Object Distinction Technology for Information Retrieval by Mobile Cameras

Takayasu Yamaguchi, Hiroshi Aono and Sadayuki Hongo

Technology is proposed to enable the information retrieval through the use of images on advertisements or signs around town that are captured by a camera equipped in a portable terminal such as a mobile terminal. Object distinction technology has suffered from drops in accuracy caused by a variety of factors, and from long processing times due to the relatively large amount of information in images. The proposed technology enables more accurate and faster object-distinction processing.

1. Introduction

Information retrieval using the mobile Internet is becoming widespread by the diffusion of mobile terminals. Today, there are much information on the Internet related to real-world objects such as advertisements and announcements. It would be extremely convenient for users if opened information on the Internet could be searched for via real-world objects related to the information.

At the same time, the functionality of mobile terminals has been progressing rapidly and terminals equipped with a camera are flooding the market. A camera incorporated in a portable terminal is called a “mobile camera,” which, in addition to taking pictures, is also being used as a new means of inputting information such as scanning two-dimensional (2D) codes. This is much more convenient for users than traditional key input.

Against this background, the mobile camera is expected to be used as a sensor for acquiring information from one’s vicinity. Specifically, if objects in pictures taken with a mobile camera could be distinguished, and multimedia information related to those objects retrieved, not only would the effort required to search for and retrieve information be greatly reduced but a variety of new information retrieval applications could be created. In short, the user could be provided with an extremely con-

venient means of retrieving multimedia information. A variety of items can be considered as targets to be distinguished by users in mobile environment, but we have chosen flat objects such as signs and posters for the following reasons.

- Signs are a widely used by means of informing users about what services are available at the location in question.
- Taking a picture of a sign related with information that a user desires is an intuitive action with a high degree of service potential.
- A high degree of correlation can be expected between information conveyed by signs and information placed on the Internet.

In this article, we first examine the problems with current technology and present preconditions for the proposed technology. We then describe new object distinction technology focusing on image representation [1], learning and distinction techniques [2], and use of location information [3]. We also introduce a prototype system that we have constructed to apply this object distinction technology, and we touch upon the future outlook for this technology.

2. Object Distinction Technology

2.1 Problems with Existing Distinction Techniques

A number of distinction techniques are currently in use such as 2D codes and digital watermarks[*1]. These techniques suffer from the problems described below when attempting to distinguish a sign.

1) 2D Codes

Large 2D codes must be attached to signs when dealing with pictures taken from afar, and it might detract the design of the sign.

2) Digital Watermarks

Thought it is inevitable that pictures of signs around town will be taken from a variety of angles, many systems are sensitive to the angle between the position of the distinction target and the user's position when shooting the picture. Signs that cannot be shot from straight on would make it difficult to retrieve the digital watermark.

3) Radio Frequency IDentification (RFID)

The distance up to which an RFID tag can be read is usually limited, making it difficult to distinguish highly elevated signs.

4) Character Recognition

It is difficult to distinguish signs not having expected character fonts. There are many signs that use special character fonts or signs that consist of only the product images with no characters.

In contrast, a technique for distinguishing images using image processing would be advantageous since it could directly register and distinguish target images on existing signs without having to make any modifications to those signs. Image processing, however, is generally time consuming, and target images situated outdoors (hereinafter referred to as "outdoor images") will undergo changes every moment according to weather, time, etc. Making accurate identifications of images is therefore difficult.

2.2 Preconditions

For this proposal, it is assumed that object distinction technology is to be applied to a "town directory" or other means of providing public information. We therefore begin by limiting the objects to be distinguished to flat objects like signs and posters while aiming to apply this technology to more general objects in the future. Up to now, various techniques have been proposed for distinguishing objects taken by a camera. But since lighting conditions in outdoor can change drastically, there is no guarantee that any of these techniques can accurately distinguish outdoor objects. Our aim, therefore, is to establish general-purpose object distinction technology that stresses robustness to environmental changes such as ever-changing lighting conditions.

Since flat objects like signs and posters have a variety of shapes and colors, we here target diverse types of signs and posters as objects of distinction instead of limiting ourselves to specific colors and shapes. The difficulty of distinguishing flat objects like signs and posters will differ according to types, set up numbers, and conditions. The results of a survey conducted on signs set up along Shibuya Center Street in Shibuya district of Tokyo revealed that most signs are the following three types: ordinary signs with no light source of their own, signs that are purposely lighted by lamps, and signs within which light sources such as fluorescent lights have been installed resulting in a sign that emits light itself. Even if it is a same sign, differences were noticeable between daytime and nighttime lighting regardless of the type. We therefore treat pictures of signs taken

*1 Digital watermark: Technology for embedding information in images, moving pictures and audio with hardly no effect on picture or sound quality. This information can be retrieved by digital-watermark detection software.

during the day and night as two different data groups (classes) having different characteristics.

In addition, the proposed technology assumes that a sign is to be distinguished based on a still picture taken only once by the user. We therefore exclude signs whose light source or the sign itself changes over time (such as dynamic neon lighting, scrolling signs and television signs). When taking a picture of a sign, some parts of the sign may not be visible due to the shape of the sign or the angle of shooting. To prevent such hidden parts from occurring, we limit the targets of distinction to flat objects as described before. We also set shooting requirements to be uniform in the position and size of target objects contained within pictures taken by a camera, so when taking a picture, the object should be centered on the screen and made as large as possible without any part of the object sticking out beyond the screen (the vertical length of the object should match the vertical length of the screen, or the horizontal length of the object should match the horizontal length of the screen). Furthermore, the equipment performing object distinction must register learning images beforehand for use as reference, and store owners will be expected to register photos of their shop signs. Therefore, to minimize the work of such registration, a small number of learning images is to be used for each sign.

Finally, the results of distinction processing will be presented to the user on the screen of the mobile terminal. These results may include multiple candidates, and up to ten candidates will be displayed in order of ranking.

3. Image Representation

3.1 Problem Analysis and Extraction of Requirements

When displaying the results of distinction processing to the user, the target object of distinction may not be ranked high on the displayed results if the overlap between class distributions in feature space[*2] rises up, the Bayes error occurs. To make distinction processing more accurate, it is necessary to represent image features in such a way that separates the distributions of each class in feature space.

To improve the degree of distribution separation, we set the following requirements.

1) Unavoidable background effects that occur even when observing the shooting requirements, and it must be removed as much as possible to achieve image representation robust to environmental changes.
2) Attention must be given to features such as brightness, color and texture (pattern) unique to a sign to provide a better representation of its image. While those features such as these will fluctuate depending on the shooting environment, the improvement in image representation that they provide should result in a greater degree of separation between class distributions and improve the performance of sign distinction, which is our objective here.

*2 Feature space: The arrangement of quantified features (in this study, characteristics such as brightness, color and texture) along coordinate axes.

3.2 Representation Method

Given the above requirement that background effects must be removed as much as possible, we investigate methods of image representation when taking pictures while observing shooting requirements. As described in “Preconditions” in Section 2.2, the user is expected to shoot the target object so that it appears as large as possible in the center of the screen. Assuming, therefore, that a sign targeted for distinction processing will exist in the center of the camera picture, we can assign different weights to the center region and peripheral region of the picture and represent that picture with an emphasis on the former region. Also, given the requirement that the unique features of a sign such as brightness and color need to be better represented, we can divide the camera picture into a number of cells to form a grid and represent the picture in each cell by a $L^*a^*b^*$ histogram.

For purposes of comparison and evaluation, we represent pictures taken with a camera using several methods. These are a newly proposed $L^*a^*b^*$ histogram method that divides the picture into 7 × 7 cells and emphasizes the center region, the traditional $L^*a^*b^*$ histogram method, the gray method, the frequency method, and the morphology method [1]. The color histogram is a simple method of image representation using color. Although it is weak in representing change in lighting conditions, the fact that the color scheme of many signs is chosen for maximum psychological effect makes this a good technique for representing color features that can facilitate sign distinction. In this article, we use an equivalent color space in which the distance between colors corresponds to perceptual differences in those colors, and adopt, in particular, the $L^*a^*b^*$ color space, which is a colorimetric system representative of equivalent color spaces.

Image-representation methods that use gray, frequency and morphology do not take color changes into account. Here, methods that treat pixel value in a gray image as a feature quantity

suffer to some extent from high-dimension feature vectors. But dramatic increases in computer processing speed have made it possible to handle feature vectors even of the order of several thousand dimensions. Likewise, methods that treat picture frequency as a feature quantity also tend to produce high-dimension feature vectors and also demand a particular vertical and horizontal picture size in order to apply a Fast Fourier Transform (FFT) for extracting the feature immediately. Finally, there are methods that apply morphology processing to the picture of a sign in order to extract the features of the object. But morphology processing can be achieved by a wide variety of operations, and specific extraction methods must be examined to obtain optimal features from the picture of a sign.

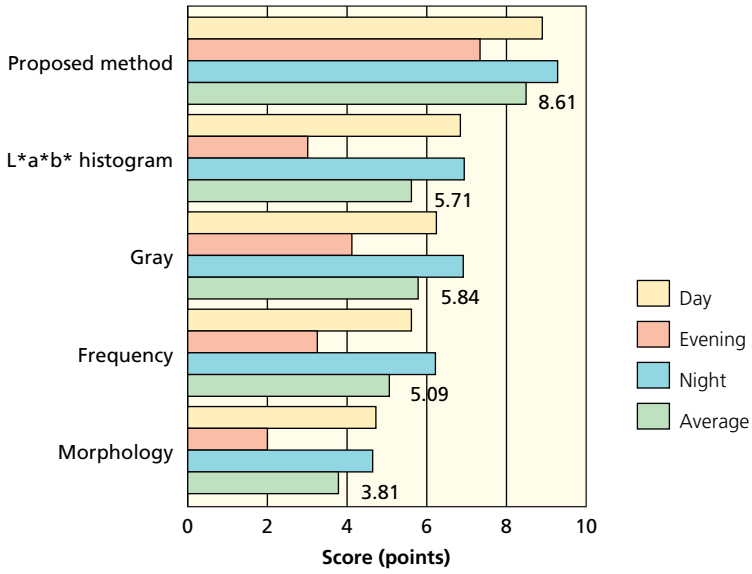

Figure 1 Evaluation of representation methods

3.3 Evaluation Scheme for Representation Methods

To evaluate the above representation methods, we assume to display ten candidates on the screen of the mobile terminal as in preconditions described in Section 2.2. The result of distinction processing for a particular object can therefore be scored by assigning 10 points if the object appears at the top of the candidate list, 9 points if it appears second on the list, and 0 points assigned if it fails to make the list at all.

3.4 Evaluation of Representation Methods

For this evaluation, we targeted 520 flat objects such as signs and posters all within 25 meters of each of five intersections on Shibuya Center Street near the Shibuya station. **Figure 1** shows the results of distinction processing for each of the representation methods described above using the *k*-Nearest Neighbor (*k*-NN) distinction method with k=1. The *k*-NN method is said to be effective for separating high-dimension data in the case of a small number of samples [2].

On comparing the average scores for day, evening and night obtained by the evaluated methods, we see that the proposed method scores the highest indicating that separation performance is good in distinction processing. Compared with the next best performance achieved by the gray method, the proposed method is about 2.8 points better. A score of 8 points or better means that the object searched for by the user can be displayed on the mobile terminal's screen as one of the top three candidates on average.

4. Learning/Distinction Techniques

4.1 Requirements of a Good Learning/Distinction Technique

We can set specific requirements based on the following three points of view under the preconditions described in Section 2.2.

1) Display Correct Answers on the top

To make it easy for the user to select the target object from the candidates displayed, the correct result should be displayed on the top of the screen. In particular, the target object should be displayed, on average, as one of the top five candidates at the least and as one of the top three candidates if possible.

2) Short Learning Time

When performing registration and learning for a certain sign, the store owner must be able to confirm that the sign has been correctly registered. Learning time should therefore be within one minute at the least and no more than one second if possible.

3) Short Distinction Time

A user who takes a picture of a target object at a street corner should be immediately provided with the information related with that object. Distinction time should therefore be within one second at the least and no more than 100 ms if possible.

4.2 Learning/Distinction Techniques

Now, using the image-representation method proposed in Section 3, we can perform distinction processing on pictures of

signs. But first, to determine a learning/distinction algorithm applicable to that method of image representation, we evaluate typical learning/distinction algorithms currently in use [2]. These are Fisher's Linear Discriminant (FLD), the subspace method, Learning Vector Quantization (LVQ), Support Vector Machine (SVM), *k*-NN and Naive Bayes (NB).

In FLD, the ratio of the variance between two classes to the variance within each class is called the "Fisher criterion." Maximizing this ratio separates the two classes. The FLD method is of the "binary distinction" type and comes in two forms: In one form, samples of a sign are dealt with as positive samples and samples of complementary signs are dealt with as negative samples only once. In the other form, samples of a sign are dealt with as positive samples and samples of the other sign are dealt with as negative samples, and this is repeated for the number of class combinations. In this study, we tried the latter form performing dimension compression for every two classes.

As for the subspace method, many variations are known such as one that can configure complex separation boundaries. We here use CLAss-Featuring Information Compression (CLAFIC), the most common subspace method. The CLAFIC method creates subspaces through KL expansion[*3] of each class, and determines the class of certain unknown data by the extent to which that data matches the subspaces of each class.

The LVQ method assigns categories to input data vectors and coupled-weighted vectors. It learns by repeatedly comparing categories and making the distance between input vectors and coupled-weighted vectors closer if those categories agree and making that distance farther if they don't.

The SVM method configures a distinction boundary that maximizes the margin between two classes of data. It can separate high-dimension data at high speed by introducing a kernel. In this study, we apply SVM by simple distinction and rank classes by the distance between the distinction boundary and unknown data.

The *k*-NN method makes a decision on a target of distinction by majority rule based on the distance between unknown data and learning prototypes. Here, *k* represents the number of learning prototypes to be used for majority-rule decision making. The value of *k* may be set by one of two methods: estimate optimal *k* for distinguishing the learning prototypes or make *k* a constant. Here, considering that distinction processing can usually be performed correctly even if the number of learning prototypes per class is small and *k* is made fixed, we tried using the constant-*k* method.

Finally, the NB method learns what parameters for each class distribution maximizes the a posteriori probability based on learning data and performs distinction processing by comparing the distribution of unknown data with that of each class. This method adjusts a hyperparameter (ξ) of a multinomial distribution for each class to achieve the best separation between learning samples.

*3 Karhunen-Loéve expansion: A method of performing function expansion using statistical properties.

4.3 Evaluation Scheme for Learning/Distinction Techniques

We use the scheme shown in **Table 1** for evaluating learning/distinction techniques based on the requirements described in Section 3.1., "learning time" refers to the time needed for learning all signs within a certain area on Shibuya Center Street, and "distinction time" means the time needed to perform distinction processing on one picture taken of an unknown sign (distinction time per object).

4.4 Evaluation of Learning/Distinction Techniques

We selected 145 objects within the periphery (a radius of 20 m) of one intersection as an effective means of performing a comparison experiment, and evaluated the performance of the above learning/distinction techniques within Shibuya Center Street area. **Table 2** lists the results of this evaluation.

To begin with, we see that the *k*-NN (*k* fixed) method achieved an average score of 8.8 points (indicating that an object would be displayed as one of the top 2.2 candidates),

Table 1 Evaluation scheme for learning/distinction techniques

Result	Score for 145 objects	Learning time for 145 objects	Distinction time per object
○	8 or more points	Under 1 s	Under 100 ms
△	6-7 points	1-59 s	100 ms–999 ms
×	Under 6 points	1 min or more	1 s or more

Table 2 Evaluation of learning/distinction techniques

Technique	Average score for 145 objects	Learning time for 145 objects	Distinction time per object
FLD	○8.3 points	×5.7 h	△454ms
Subspace	○9.0 points	△24 s	○6ms
LVQ	△7.6 points	×1 min, 53 s	○7ms
SVM	○9.1 points	×3 min, 44 s	△280ms
NB	○8.8 points	○0.4 s	○46ms
k-NN	○8.8 points	○No learning	○94ms

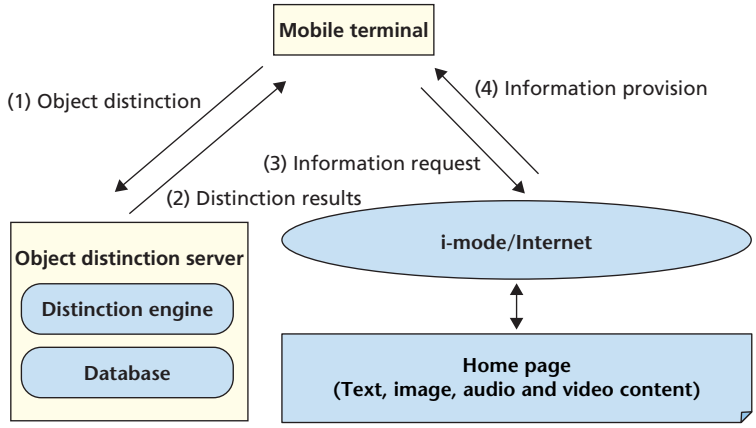

Figure 2 Configuration of object-distinction prototype system

required no learning, and achieved a distinction time per object of 94 ms. The NB method (with ξ adjustment) also achieved an average score of 8.8 though with a learning time of 0.4 s but a distinction time of 46 ms. These two techniques therefore satisfy the requirements described in Section 4.3, namely, display of the target object as one of the top three candidates, learning time within one second, and distinction time within 100 ms.

5. Prototype System Construction Using Location Information

A proposal has been made for a method to achieve accurate and high-speed object distinction by using location information of mobile characteristic to limit search candidates to the relatively a small number of objects in the user's vicinity [3].

The object-distinction prototype system that we have constructed consists of a server and mobile terminal. The server has at least a database and distinction engine. It performs object distinction, registration, and modification in accordance with location information received from the mobile terminal and provides object-related information. The mobile terminal accepts operations by the user, sends location information to the server, and requests object distinction, registration, and modification and the provision of object-related information. After object distinction, the user can jump to the home page related to the object via i-mode enabling smooth access to text, image, audio and video content. **Figure 2** shows the configuration of this object-distinction prototype system.

Distinction time when using a FOMA terminal is about 10 s including communication time despite the fact that positioning time is not considered as this terminal has no location-information acquisition function. In the future, greater memory capacity in mobile terminals may enable distinction processing to be performed within the terminal itself. We can envision, for example, how entering an amusement park could trigger the download of learned data to the mobile terminal all at once enabling distinction processing to be performed entirely at the amusement park without sending pictures to a server.

6. Conclusion

Focusing on the identification of outdoor signs, this article described image representation, learning and distinction techniques, and use of location information as the prime components of new object distinction technology developed to solve the problems of existing distinction techniques. This newly proposed technology was used in combination with i-appli and a server connected to the Internet to construct a prototype object distinction system. For the future, we plan to improve distinction performance, improve positioning accuracy, shorten positioning and communication time, and enhance mobile terminal functions with the aim of simplifying information retrieval by portable terminals when outside the home and expanding the business aspects of this technology to advertising, games, events, etc.

References

[1] T. Yamaguchi, H. Aono and S. Hongo: "Feature of signboard pictures using a Mobile Camera," Technical Report of IEICE, PRMU2004-105, Vol. 104, No. 448, pp. 1–6, 2004 (In Japanese).

[2] T. Yamaguchi, H. Aono and S. Hongo: "Discrimination of signboard pictures using a Mobile Camera," Technical Report of IEICE, PRMU2004-106, Vol. 104, No. 448, pp. 7–12, 2004 (In Japanese).

[3] T. Yamaguchi, M. Takahata and S. Hongo: "The Information Handling Technology using Location Information," Technical Report of IPSJ, MBL02021017, Vol. 2002, No. 49, pp. 101–106, 2002 (In Japanese).

ABBREVIATIONS

CLAFIC: CLAss-Featuring Information Compression
FFT: Fast Fourier Transform
FLD: Fisher's Linear Discriminant
k-NN: *k*-Nearest Neighbor
LVQ: Learning Vector Quantization
NB: Naive Bayes
RFID: Radio Frequency IDentification
SVM: Support Vector Machine