Image Based Search System Using Hierarchical Object Category Recognition Technique

Takuya Minagawa
Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi,
Kanagawa-ken, 223-8522, Japan
takuya@hvrl.ics.keio.ac.jp

Hideo Saito
Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi,
Kanagawa-ken, 223-8522, Japan
saito@hvrl.ics.keio.ac.jp

Abstract

In this paper, we present the information search system using object category recognition, which is queried by image from mobile phone camera or from photo sharing service on internet. In such system, processing speed is an important requirement. We adopted "Standard Model" proposed by T. Serre in 2005, and improved processing speed by replacing Gabor filter to Haar wavelet, vector quantization of feature patch, and restriction of calculation area. In addition, by retaining the information of each feature's position, it compensates the accuracy which is a little reduced in exchange of processing speed. We implemented this method to server system, and proved this system can work in practical processing time. Through the experiment for Caltech-101 image database, we confirmed value of this system.

1. Introduction

Recently image based search system has been developed, which uses images for query instead of text. This means that user can obtain information from what he or she sees.

A lot of computer vision technologies are applied to this purpose. For instance "similar image matching" which uses color, composition, texture as image features are used to retrieve images[15], because these features have something to do with the impression of images. Word-image translation model is also popular method of this application, which binds segmented image area to word[1][17], therefore user can get words from images and images from words. Local feature point matching approach, e.g.[13], is used to retrieve images which includes same object like logos, magazines, CD jacket, etc, through internet[8][9]. Face recognition/certification technique is used to retrieve image of same person as well[16], and optical character recognition is mainly used to add index, or tags, to image database.[4]

Our main interest is object category recognition technique for information search system. At internet services like photo sharing service, social network service, etc, this technology may help users to find people who uploaded the same category of image, or helps service provider to make profit by advertisement which fit to image content.

Object category recognition is hot topic these years and many types of method are proposed: "bags of features" is one of the most popular approach which compare the histogram of number of each feature[3][20], and "constellation model" is also major which uses Bayes model to train feature, position, and scale[5][6][7].

Hierarchical object recognition method is effective approach as well[14][18], which is inspired by knowledge of bioinformatics. This architecture is simple and possible to be expanded to understand not only object but context of scene[2]. Thus, we adopt this hierarchical architecture proposed by Serre et al, called "standard model of visual cortex"[18], and improve it mostly at processing speed for the purpose of search system.

2. Standard Model

"Standard model of visual cortex" is illustrated in Figure.1. This model has hierarchical 4 layers, in which S layer, representing selectivity, and C layer, representing invariance, appears alternately.

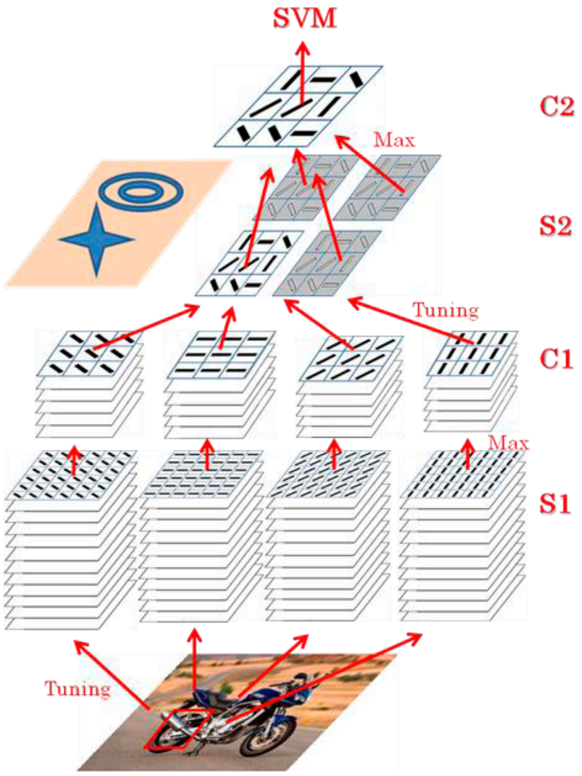


Figure 1. Standard Model

2.1. S1 unit

Input image is transferred to gray scale, and entered into S1 layer. Each unit on S1 reacts to a certain orientation and width of line feature. This reaction is represented as following function.

$$F(x,y) = \exp\left(-(x_0^2+\gamma^2 y_0^2)/2\sigma^2\right)\times\cos\left(2\pi x_0/\lambda\right)$$

$$x_0 = x\cos\theta + y\sin\theta, \qquad y_0 = -x\sin\theta + y\cos\theta \qquad (1)$$

This is 2D Gabor filter of aspect ratio γ, orientation θ, wavelength λ, effective width σ. In this layer, 16 scale bands and 4 orientations (θ=0, π/4, π/2, 3π/4) are used, thus 64 values are generated at each unit.

2.2. C1 unit

C1 layer add the invariance of scale and position to signals from S1 units. Each unit on C1 layer receives outputs of S1 units from nearest Ns×Ns area and 2 scale bands, then passes maximum one as output of this unit. This max operation puts small differences of each signal together, about position and size, at each orientation of line feature.

If larger wavelength λ of Gabor filter is, then larger Ns is on each scale band. S1 areas which enters into adjacent C1 unit are overlapped ΔS each other.

C1 finally generates the output of 8 scale bands and 4 orientations, which has the smaller number of units than S1 output.

2.3. S2 unit

S2 unit reacts to a specific feature patch which has been studied by unsupervised manner. S2 layer has N feature patches (N=1000), which are represented same as C1 format.

Each S2 unit receives signals from C1 layer in n×n area, and outputs responses which are calculated from distance between input signal and each patch by radial basis function (RBF).

$$r = \exp(-\beta\|\mathbf{X}-\mathbf{P}_i\|^2) \qquad (2)$$

X is the input vector from C1 unit, P_i is i_{th} feature patch, and β is sharpness of reaction. Each patch has been learned as following way.

1. C1 signals are calculated from training image.
2. One S2 unit is selected randomly.
3. Inputs of selected S2 units from C1 layer are saved as feature patch.
4. Repeat 1-3 for N times.

Thus, each patch is a vector which has n×n×4 elements. These N patches are used over all 8 scale bands.

2.4. C2 unit

C2 unit integrates S2 outputs of all position and scale by taking maximum signal of them. Therefore, response of C2 is a vector of N elements, in which each element represents a maximum response of each C1 feature patch over all scales and positions.

Finally, this vector is used for training and discrimination of machine learning algorithm in order to recognize object category; we use linear support vector machine at this system.

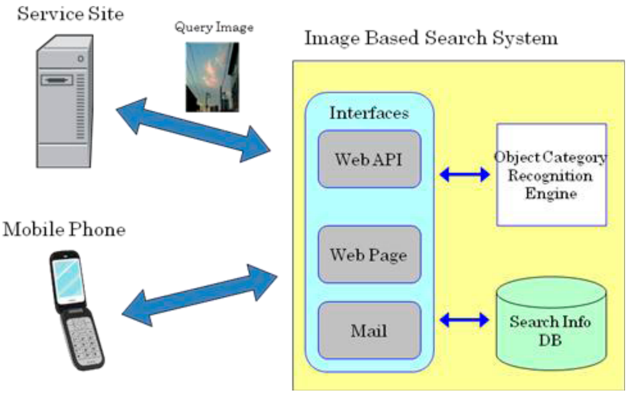


Figure 2. Search System Overview

Table 1. Processing Speed

Process	Standard model[18] (sec)	Our approach (sec)
S1	1.23 ± 0.02	0.31 ± 0.09
C1	0.24 ± 0.01	0.27 ± 0.05
S2	11.04 ± 0.07	0.47 ± 0.09
C2	0.05 ± 0.01	0.05 ± 0.01
SVM	0.00 ± 0.00	0.00 ± 0.01
Total	12.56 ± 0.08	1.11 ± 0.11

3. Our System

3.1. System Architecture

System flow is shown in Fig.3. At first, end user send query image to the system via web or mail interface. Then system forwards it to object category recognition engine, and obtain the category information of image. Finally, this system search information from database or internet by its category, and returns it to user.

We have implemented several two-class classifiers, not multi-class, in order to show end user all candidates, which is better flexibility as search system. In addition, it is not cost effective to train all images in the case that one more class is added to the system of multi-class.

Considering the purpose of this system, its response time should be in a few seconds.

3.2. Recognition Algorithm

We measure processing speed of standard model at the following environment:

- CPU: Intel Core Duo 1.8GHz
- RAM: 2G Bytes
- Image Size: QVGA(240×320)
- Programming Language: C/C++ with OpenCV

As shown in table.1, standard model is very time consuming, especially at S1 and S2. It takes about 12 sec, which is not acceptable for search system. Therefore, we improved it at 4 points as follows:

1. For decreasing the number of feature patches in S2, summarize features by vector quantization.
2. For restricting the area of calculation of S2, inhibit signals of S2 other than the point that C1 layer

takes local maxima and minima.

3. For simplify the process of S1, replace Gabor filter of by Haar wavelet feature.
4. For improving accuracy, retain the position information of each feature of C2.

"1" and "2" of the above is reducing the computation time of S2, and "3" reduces the time of S1 calculation.

3.2.1. Reduction of feature patches

As described in section 2.3, all distances between input signals and trained feature patches are calculated on S2 layer. Accordingly, the number of patches has much influence on processing time of S2. These feature patches are obtained by "imprinting" way; C1 signals from random area of training image are just saved directly. Therefore, these patches might include several similar features, which are redundant for recognition. To avoid this redundancy, we try to cluster feature patches using Linde-Buzo-Gray (LBG) algorithm[12], a major vector quantization approach, and integrate similar ones to one representative. In our system, they are eliminated from N=1000 to N=200.

3.2.2. Restriction of S2 calculation area

At C2 layer, most of signals from S2 layer are ignored other than maximum response of each feature. Neighbor outputs from S2 are assumed to be similar, thus it is reasonable to restrict S2 calculation on some interest points. Local feature based object categorization approaches, like bags of features, usually adopt an interest point detector, (e.g., Harris operator, difference of Gaussian)[10][13]. To reduce process time, we avoid these well-known detectors, but use the positions, in which C1 takes local maxima or minima.

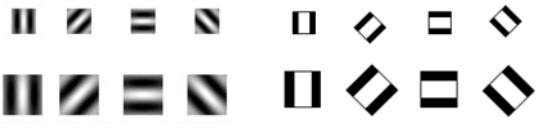
3.2.3. Simplification of S1 process

Haar wavelet is much simpler than Gabor function. Viola et al stated that Haar-like features can be calculated very rapidly using "integral image"[19], and Lienhart et al expanded Viola's approach by adding skewed Haar feature.[11] Therefore, we approximate Gobor function by Haar features.

Example of Haar features are shown in Figure 3. Output value of Haar filter is obtained by subtracting summed pixel value of black area from white one.

3.2.4. Retaining feature position

At C2 layer, all S2 signals are integrated to N elements, which mean information of feature location is lost. Mutch *et al.* expand standard model to retaining feature positions in order to improve recognition rate.[14] This is effective approach because this hierarchical approach keep a certain level of invariance of position at C layer. In our approach, C2 signals are divided to $N_{c2} \times N_{c2}$ areas, and integrated vector of N elements are calculated at each area. These areas are overlapped by ratio r_{c2} (0-1)



(a) Gabor filter (b) Haar-like feature

Figure. 3. Gabor filter and Haar-like feature

4. Experiments

4.1. Comparison with standard model

We tested this system and comparing it to the other object category recognition methods. Feature patches at S2 layer are trained by 1200 images downloaded through internet.

Table.1 describes the processing speed of our approach. In this table, we have succeeded to reduce processing time from more than 12 sec to about 1 sec. These results prove that our approach is fast enough.

Our approach was also evaluated with 4 categories of Caltech-101 image database, shown its example in Figure.4. Four object categories (airplane, car side, face, motorbikes) were trained with 40 positive images and 50 negative ones in "background" category of Caltech-101. This trained classifier was tested on 50 positive images and 50 negative images[18]. Its result is appeared in Table 2, which shows that our approach improves process much faster with only small decrease of accuracy, except for "faces". This reduction of accuracy in "faces" category might be caused by relatively various background, size, position, and illumination. This issue would be addressed in the future work.

4.2. Evaluation of each modification

Each change of this method was evaluated in recognition rate at Table 3, and in processing speed at Table 4 so

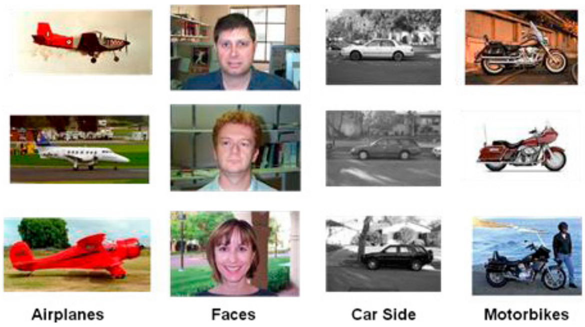


Figure 4. Example of Caltech-101 image set.

Table 2. Result of object category recognition (%)

Category	Our approach	Serre[18]
Cars(Side)	95.0	99.8
Faces	88.0	98.1
Airplanes	96.0	94.9
Motorcycles	96.0	97.4

Table 3. Evaluation of our modification in recognition rate.(%)

Category	Our approach	Without VQ	Full area calc of S2	Gabor Filter
Cars(Side)	98.8	97.7	98.8	100.0
Faces	91.5	86.3	91.5	93.5
Airplanes	95.3	96.3	96.5	95.3
Motorcycles	96.3	95.3	96.3	98.5
Average	95.5	93.7	96.1	97.3

Table 4. Evaluation of our modification in process time. (sec)

Process	Our approach	Full area calc of S2	Gabor Filter
S1	0.31	0.31	1.23
C1	0.27	0.30	0.23
S2	0.48	1.81	0.41
C2	0.05	0.04	0.05
SVM	0.00	0.00	0.00
Total	1.11	2.46	1.92

that the trade-off can be seen. Here we use 80 positive samples and 100 negative samples for each training.

"Without VQ" of Table 3 shows that when the number of S2 feature is the same, vector quantization improves recognition rate totally. In this case, their processing steps are also the same. Restriction of calculation area reduces processing time less than 1/3 at S2 layer, in spite of only 0.6% reduction of recognition rate.(See "Full area calc of S2") Implementation of Haar-like feature improves the processing speed of S1 4 times faster than Gabor filter, in exchange for 1.8% of recognition rate.

5. Conclusions

In this paper, we proposed image based search system using hierarchical object category recognition algorithm. To achieve practicality of system, we improve processing speed of standard model significantly with small accuracy decrease.

For more practicality, more challenges are required: more elimination of processing speed, improvement of recognition rate, multiple objects in one image, etc

References

[1] K. Barnard, D. Forsyth:"Learning the Semantics of Words and Pictures", Proc. of IEEE International Conference on Computer Vision, pp.408-415, July 2001

[2] S. Bileschi, L. Wolf: "A Unified System For Object Detection, Texture Recognition, and Context Analysis Based on the Standard Model Feature Set", Proc. of British Machine Vision Conference, Sept 2005.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, C. Bray: "Visual Categorization with Bags of Keyponts", Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision, pp.1-22, May 2004.

[4] Evernote: "EVERNOTE", http://evernote.com/

[5] L. Fei-Fei, R. Fergus, P. Perona: "Learning generative visual models from few training examples: an incremental Baysian approach tested on 101 object categories", Workshop on Generative-Model Based Vision, Computer Vision and Pattern Recognition, June 2004

[6] L. Fei-Fei, P. Perona: "A Bayesian Hierarchical Model for Learning Natural Scene Categories", Proc. of Conference on Computer Vision and Pattern Recognition, June 2005

[7] R. Fergus, P. Perona, A. Zisserman: "Object Class Recognition by Unsupervised Scale-Invariant Learning", Proc. of Conference on Computer Vision and Pattern Recognition, volume 2, pp.264-271, 2003

[8] idée: "TinEYE", http://tineye.com/

[9] Y. Jing, S. Baluja: "PageRank for Product Image Search", Proc. of International conference on World Wide Web, April 2008

[10] G Kim, C. Faloutsos, M. Hebert: "Unsupervised Modeling of Object Categories Using Link Analysis Techniques" Proc. of Conference on Computer Vision and Pattern Recognition, June 2008

[11] R. Lienhart, J. Maydt: "An extended set of Haar-like features for rapid object detection", Proc. of International Conference on Image Processing, Vol.1, pp.900-903, Sept 2002

[12] Y. Linde, A. Buzo, R. Gray: "An algorithm for vector quantizer design", IEEE Trans.Commun, COM-28, pp.84-95, 1980

[13] D. Lowe: "Object Recognition from Local Scale-Invariant Features", Proc. of the International Conference on Computer Vision, pp.1150-1157, Sept 1999

[14] J .Mutch, D. Lowe: "Multiclass Object Recognition with Sparse, Localized Features", Proc. of Conference on Computer Vision and Pattern Recognition, June 2006

[15] G. Pass, R. Zabih: "Histogram Refinement for Content-Based Image Retrieval", IEEE Workshop on Applications of Computer Vision, 1996

[16] Polar Rose: "Polar Rose", http://www.polarrose.com/

[17] G. Qi, X. Hua, Y. Rui, T. Mei, J. Tang, H. Zhang: "Concurrent Multiple Instance Learning for Image Categorization", Proc. of Conference on Computer Vision and Pattern Recognition, June 2007

[18] T. Serre, L. Wolf, T. Poggio: "Object Recognition with Features Inspired by Visual Cortex" , Proc. of Conference on Computer Vision and Pattern Recognition, June 2005

[19] P. Viola, M. Johnes: "Rapid Object Detection using a Boosted Cascade of Simple Features", Proc of Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii USA, I-511- I-518 vol.1, June 2001

[20] H. Zhang, A. Berg, M. Maire, J. Malik: "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition", Proc of Conference on Computer Vision and Pattern Recognition, June 2006