ν is sensitive to fine edges, the outputs of Gabor filters in Fig.1 is more clear than those in Fig.3.

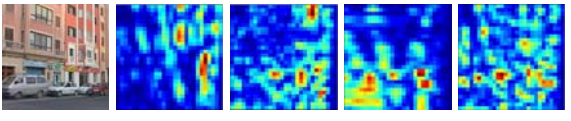

Fig. 1. Example of Gabor feature of ν=0

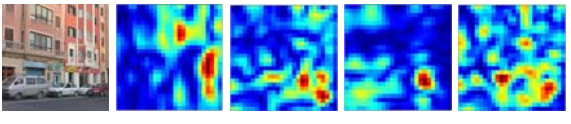

Fig. 2. Example of Gabor feature of ν=1

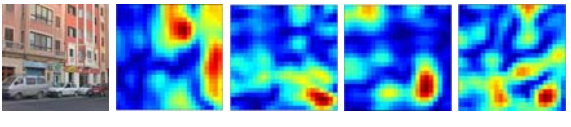

Fig. 3. Example of Gabor feature of ν=2

The output images of Gabor filters depend heavily on the position of objects in images. Namely, they are not robust to position changes of objects. Since the composition and the position of objects in the same scenes are not stable, the robustness to position of objects is required. Thus we develop the histogram from the outputs of Gabor filters. However, we do not know appropriate sizes for computing histograms. If we develop the histogram from whole output image, it is robust to position changes of objects but does not work well to estimate the position. We can consider that both the robustness to position changes and rough topological information are effective. Therefore, we prepare some subregions for

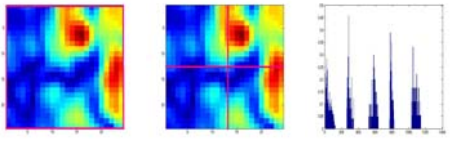

Fig. 4. Subregions for shape histogram and example of histogram

compute histograms. Fig.4 shows the subregions used in the experiments. We divide the output images of Gabor filters into 1×1 and 2×2 subregions, and compute one histogram from one subregion. The histograms are robust feature to position changes. In addition, they have rough topological information. The number of bins of one histogram is set to 256 because 256 bins are appropriate for Pyramid Match Kernel. Each histogram with 256 bins is normalized so that sum of value is 1. The image on rightside in Fig.4 shows the example of shape histogram. The dimention of multi-resolution histogram obtained from an input image is 15,360 (=256(bins)×3(scales)×4(orientations)×5(subregions)).

2.2 Color Contextual Information

The color information is also important to describe scenes as well as shape information. We convert images from RGB color space to HSV color space and histogram is computed in each color independently. To use the local and global feature, we develop the color histogram from subregions of various sizes. In this paper, the image is divided into 1×1, 2×2 and 4×4 subregions, and histogram with 256 bins of each color is developed from each subregion. Each histogram is normalized so that the sum of values is 1. Fig.5 shows examples of the subregions and color histograms. We define color histograms as color contextual information.The dimention of multi-resolution color histogram obtained from an input image is 16,128 (=256(bins)×3(H, S, V)×21(subregions)).

Fig. 5. Subregions for color histogram and example of histogram

3 Position Estimation Using SVR

We select SVR to estimate the position of object, because Suzuyama [2] reported that SVR outperforms generative model [1]. SVR must be trained to estimate the position of objects from only contextual information. To develop the estimator by SVR, the teacher signal (correct position of objects) obtained manually and contextual information are used.

We want to develop the good estimator with high generalization ability. However the generalization ability of SVR depends on the selection of a kernel function. We choose the Pyramid Match Kernel [3] based on the similarity between histograms, because our contextual information is based on histogram. When the Pyramid Match Kernel is adopted to a feature which consists of color and shape contextual information, the similarity of each color or shape is not used effectively. To use the similarity of each feature, we use the weighted summation kernel. First, we explain the Pyramid Match Kernel in section 3.1. Section 3.2 explains the weighted summation kernel.

3.1 Pyramid Match Kernel

Pyramid Match Kernel [3] measures the similarity between histograms, by changing the bin size in a hierarchically fashion. Kernel function is defined as

$$K(y,z) = \sum_{i=0}^{L} \frac{1}{2^i}(L(H_i(y), H_i(z)) - L(H_{i-1}(y), H_{i-1}(z)), \tag{2}$$

where H_i is histogram at level i. $L(H(y), H(z))$ is the function for measuring the similarity between histograms, and counts the overlapped value in the correspoinding bin. In the experiments, L is set to 3.

3.2 Weighted Summation Kernel

We can adopt one kernel to a feature vector which consists of color and shape contextual information. However, in that case, the similarity of each feature is not used effectively. To avoid this problem, we use the summation kernel [4][5]. By assigning kernels to each feature independently and calculating the summation of each output, we obtain a new kernel function which uses the similarity of each feature effectively. The summation function K_{sum} is defined as

$$K_{sum}(x,z) = K_c(x_c, z_c) + K_s(x_s, z_s), \tag{3}$$

where K_c and K_s are the kernels for color and shape. In this paper, Pyramid Match Kernel is used as kernel function. In [5], all kernels are integrated with equal weights. However, the summation kernel with non-negative weight also satisfies Mercer's theorem [6]. In this paper, we use weighted sum of the two kernels as

$$K_{sum}(x,z) = \alpha K_c(x_c, z_c) + (1-\alpha)K_s(x_s, z_s), \tag{4}$$

where α is a constant between 0 and 1. In the following experiments, we set α to 0.5 empirically.

4 Experiments

In the experiments, we estimate the position of car in an image. 990 images are used for training and 330 images are used for evaluation. The size of an input

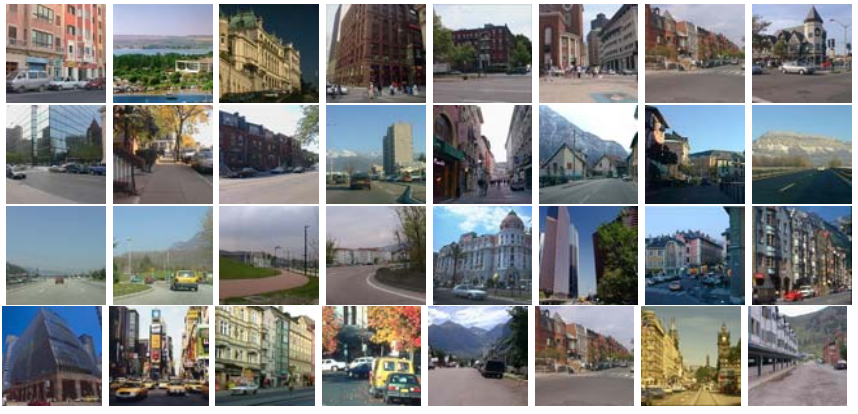

Fig. 6. Examples of images used in test

image for extracting color contextual information is 128 128 pixels. This is the same images as the conventional method [2]. The size of an input image for extracting shape contextual information is 64 64 pixels to reduce the computational cost. Fig.6 shows the examples of images. A variety of scenes can be seen, and cars appear in different poses, scales and positions. The estimation of car position from those images is not easy task.

First, we explain how to train SVR. To develop the estimator of car location by SVR, we need the teacher signal of car location and contextual information obtained from images. The same teacher signals as Suzuyama's method are used. They were obtained manually by positioned the center point of cars in images. Suzuyama had labeled the location of car on the same manner so as not to make a large difference between training and test. SVR is trained by the teacher signals and contextual information of training images. In general, the output of SVR is only one. Thus, we train two SVRs for X and Y coordinates. As previously mentioned, we use weighted summation of Pyramid Match Kernel of color and shape contextual information.

Table 1. Comparison of the proposed method and convention methods

	X-coordinate	Y-coordinate
Proposed method with Pyramid Match Kernel	22.08(pixel)	7.54(pixel)
Suzuyama [2]	22.23(pixel)	8.68(pixel)
Generative model	26.61(pixel)	11.04(pixel)

Table 2. Evaluation of X-Y root mean squared error while changing α

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
X-coordinate	24.42	24.03	23.14	22.61	22.29	22.08	21.95	21.88	21.78	21.67	21.64
Y-coordinate	8.51	7.70	7.57	7.54	7.54	7.54	7.70	7.64	7.71	7.81	8.09

Fig. 7. Examples of in which the proposed method works well

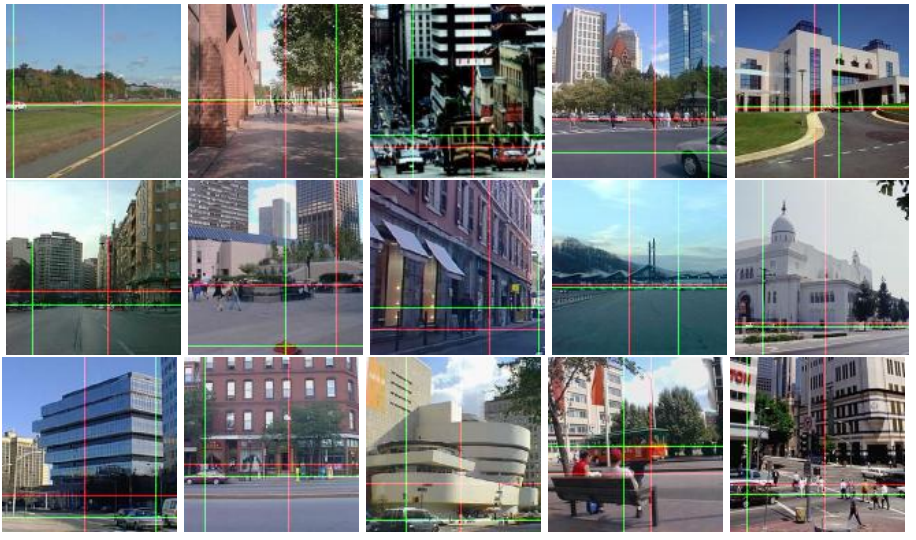

Fig. 8. Examples of failure in prediction

The proposed method estimates the center position (X and Y coordinates) of cars in test images by using SVR. Since we have the correct position of cars in test images, we can calculate root mean squared error between estimated position and correct position. That is used as a measure of accuracy. The conventional method [2] also evaluates the accuracy with the same manner. Table 1 shows result of the root mean squared error of both methods. Since the cars in images line up side-by-side, it is difficult to estimate the position in X-coordinate. On the other hand, the estimation in Y-coordinate is easier than that in X-coordinate because it is rare that cars line up vertically. Thus, the error in X-coordinate is larger than those in Y-coordinate. Table 1 demonstrates that the proposed method outperforms the conventional method. To compare the accuracy with the another method [1], Suzuyama developed the estimator based on generative model of shape contextual information [2]. The result is also shown in Table 1. Our method also outperforms the approach using generative model.

Fig.7 shows the examples in which the proposed method works well. Cross point of green lines on the images indicates the correct position selected manually. The correct position is set to the center of group of cars because cars line up side-by-side. Cross point of red lines on the images indicates the car location estimated by the proposed method. The proposed method works well when it is easy for human to recognize the car location in color. This result suggests the importance of color contextual information as well as shape information. Fig.7 demonstrates that our method works well, though the images used in test include quite different scene such as weather changes, brightness, point of view, graphic resolution and cars appearance, poses, scales and positions.

Next, we show the examples of failure in prediction in Fig.8. The SVR estimates the position by the weighted sum of kernels. Since the kernel computes

the similarity with training samples, the proposed method estimates the position based on the similarities with contextual information of training images. Thus, it does not work well to the images which are not included in training samples.

Table 2 demonstrates the evaluation result of X-Y root mean squared error between estimated position and correct position while changing the weight α in equation (4). $\alpha = 0$ means that only shape contextual information is used, and $\alpha = 1$ means that only color contextual information is used. Estimation of Y-coordinate works well when $\alpha = 0.5$ or 0.4, which suggests the importance of using both color and shape contextual information to estimate the position of Y-coordinate. On the other hand, estimation of X-coordinate works well when $\alpha = 1.0$ which suggests the importance of using color contextual information to estimate the position of X-coordinate.

5 Conclusion

We proposed the method to estimate the position of object based on only color and shape contextual information obtained from the images. In conventional methods, only shape contextual information was used to estimate object position but color information is effective to describe scenes. Thus, we use color and shape contextual information. Since they are described as histograms, we use Pyramid Match Kernel to reflect the similarity of histograms effectively. The results show that the proposed method outperformed the conventional method using only shape contextual information.

Since the proposed method estimates the rough position of objects, it is useful to speed up the object detection. However, current method estimates only one position. We will extend our method to estimate all positions of objects in an image. This is a subject for future works.

References

1. Torralba, A.: Contextual Priming for Object Detection. International Journal of computer Vision 53(2), 169–191 (2003)
2. Suzuyama, Y., Hotta, K., Haruhisa, T.: Context Based Prior Probability Estimation of Object Appearance. The Institute of Electrical Engineers of Japan 129-C(5) (2009) (in Japanese)
3. Grauman, K., Darrell, T.: The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. In: Proc. International Conference on Computer Vision, pp. 1458–1465 (2005)
4. Nomoto, A., Hotta, K., Takahashi, H.: An asbestos counting method from microscope images of building materials using summation kernel of color and shape. In: Proc. International Conference on Neural Information Processing (2008)
5. Hotta, K.: Robust Face Recognition under Partial Occlusion Based on Support Vector Machine with Local Gaussian Summation Kernel. Image and Vision Computing 26(11), 1490–1498 (2008)

6. Shawe-Taylor, J., Cristianini, N.: Kernel methods for Pattern Analysis. Cambridge University Press, Cambridge (2004)
7. Label Me, http://labelme.csail.mit.edu/
8. Hotta, K.: Object Categorization Based on Kernel Principal Component Analysis of Visual Words. In: Proc. IEEE Workshop on Applications of Computer Vision (2008)
9. Lades, M., Vorbruggen, J.C., Buhmann, J., Lange, J., Van de Malsburg, C.: Distortion invariant object recognition in the dynamic link architecture. IEEE Transactions on Computer 42(3), 300–311 (1993)