

Master thesis

# Extraction and Recognition of Shoe Logos with a Wide Variety of Appearance using Two-Stage Classifiers

March, 2018

**Kazunori Aoki**

Human Interface Laboratory

Division of Information Engineering

Graduate School of Engineering

Mie University

# Abstract

A logo is a symbolic presentation that is designed not only to identify a product manufacturer but also to attract the attention of potential buyers. Manufacturers carefully design their logos so that their characteristics, impressions and philosophies are expressed. Moreover, logos on a person's belongings can play an important role in characterizing and identifying the person. Extraction and recognition of logos from images captured by multiple surveillance cameras could provide useful information for identification and tracking of individuals.

Automatic extraction and recognition of shoe logos using image analysis techniques is challenging because they have characteristics that distinguish them from those of other products, and their appearance can vary substantially. Automatic extraction and recognition techniques must handle these problems properly, because shoes are usually worn on feet and, therefore, move frequently with respect to stationary cameras. Additionally, since shoe logos are usually appeared as an integrated design component of shoe design, they have significant within-class appearance variation due to the variation in color, fabric material, shape of shoe and dirt or aging of shoe.

The extraction and recognition of logos in images has attracted the attention of many researchers. Several studies on the automatic extraction of logos have been reported. Affine and non-rigid transformation occurs frequently in real-world images. This makes logo detection and recognition complex, especially for model-based approaches, owing to the difficulty in collecting sufficient samples to obtain a robust model. Several related works [8, 9, 14, 15] address this problem. While these related works were carefully designed for handling appearance variations, they did not pay enough attention for large within-class appearance variation. Therefore, extending the target of these methods for shoe logos is quite difficult. Moreover, deep neural network architectures have been employed in image recognition tasks due to

their promising performance and high adaptivity. However a deep learning method requires a large-scale, accurately annotated dataset for training. Creating such a dataset for deep learning is difficult because shoe logos have a wide variety of appearances even in the same brand.

In the present paper, we propose an automatic extraction and recognition method for shoe logos using a limited number of training samples. The proposed method employs maximally stable extremal regions (MSERs) [4] for the initial region extraction, an iterative algorithm for region grouping and gradient features, and two-stage support vector machines (SVM) for logo recognition.

For performance evaluation, we use the IEICE-PRMU shoe logo dataset which consists of shoe logo images captured in uncontrolled condition. The results of performance evaluation experiments show that the proposed method achieves promising performance for both logo extraction and recognition.

In the present paper, Chapter 1 gives the introduction of this research. Chapter 2 expresses related works. Chapter 3 shows the proposed method in detail. Chapter 4 represents evaluation experiments. In the last Chapter, a conclusion is explained. Appendix A submits the method which proposed methods are employed. Appendix B describes individual diversification extraction strategies. Postscripts shows where the program is located and presentation slide which is used on master research presentation.

# Contents

Abstract	i
Chapter 1 Introduction	1
1.1 Background . . . . .	1
1.2 The purpose of this research . . . . .	1
Chapter 2 Related works	4
2.1 Logo datasets . . . . .	4
2.2 Related works . . . . .	5
2.2.1 Model-based approaches . . . . .	5
2.2.2 Efficient searching based on feature descriptor . . . . .	6
2.2.3 Method based on invariant similarity measure . . . . .	7
2.3 Deep neural networks . . . . .	7
Chapter 3 The proposed method	9
3.1 Extraction of shoe logos . . . . .	10
3.1.1 Initial region extraction . . . . .	10
3.1.2 Iterative region integration . . . . .	11
3.2 Two strategies for extraction method . . . . .	12
3.2.1 Complementary color spaces . . . . .	12
3.2.2 Complementary similarity measures . . . . .	13
3.3 Recognition of logos . . . . .	14
Chapter 4 Evaluation Experiments	16
4.1 Dataset . . . . .	16



4.2	Training of classifiers . . . . .	17
4.3	Evaluation . . . . .	18
4.4	Results and Discussion . . . . .	20
4.4.1	Comparison of logo extraction performance . . . . .	20
4.4.2	Logo extraction and recognition performances . . . . .	21
Chapter 5	Conclusion	25
5.1	Conclusion . . . . .	25
5.2	Feature works . . . . .	25
AppendixA	The overview of employing methods	27
A.1	Maximally Stable Extremal Region(MSER) method . . . . .	27
A.1.1	Extraction of MSERs . . . . .	27
A.2	The grayscale gradient histogram features . . . . .	30
A.2.1	Extraction of Gradient histogram features . . . . .	30
A.2.2	The dimensionality of extracted gradient feature vector . . . . .	32
A.3	SVM (Support Vector Machine) . . . . .	32
A.3.1	Kernel trick . . . . .	34
A.3.2	An SVM scaling . . . . .	34
AppendixB	Individual strategies for extraction method	35
B.1	A strategy which varies color spaces . . . . .	35
B.2	A strategy which combines similarity measures . . . . .	35
AppendixC	Programs file location	37
AppendixD	Briefing paper	38
Acknowledgements		41

# Chapter 1

## Introduction

In this chapter, Chapter 1.1 describes the background of this research, Chapter 1.2 describes the purpose of this research.

### 1.1 Background

A logo is a symbolic presentation that is designed not only to identify a product manufacturer but also to attract the attention of potential buyers. Manufacturers carefully design their logos so that their characteristics, impressions and philosophies are expressed. Moreover, logos on a person's belongings can play an important role in characterizing and identifying the person. Logo recognition aims to recognize the logo brand of the input image, and logo extraction aims to find the locations of logo objects in the input image. The combination of extraction and recognition of logos from images captured by multiple surveillance cameras could provide useful information for identification and tracking of individuals.

### 1.2 The purpose of this research

Automatic extraction and recognition of shoe logos using image analysis techniques is challenging because shoe logos have characteristics that distinguish them from the logos of other products, and their appearance can vary substantially. Fig.1.1 shows examples of shoe logos captured by standard still cameras. The logos shown in Fig.1.1(a) and (b), which belong to the same company, have the same shape but different colors. Fig.1.1(c) and (d) are exam-

ples of logos consisting of multiple components. Fig.1.1(e) and (f) show the most common appearance variations of shoe-logo images, i.e., rotation, occlusion, and perspective distortion. Automatic extraction and recognition techniques must handle these problems properly, because shoes are usually worn on feet and, therefore, move frequently with respect to stationary cameras. Additionally, since shoe logos are usually appeared as an integrated design component of shoe design, they have significant within-class variation due to the variation in color, fabric material, shape of shoe and dirt or aging of shoe.

In this research, we propose an automatic extraction and recognition method for shoe logos using a limited number of training samples. The proposed method employs maximally stable extremal regions (MSERs) [4] for the initial region extraction, an iterative algorithm for region grouping and gradient features, and two-stage support vector machines (SVM) for logo recognition. For performance evaluation, we use the IEICE pattern recognition and media understanding (IEICE-PRMU) shoe logo dataset [5]. This dataset consists of shoe logo images captured in uncontrolled condition.

The main contributions of this research are the following:

1. We propose a method which extracts and recognizes shoe logos which contains wide within-class variety due to the reasons mentioned above.
2. In order to improve the performance of extraction and recognition, two classifiers of which objectives are different are introduced.
3. For performance evaluation, IEICE-PRMU shoe logo dataset is used. While the dataset is originally created for a PRMU algorithm contest, its property where real shoe logos captured in uncontrolled condition enables us to evaluate the method for logo detection and recognition in real scenario.

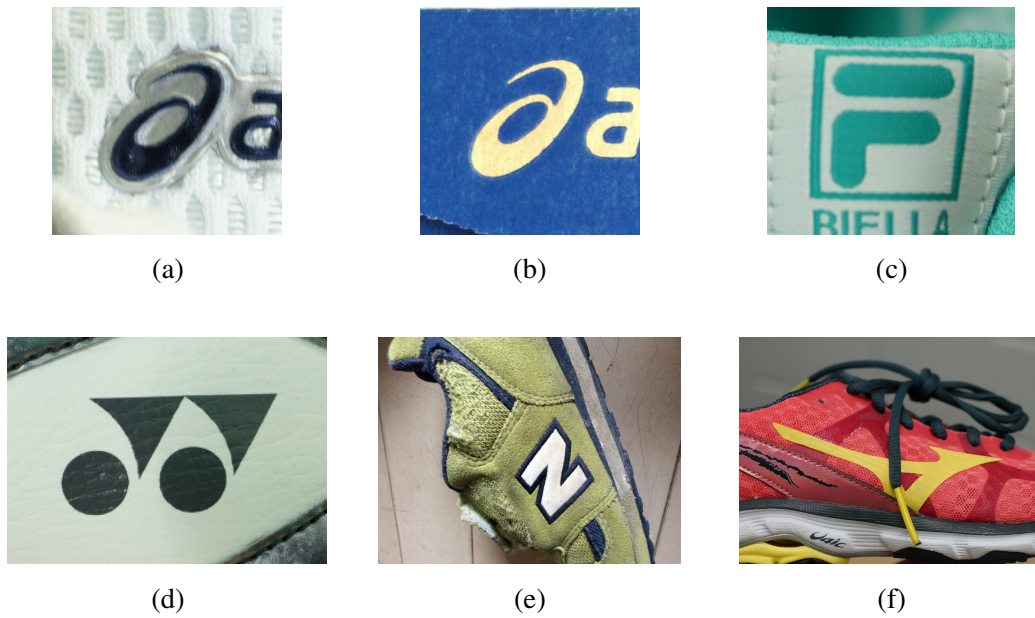


Fig. 1.1: Examples of shoe logos [5]: (a) and (b) examples belonging to the same brand but with different colors, (c) and (d) examples containing multiple connected components, (e) rotation, and (f) occlusion and rotation.

## Chapter 2

# Related works

The extraction and recognition of logos in images has attracted the attention of many researchers. Several studies on the automatic extraction of logos have been reported. Chapter 2.1 describes logo datasets which is often used. In Chapter 2.2, we provide brief review for logo detection and recognition methods and discussion about relation between the present and conventional works.

### 2.1 Logo datasets

There are multiple logo datasets (1) the BelgaLogos [1], (2) the FlickrLogos [2] and (3) the LOGO-NET [3] dataset which attract attention in image analysis.

1. The BelgaLogos dataset is composed 10,000 images. The dataset consists of 37 different brand logos. An image contains multiple logos and ground-truths. The annotated instances have then visually classified as “OK” or “junk” by a set of 3 users. Instances of which annotation is classified as “OK” are used.
2. The FlickrLogos dataset consists of 8,240 real-world images and 32 different brand logos. The dataset has also 6,000 non-logo images. The dataset is split into three disjoint subsets: (1) the training set which consists of 10 images per one brand that were extracted manually, (2) the validation set and (3) the test set which contains 30 images per one brand. The validation set is used for investigating the parameter of the method.

3. The LOGO-NET dataset consists of 73,141 images and 130,608 annotated logo objects. The dataset consists of 160 logo classes with 100 brands. All images of the dataset is collected from online retail markets.

Several studies which used above datasets for evaluation are shown in the following chapter.

## 2.2 Related works

### 2.2.1 Model-based approaches

Affine and non-rigid transformation for object appearance occur frequently when we take a picture of an object in real-world images. This makes logo detection and recognition complex, especially for model-based approaches, owing to the difficulty in collecting sufficient samples to obtain a robust model. To address this problem, Farajzadeh et al. [8] proposed an exemplar-based method for logo or trademark recognition. Their method uses new sample synthesizing which generates training logo images from standard logos with different tilts and rotations. This approach employs a two-stage strategy where initial candidate regions are extracted using efficient sub-window search (ESS) [10]. To perform localization, one can take a sliding window approach, but this strongly increases the computational cost, because the classifier function has to be evaluated over a large set of candidate subwindows. The underlying intuition of ESS is the following: even though there is a very large number of candidate regions for the presence of the objects we are searching for, only very few of them can actually contain object instances. One should target the search directly to identify the regions of highest score, and ignore the rest of the search space where possible. The branch-and-bound framework allows such a targeted search. It hierarchically splits the parameter space into disjoint subsets, while keeping bounds of the maximal quality on all of the subsets. This way, large parts of the parameter space can be discarded early during the search process by noticing that their upper bounds are lower than a guaranteed score from some previously examined state. ESS extracts regions very fast because it relies on a branch-and-bound search instead of an exhaustive search. These regions are then recognized using a linear SVM trained using new synthesized samples. Experiments is based on the FlickrLogos dataset. The dataset for classifier training includes logo images synthesized using gamma correction, difference of

Gaussian filtering, equalization of variation, and size normalization. The main disadvantage of ESS is the significant tradeoff between recognition accuracy and false-positive detection. When the number of synthesized samples is increased to improve recognition accuracy, the number of false-positive detections drastically increases.

Chu et al.[9] proposed a method using visual patterns. This approach first extracts scale-invariant feature transform (SIFT) features [11] from both test images and a logo image. Features with high similarity in both test images and a logo image are found using locality sensitive hashing (LSH) [12]. LSH reduces the dimensionality of high-dimensional data such as SIFT features. LSH hashes input items so that similar items map to the same “buckets” with high probability. LSH differs from conventional and cryptographic hash functions because it aims to maximize the probability of a “collision” for similar items. The main purpose of their method is to improve computational efficiency by eliminating outliers in a test image obtained from an exhaustive sliding-window search. The extracted feature points are combined using the mean-shift algorithm [13] to extract local logo regions. The extracted regions are classified using visual word histograms (bag of words) and visual patterns. Experiments is based on two recently-proposed datasets: the BelgaLogos dataset and the FlickrLogos dataset. This method obtains high computational efficiency at the sacrifice of extraction accuracy. The authors reported that the method obtains only 19.0% recall and 30.0% precision.

### 2.2.2 Efficient searching based on feature descriptor

As another application of LSH for logo detection, Romberg et al. proposed the bundle min-hashing[14, 15]. This method calculates features which are more robust to appearance variation than single visual word by bundling visual word and spacial neighborhood features. The recognition of logos in novel images is then performed by querying a dataset of reference images. Logos in reference images is ranked regarding logos in novel images based on the similarity between a reference image and a novel image. Moreover, this approach retrieve logos more robust in order to use new Random Sample Consensus (RANSAC) for extremely fast reranking. This approach outperformed existing another approaches in logo recognition performance by combining synthetic data augmentation on FlickrLogos dataset.

While these three methods were carefully designed for handling appearance variations due

to affine and non-rigid transformation which occur frequently in real-world images, they did not pay enough attention for large intraclass variations. Therefore, it is quite difficult for these methods to extend for detection and recognition of shoe logos which have significant appearance variations.

### 2.2.3 Method based on invariant similarity measure

Logo detection and recognition can be categorized as a sub problem of object retrieval. In the literature of object retrieval, evaluating similarity between two images, a query and a gallery, is important. Shen et al. [16] proposed spatially-constrained similarity measure (SCSM) for large-scale object retrieval. The method focus one fundamental problem: the loss of spatial information when representing the images as histograms of quantized features. Theremore, the method incorporate spatial information to quantized features. SCSM could handle object rotation, scaling, view point change and appearance deformation. Furthermore, based on the retrieval and localization results of SCSM, they introduced a robust re-ranking method with the  $k$ -nearest neighbors of the query for automatically refining the initial search results. While high performance of this method on object retrieval and object categorization was confirmed by experiments, such high performance may not be expected for shoe logos due to its variation of design and colors.

## 2.3 Deep neural networks

Deep neural network architectures have been employed in image recognition because of their promising performance and high adaptivity. Girshick et al. [17] proposed a method called regions with convolutional neural network (R-CNN) for generic object recognition. R-CNN is an approach that combines selective search [18] and CNN. The method extracts initial regions using efficient graph-based image segmentation [19]. It then iteratively groups regions using similarity calculated from appearance features (color, texture, size and fill features) in the regions. Although R-CNN achieved a high performance score on PASCAL2010 [20], the method requires a large-scale, accurately annotated dataset for training. Creating such a dataset for shoe logos is difficult because they have a wide variety of appearances even



in the same bland.

To overcome the problem of deep-learning based method where the performance depends on the size and quality of training dataset, synthesizing dataset is the most common approach. Su et al. [21] proposed an algorithm for generating synthetic context logo (SCL) training images for deep learning technique. SCL training data generation method synthesis logo images in context by overlaying a transformed logo exemplar (scaling, shearing, rotation and coloring) at a random location in non-logo context images to deal with unknown background clutters. The experiment are employed Fast R-CNN method [22] which significantly improves computational efficiency of R-CNN method on FlickrLogos dataset and handcrafting dataset. SCL training data generation method achieve better score than no SCL method. While this method is considered to be applicable for shoe logo detection and recognition, difficulties on preparing exemplar of shoe logo of which design is integrated with shoe itself will happen.

From the dataset perspective, a dataset for evaluating logo detection and recognition have been proposed. LOGO-NET proposed by Hoi et al.[3] is one of the largest dataset for logo detection and brand recognition. To facilitate deep-learning based logo detection, they built a large scale dataset by exhaustively collecting images from online retailer website. They employ R-CNN, Fast R-CNN and Spatial pyramid pooling (SPP) net in order to compare these methods. Consequently, R-CNN obtained the best logo detection and brand recognition results among three methods. This is mainly because company logos tend not to be the primary object the photographer intended to depict and just happen to get recorded in the background along with the scene. Fast R-CNN and SPPnet might fail to detect if the convolutional feature map is not large enough. In contract, R-CNN suffers less from this issue because R-CNN first takes an RoI (a small region) and then resize it to a fixed size (essentially enlarged) before it is passed to the convolutional network. However, the inherent property of these images in which the background is simple and plane is sometimes different from real-world image.

## Chapter 3

# The proposed method

This chapter describes the outline of the proposed method.

Figure 3.1 shows the flow of the proposed method. In the proposed method, we input one still image of a single color space. The method outputs regions in which a logo appears and the class (name of brand) corresponding to each region. The proposed method consists of main two stages: extraction of shoe logos and recognition of the extracted logo regions. The following sections describe each stage in detail.

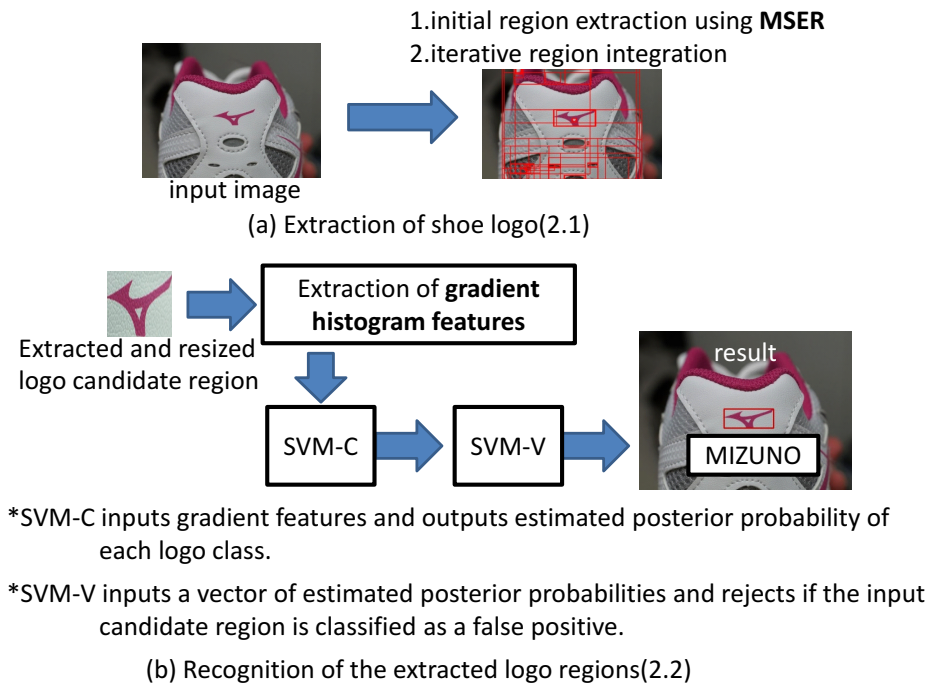


Fig. 3.1: Overview of the proposed method

## 3.1 Extraction of shoe logos

The purpose of this stage is to extract all regions that might contain a logo from an input image. The extraction stage consists of an initial region extraction using MSERs and iterative region integration using three similarities.

### 3.1.1 Initial region extraction

The initial region extraction for shoe logos employs MSERs [4] for each color plane of the input image. The color space which is used in the initial region extraction is described in the Chapter 3.2. The MSER method is a region segmentation based on pixel values in a grayscale image. Appendix A.1 describes the MSER method in detail. To employ MSER sufficiently for the input shoe logo image, preprocessing consisting of image smoothing and histogram equalization is performed.

First, we smooth the image to reduce noise. It is necessary to preserve region edges for extracting sharp-shape regions such as logos. Therefore, we use a bilateral filter for noise reduction and edge preserving. Early trials and investigation suggested that image smoothing using median or Gaussian filters is unsuitable for segmenting regions for logo extraction because these filters remove edges as well as noise. We adjusted the parameters for the bilateral filter so that they extract the smallest logo in the reference image.

Second, we use grayscale histogram equalization to reduce the effects of illumination variation.

Finally, we separate the smoothed and equalized images into three color-plane images. The MSER segmentation algorithm is applied to each color-plane image. This color-plane separation enables extraction of shoe logos with multiple colors. We then combine regions extracted from the three color-plane images to reconstruct the initial candidate regions. The parameters for the MSER algorithm were determined by preliminary investigation using reference images such that the smallest logo is extracted correctly. Variations in another parameters contained in the algorithm of MSER do not influence the result of initial region extraction.

### 3.1.2 Iterative region integration

The above extraction process extracts single connected components as initial regions. Because some logos contain multiple connected components, we perform region integration using three similarities to extract these multiple connected components as one integrated candidate region. The basic concept of region integration is similar to hierarchical grouping algorithm in selective search [18]. The method continues grouping candidate regions until all the MSERs are merged into a single connected component. Figure 3.2 shows an example of this process. The region integration works iteratively as follows.

The input and output of the region integration are a set of extracted initial regions  $R^0 = \{r_1, \dots, r_n\}$  and a set of integrated regions  $R^C$ , respectively. In the example shown by Figure 3.2,  $r_1$  to  $r_4$  in (b) are the initial regions extracted from (a).

1. Initialize  $R^C$  as  $R^0$ :

$$R^C = R^0. \quad (3.1)$$

Let  $k = 0$  and proceed to the next step.

2. Determine two different regions  $(r_i, r_j) \in R^k, (i \neq j)$  that maximize similarity  $s(r_i, r_j)$ . In the example (Figure 3.2 (b)),  $r_1$  and  $r_2$  are selected as the regions which maximize the only color similarity.
3. Subtract  $r_i$  and  $r_j$  from  $R^k$  and create  $R^{k+1}$  using  $R^{k+1} = R^k - \{r_i, r_j\}$ . Create a new integrated region  $r^{(ij)} = r_i \cup r_j$  and add  $r^{(ij)}$  to  $R^{k+1}$  and  $R^C$ .
4. Increment  $k$  and iterate Steps 2 to 4 until the number of elements in  $R^k$  equals one. In the example, sub figures (c) and (d) show the components and integration at each iteration.
5. Output the region set  $R^C$  is the final integrated region result.

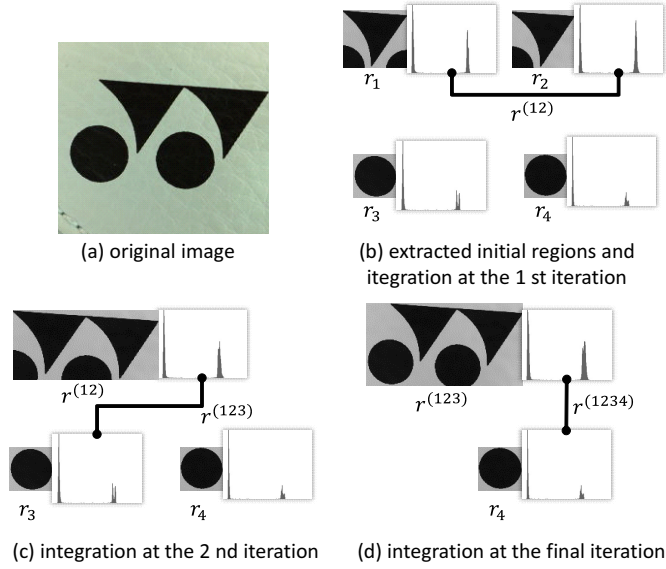


Fig. 3.2: Iterative region integration example for a logo consisting of multiple connected components (e.g. the YONEX logo). In each subfigure, extracted regions and their grayscale histograms are shown. The connected lines in each subfigure denote the combinations of connected components which provide the maximum similarity at each iteration. The region integration is repeated until adequate logo region is obtained.

## 3.2 Two strategies for extraction method

The design criterion for the best extraction of shoe logos is to create a set of complementary strategies are combined afterwards. We diversify extraction of shoe logos (1) by using a variety of color spaces with different invariance properties and (2) by using different similarity measures  $s(r_i, r_j)$ .

### 3.2.1 Complementary color spaces

We want to account for different scene and lighting conditions. Therefore we perform extraction of shoe logos in a variety of color spaces with a range of invariance properties. Specifically, we apply the following color spaces: *RGB*, the intensity (gray-scale image) *I*, *Lab*, *YCbCr* and *HSV*. In the present paper, we always use a single color space throughout extraction of shoe logos, meaning that both the initial region extraction and iterative region integration are performed in this color space. We apply *HSV* color space for extraction method, where we maximized the extraction performance. An applying color space of ex-

traction method are determined by Appendix B.1.

### 3.2.2 Complementary similarity measures

We define three complementary similarity measures. These measures are all in range 0 to 1.

1.  $s_{color}(r_i, r_j)$  measures color similarity. Similarity  $s_{color}(r_i, r_j)$  is calculated by

$$s_{color}(r_i, r_j) = \sum_{l=0}^{L-1} \min(h_l^{(i)}, h_l^{(j)}), \quad (3.2)$$

where  $h_l^{(i)}$  and  $h_l^{(j)}$  denote the  $l$ -th value of histograms obtained in regions  $r_i$  and  $r_j$ , respectively. We obtain the color histogram in (3.2) from each color plane with 16 bins. For example, when regions are detected in the red color-plane image, we obtain a color histogram using the red value. The color histograms are normalised in range 0 to 1. The connected components which construct a multi-component logo have similar color distribution. The  $s_{color}(r_i, r_j)$  is expected to properly capture this similarity.

2.  $s_{distance}(r_i, r_j)$  measures a distance between a centroid coordinate of  $r_i$  and  $r_j$ . Similarity  $s_{distance}(r_i, r_j)$  is calculated by

$$s_{distance}(r_i, r_j) = 1 - \frac{|centroid(r_i) - centroid(r_j)|}{diagonal(im)}, \quad (3.3)$$

where  $centroid(r_i)$  and  $centroid(r_j)$  denote a centroid coordinate of  $r_i$  and  $r_j$ , respectively. The  $diagonal(im)$  denotes a diagonal of the input image in pixels. The connected components which construct a multi-component logo are close to each other. The  $s_{distance}(r_i, r_j)$  is expected to properly capture this similarity.

3.  $s_{size}(r_i, r_j)$  encourages small regions to integrate early. This forces regions in  $R$ , *i.e.* regions which have not yet been merged, to be of similar sizes throughout the algorithm. For example, it prevents a single region from gobbling up all other regions one by one, yielding all scales only at the location of this growing region and nowhere

else. Similarity  $s_{size}(r_i, r_j)$  is calculated by

$$s_{size}(r_i, r_j) = 1 - \frac{size(r_i) + size(r_j)}{size(im)}, \quad (3.4)$$

where  $size(im)$  denotes the size of the image in pixels.

In the present paper, similarity measure is a combination of the above three:

$$s(r_i, r_j) = a_1 s_{color}(r_i, r_j) + a_2 s_{size}(r_i, r_j) + a_3 s_{distance}(r_i, r_j), \quad (3.5)$$

where  $a_i$  is investigated by Appendix B.2. We apply combined three similarities for extraction method because combined three similarities maximize extraction performance.

### 3.3 Recognition of logos

The extracted candidate regions are evaluated and classified in the recognition stage. We employ a two-stage strategy using two SVMs, one for classification (SVM-C) and one for verification (SVM-V). These SVMs are used consecutively for logo recognition, as shown in Figure 3.1. We employ the grayscale gradient histogram features [24] and an SVM in the recognition stage. The grayscale gradient histogram features method and an SVM method are described by Appendix A.2 and Appendix A.3, respectively.

We first clip the input image using the extracted logo candidate regions and convert them to a fixed size ( $50 \times 50$  pixels). We extract gradient histogram features. The parameters of the gradient histogram features are determined by our initial investigation (Appendix A.2.2) with the training dataset, where we maximized the recognition performance.

The role of SVM-C in the first stage is to estimate the posterior probability where the input logo candidate clipped by the extracted candidate region belongs each logo category. When we have  $n$  categories to classify, SVM-C outputs  $n + 1$  posterior probability values, where  $n$  logo categories and one “non-logo” category.

The SVM-V in the second stage verifies the input logo candidate region using the posterior probability values estimated by SVM-C in the first stage. When a logo candidate region accurately captures a shoe logo, the posterior probability value of the logo class corresponding to

the captured logo is expected to become large while others are low. In contrast, if the candidate region does not contain the logo or the class to classify the logo is confusing, the values of the posterior probability is expected to show a unique distribution. The task of SVM-V is to input  $n + 1$  posterior probability values as a feature vector and to classify the vector into logo or non-logo classes.

SVM-C and SVM-V employ radial basis function kernels and are trained as multiple-class and two-class classifiers, respectively. The training schemes of the SVMs are described in the Chapter 4.2.



## Chapter 4

# Evaluation Experiments

This chapter describes evaluation experiments for the proposed method. Chapter 4.1 describes the dataset used in the experiments, and training of classifiers is described in Chapter 4.2. Chapter 4.3 details evaluation. Chapter 4.4 describes results and discussion.

### 4.1 Dataset

We used the IEICE pattern recognition and media understanding (IEICE-PRMU) shoe logo dataset [5]. To the best of our knowledge, IEICE-PRMU shoe logo dataset is one of the most difficult dataset which consists of real-world shoe images. The dataset consists of 661 images and ground-truth annotations are given for each image. The logos of eight brands with a wide variety of appearances are contained in the dataset. Figure 4.1 shows the appearance variations of the shoe logos in the dataset. Some images were not captured under controlled conditions and the images contain blur, rotation, occlusion, and perspective distortion. Samples of the eight brands in the dataset are shown in Figure 4.2. Four brands of ASICS, FILA, MIZUNO and New Balance have two types of appearance such as figure 4.2. ASICS(1) and ASICS(2) are designed in the motif of character “A” and “a”, respectively. FILA(1) and FILA(2) are designed by character “FILA” and “F”, respectively. MIZUNO(2) is designed by combination of MIZUNO(1) graphic symbol and character “MIZUNO”. New Balance(1) and New Balance(2) are designed in the motif of character “NB” and “N”, respectively.

We employed a three-fold cross-validation for performance evaluation. The dataset was divided into three groups at random. Two were used for training and the remaining one

was used as a test set. We conducted this evaluation ten times and calculated the mean performance.



Fig. 4.1: Samples from the IEICE pattern recognition and media understanding (IEICE-PRMU) shoe logo dataset [5].



Fig. 4.2: Examples of shoe logo of eight brands contained in the IEICE pattern recognition and media understanding (IEICE-PRMU) shoe logo dataset. The brands are ASICS, FILA, Le Coq Sportif, Syunsoku, MIZUNO, New Balance, Under Armour, and YONEX. The brands of ASICS, FILA, MIZUNO and New Balance have two types.

## 4.2 Training of classifiers

As described above, we employ a two-stage method using two SVMs, and the objective of each is different. Each SVM is trained using independent training data.

SVM-C in the first stage is trained with logo images extracted using the ground-truth anno-

tation. The training SVM-C dataset contains 11 logo classes. The dataset originally contains eight brands, but three of these, ASICS, MIZUNO and New Balance, have two types of appearance. Thus, we divide each of these into two separate classes. FILA(1) and FILA(2) are classified as same class, because both types aim to recognize only character “F”. The images belonging to the “non-logo” class is extracted from detected regions that do not overlap with annotated logo regions.

For training SVM-V in the second stage, we collect the dataset containing posterior value vectors of logo and non-logo regions. Applying the algorithm for logo candidate region extraction described in Chapter 3.1 to training images, we obtain several candidate regions. For logo verification, in which the extracted candidate region is identified as containing or not containing a logo, negative samples are necessary for SVM training. The negative samples were generated from extracted regions that do not overlap the annotated logo regions.

### 4.3 Evaluation

We evaluated the overall performance of the method using recall  $R$ , precision  $P$ , and  $F$ -measure  $F$ . These are defined by the following:

$$R = \frac{1}{N_{\text{GT}}} \sum_{i=1}^{N_{\text{GT}}} \delta(S_c^{(i)}, S_o^{(i)}) O(S_c^{(i)}, S_o^{(i)}), \quad (4.1)$$

$$P = \frac{1}{N_{\text{DET}}} \sum_{j=1}^{N_{\text{DET}}} \delta(S_c^{(j)}, S_o^{(j)}) O(S_c^{(j)}, S_o^{(j)}), \quad (4.2)$$

$$F = \frac{2PR}{P + R}. \quad (4.3)$$

We also evaluated the detection performance of the method using average best overlap (ABO), calculated by

$$ABO = \frac{1}{N_{\text{GT}}} \sum_{k=1}^{N_{\text{GT}}} O(S_c^{(k)}, S_o^{(k)}), \quad (4.4)$$

where the overlap rate between regions  $S_c$  and  $S_o$  (shown in Figure 4.3) is determined by

$$O(S_c, S_o) = \frac{A(S_c \cap S_o)}{A(S_c) + A(S_o) - A(S_c \cap S_o)} \times 100. \quad (4.5)$$

For the above calculation,  $N_{GT}$  and  $N_{DET}$  denote the number of ground-truth regions and detected regions, respectively. Furthermore,  $S_o^{(i)}$  in the recall calculation in (4.1) is the detected region with the minimum distance from the  $i$ -th ground-truth  $S_c^{(i)}$ . Similarly,  $S_c^{(j)}$  is the selected ground-truth region with respect to the detected regions. Regions  $S_c^{(k)}$  and  $S_o^{(k)}$  in the overlap ratio calculation are determined as the highest overlap regions, and  $\delta(S_c^{(i)}, S_o^{(i)})$  denotes

$$\delta(S_c, S_o) \begin{cases} 1 & (\text{brands of } S_c \text{ and } S_o \text{ are same}), \\ 0 & (\text{otherwise}). \end{cases} \quad (4.6)$$

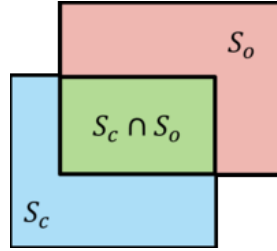


Fig. 4.3: Calculation of overlap.  $S_c$ ,  $S_o$ ,  $S_c \cup S_o$  mean regions of ground-truth, detected and overlapped regions, respectively.

After  $P$ ,  $R$  and  $F$  are calculated for each test image, we calculate the evaluation value by averaging over all test images.

To compare the extraction and recognition performances of the proposed method with those of other techniques, we implemented R-CNN and adopted the same conditions as used for the proposed method. CNN features are computed by forward propagating a mean-subtracted  $50 \times 50$  RGB image through two convolutional layers and two fully-connected layers. We then obtain a 400-dimensional feature vector using the CNN.

## 4.4 Results and Discussion

In this chapter, we evaluate the quality of our proposed method. We divide our results in two part, (1) comparison of logo extraction performance and (2) logo extraction and recognition performances.

### 4.4.1 Comparison of logo extraction performance

Table 4.1 compares the results of logo extraction performance. Each row in the table denotes the method employed for logo extraction. The naive MSER method listed in the first row is adopted for preprocessing the grayscale image. The second and third rows show the extraction performance of the proposed method. Although the initial region extraction using *HSV* color images outperforms the naive MSER, iterative integration of the extracted initial regions further improves the region extraction performance. The fourth row shows the ABO obtained by a selective search based method proposed by Uijlings [18]. These results show that the MSER algorithm extracts logos more efficiently when it is applied to color-plane images. The ABO is increased by introducing the proposed iterative region integration method. This suggests that logos containing multiple connected components are successfully reconstructed by the region integration approach.

Figure 4.4 shows ABO values for each brand of shoe logos obtained by the proposed method. It is observed that the ABO value of FILA(1) is lower than that of FILA(2). As shown figure 4.5, ground-truth annotation for FILA(1) is given as bounding is only character “F”. Therefore, iterative region integration integrate a black component of character “F” and a character “T”. For this reason, the ABO value of FILA(1) is lower than FILA(2) which is no character around “F”. Also, it is observed that the ABO value of Syunsoku is also lower than that of other logos. The reason for this is that the logos of Syunsoku consist of multiple components and appear relatively small in most images.

Table. 4.1: Comparison of logo extraction performance

method	ABO (%)
naive MSER	54.67
initial region extraction	68.23
proposed (2.1)	77.04
Uijlings et al.[18]	44.46

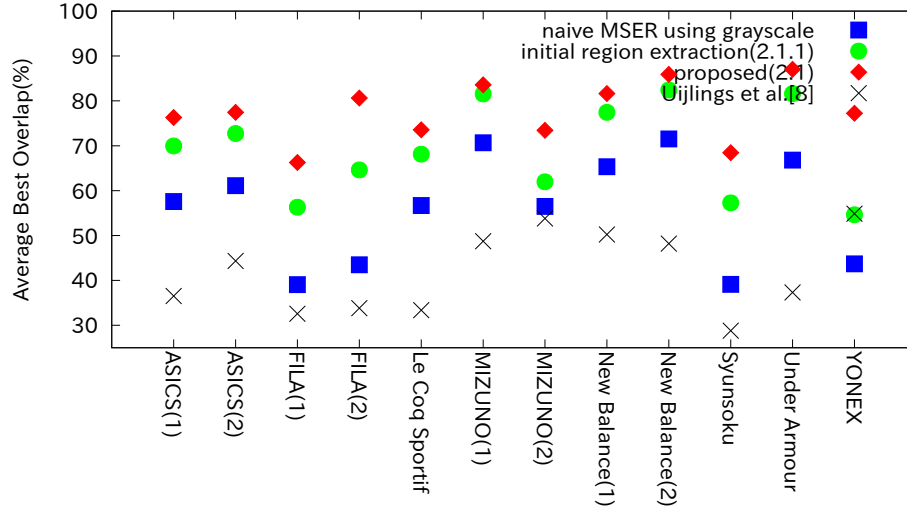


Fig. 4.4: ABO for each brand of shoe logos



Fig. 4.5: The example of FILA(1) and ground-truth annotation

#### 4.4.2 Logo extraction and recognition performances

Figure 4.6 and 4.7 show examples of shoe logos extracted by the proposed method. In Figure 4.6, the red boxes denote shoe logos which are correctly detected and recognized. Although the logos in the images are a wide variety of sizes and rotations, the proposed

method successfully recognizes these logos. A number of false-positive regions obtained by the region-extraction stage were successfully eliminated by introducing a negative class into the classifier. In Figure 4.7, the green boxes denote shoe logos which the proposed method could not detect properly. The most common reason of error is small color difference between logo and background due to shoe and logo design ((a) and (b)). When the color difference is quite small, MSER may fail to extract these logo regions in the initial region extraction stage. While the proposed method is robust against affine transformation of logo, weakness for perspective and nonrigid distortion is still remaining (c). A few cases where the proposed method could not detect a large logo region fragmented by an object like shoelace are observed (d). Although the majority false-positive regions were eliminated successfully, the recognition method may misrecognize the background clutter to the logo ((e) and (f)).

Tables 4.2 and 4.3 compare the extraction and recognition performance among the proposed method and other methods. We conducted a three-fold cross validation ten times and show the mean, standard deviation, maximum, and minimum values of each criterion in the tables. This method outperforms R-CNN. These results suggest that the proposed approach is more effective when the training dataset is small compared with the appearance variety.

Table. 4.2: Quantitative evaluation of the extraction and recognition performance of the proposed method

criterion	mean	std	max	min
Recall	30.32	0.73	31.65	29.39
Precision	59.53	1.39	60.80	58.14
F-measure	40.17	0.68	40.99	39.26

Table. 4.3: Quantitative evaluation of the extraction and recognition performance of R-CNN

criterion	mean	std	max	min
Recall	7.25	0.79	8.21	5.45
Precision	46.61	3.44	51.19	40.81
F-measure	12.54	1.29	14.08	9.64





(a) F-measure: 94.1065



(b) F-measure: 91.6602



(c) F-measure: 87.3386



(d) F-measure: 84.7152



(e) F-measure: 89.322



(f) F-measure: 70.6134



(g) F-measure: 97.6718



(h) F-measure: 90.8243

Fig. 4.6: Examples of successfully extracted and recognized logos.





Fig. 4.7: Examples of failure extracted and recognized logos.

## Chapter 5

# Conclusion

This chapter describes this research conclusion and feature works.

### 5.1 Conclusion

In this paper, we proposed an approach combining shoe logo extraction and recognition. Our approach achieves an  $F$ -measure of 40.17 %. Shoe logo extraction employing MSERs and hierarchical region integration works effectively for logos consisting of different colors and sizes. For diversification strategy, the combination of  $HSV$  color space and three similarity measures (color, distance and size) has a particularly good score. Although the initial region extraction using three color-plane images outperforms the naive MSER, iterative integration of the extracted initial regions further improves the region extraction performance. Also, the MSER algorithm extracts logos more efficiently when a selective search based method proposed by Uijlings [18] is applied to color-plane images. Logo recognition using gradient histogram features and an SVM works for both recognition and false-positive elimination of logos. The proposed two-stage approach is effective for improving performance of both extraction and recognition.

### 5.2 Feature works

Although shoe logo extraction is employed MSERs, we have to investigate a robust extraction method in order to extract a variety of shoe logos. Further study and investigation of the

performance of the proposed method on a larger dataset are required in future research.

## Appendix A

# The overview of employing methods

This appendix described the method which proposed methods are employed. Appendix A.1 described the outline of Maximally Stable Extremal Regions (MSERs). Appendix A.2 describes the outline of the grayscale gradient histogram features. The outline of Support Vector Machine (SVM) is described in Appendix A.3.

### A.1 Maximally Stable Extremal Region(MSER) method

The MSER method is a region segmentation based on pixel values in a grayscale image. The MSER algorithm extracts from an image a number of co-variant regions, called MSERs: an MSER is a stable connected component of some gray-level sets of the image. The MSER method is based on the idea of taking regions which stay nearly the same through a wide range of thresholds. The MSER method process is performed as follows.

#### A.1.1 Extraction of MSERs

The pixel  $D(\subset integer\mathbb{Z})$  of image  $I$  is a mapping the intensity value  $S = \{0, 1, \dots, 255\}$ .

$$I : D \rightarrow S \tag{A.1}$$

Extremal regions are well defined on images if:

1.  $S$  is totally ordered.
2. An adjacency relation  $A \subset D \times D$  is defined. If  $p$  and  $q \in D$  are an adjacency relation,  $pAq$  is expressed.

Region  $Q$  is a contiguous subset of  $D$ .

Outer Region Boundary  $\partial Q$  of  $Q$  is the set of pixels adjacent to at least one pixel of  $Q$  but not belonging to  $Q$ . Outer Region Boundary  $\partial Q$  is determined by

$$\partial Q = \{q \in D \setminus Q : \exists p \in Q : pAq\} \quad (\text{A.2})$$

Figure A.1(a) shows Region  $Q$  and Outer Region Boundary  $\partial Q$ . Region  $Q$  means connected pixels (black or white) in binary image. Binary images are generated by thresholding a gray level image at all possible thresholds starting from 0 and ending at 255. A series of Region  $Q$  is segmented to use binary images as shown Figure A.1(b).

Extremal Region  $Q \subset D$  is a region such that either for all  $p \in Q, q \in \partial Q : I(p) \geq I(q)$  (maximally intensity region) or all  $p \in Q, q \in \partial Q : I(p) \leq I(q)$  (minimum intensity region).

Let  $Q_1, \dots, Q_{i-1}, Q_i, \dots$  be a sequence of nested extremal regions, i.e. ( $Q_i \subset Q_{i+1}$ ). Extremal region  $Q_{i^*}$  is maximally stable if  $q(i)$  has a local minimum at  $i^*$ . A equation  $q(i)$  is determined by

$$q(i) = \frac{|Q_{i+\Delta} \setminus Q_{i-\Delta}|}{|Q_i|}, Q_i \subset Q_{i+1} \quad (\text{A.3})$$

where  $\Delta \in S$  is a parameter of the method. The equation checks for regions that remain stable over a certain number of thresholds.

Extremal regions have three important properties :

1. The approach is invariance to affine transformation of image.
2. The approach is robust monotonic transformation of image intensities.
3. Multi-scale detection without any smoothing involved, both fine and large structure is detected.

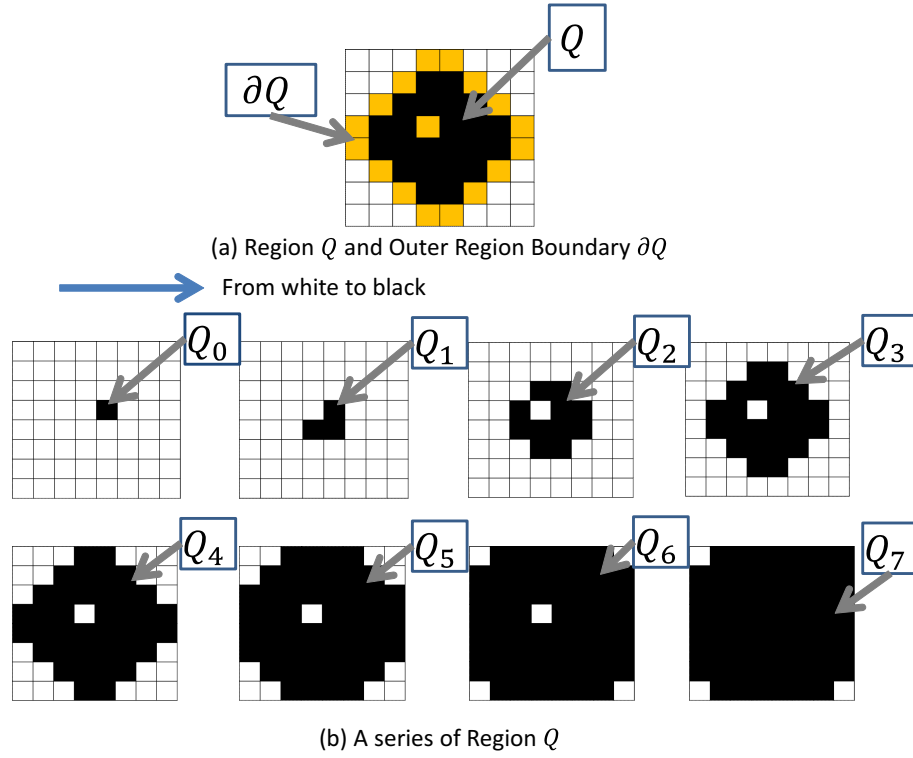


Fig. A.1: Region  $Q$ , Outer Region Boundary  $\partial Q$  and a series of Region  $Q$ . The square denotes a pixel of image. The black square and orange square denote Region  $Q$  and Outer Region Boundary  $\partial Q$ , respectively.

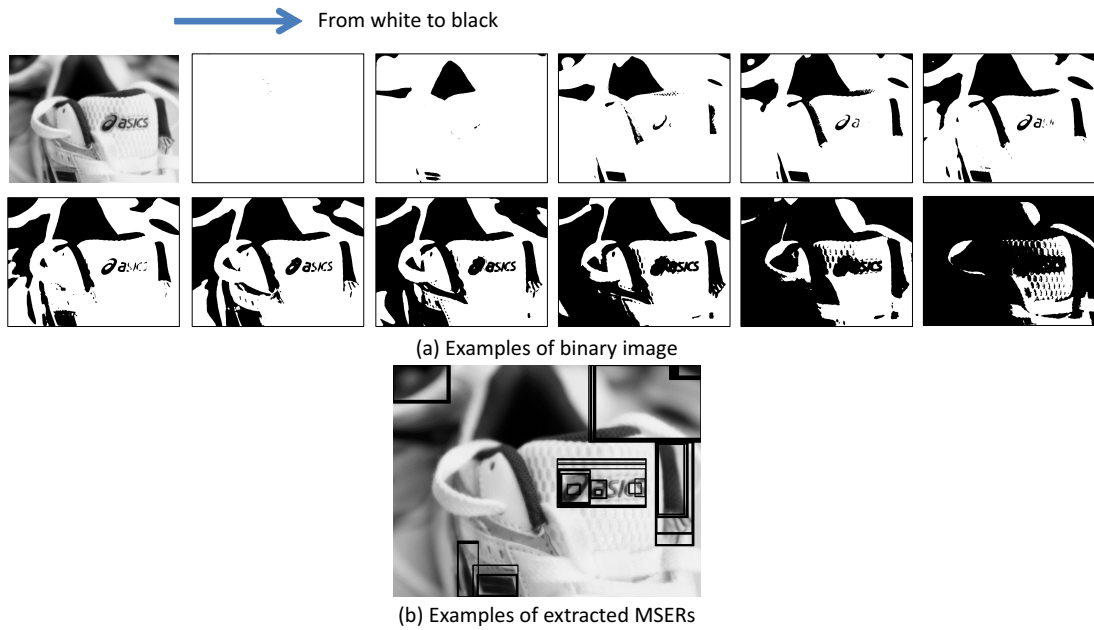


Fig. A.2: Examples of binary images and extracted MSERs

The MSER method is expected to extract shoe logos with a wide variety of appearance. Figure A.2 shows examples of binary images and extracted MSERs. Connected regions which grows slowly are extracted as MSERs.

## A.2 The grayscale gradient histogram features

The grayscale gradient histogram features, which is composed of a directional histogram of the gradient of the grayscale image, is extracted from the input image. The grayscale gradient histogram features are well-known to achieve high score in object recognition with unique shape. In the recent, histograms of oriented gradients (HOG) feature is proposed with the method using the gradient. In the present paper, the grayscale gradient histogram features are extracted by grayscale logo candidate.

### A.2.1 Extraction of Gradient histogram features

The feature extraction process is performed as follows:

1. A  $5 \times 5$  gaussian filter is applied to the input image  $I = I(x, y)$ .

$$I'(x, y) = \frac{\sum_{i=-\frac{N}{2}}^{\frac{N}{2}} \sum_{j=-\frac{N}{2}}^{\frac{N}{2}} f(i, j) I(x + i, y + j)}{\sum_{i=-\frac{N}{2}}^{\frac{N}{2}} \sum_{j=-\frac{N}{2}}^{\frac{N}{2}} f(i, j)} \quad (\text{A.4})$$

$$f(i, j) = \exp\left(-\frac{i^2 + j^2}{2\sigma^2}\right) \quad (\text{A.5})$$

where  $\sigma$  denotes kernel weighting factor, and set 3.0.

2. Sobel operators are used to obtain a gradient image. The arc tangent of the gradient (i.e., the direction of the gradient) is initially quantized into  $L$  directions and the strength of the gradient is accumulated for each quantized direction. The magnitude of the gradient  $g(x, y)$  is defined as follows:

$$g(x, y) = \sqrt{(\Delta u)^2 + (\Delta v)^2} , \quad (\text{A.6})$$

and the direction of the gradient  $\theta(x, y)$  is

$$\theta(x, y) = \tan^{-1} \left\{ \frac{\delta u}{\delta v} \right\} \quad (\text{A.7})$$

where

$$\Delta u = g(x + 1, y + 1) - g(x, y) \quad (\text{A.8})$$

and

$$\Delta v = g(x + 1, y) - g(x, y + 1) \quad (\text{A.9})$$

and  $g(x, y)$  is the gray level of the  $(x, y)$  pixel.

3. Histograms for the values of  $L$  quantized directions are computed in each block.
4. The normalized image is initially segmented into  $x$  (width)  $\times$   $y$  (height) blocks. Because of a compromise between accuracy and complexity, this block size is determined empirically.
5.  $n \times m$  blocks are downsampled into  $\frac{n+1}{2} \times \frac{m+1}{2}$  blocks using Gaussian filters.
6. A directional histogram consisting of  $L$  directions is downsampled into  $\frac{L}{2}$  directions using the weighting filter([1 4 6 4 1]). Finally, a  $\frac{(n+1)(m+1)L}{8}$  dimensional feature vector is obtained.

Figure A.3 shows the gradient image (Step 4). The direction and the magnitude are represented by the hue and the brightness, respectively, in Figure A.3.

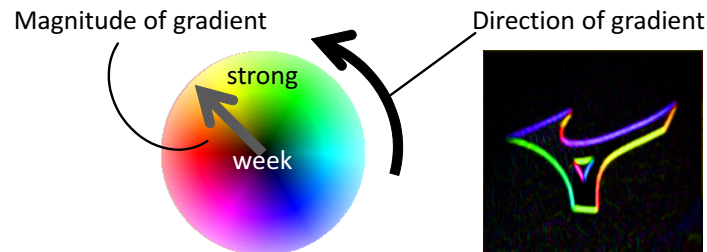


Fig. A.3: The gradient image



### A.2.2 The dimensionality of extracted gradient feature vector

We investigate the dimensionality of feature vector with the training datasets, where we maximized the recognition performance. Figure A.4 shows the result of our initial investigation. Here, we evaluated the classification performance of SVM by three-fold cross validation over the training dataset. From the results, it is confirmed that the classification accuracy is influenced by the dimensionality of feature vectors. Since the dimensionality of feature vector is calculated such as chapter A.2.1, we tested several combination of parameters for classification accuracy. The best accuracy is obtained when the dimensionality of feature vector is 400. Therefore, we use  $n = m = 9, L = 32$  of the grayscale gradient histogram features in the chapter A.2.1.

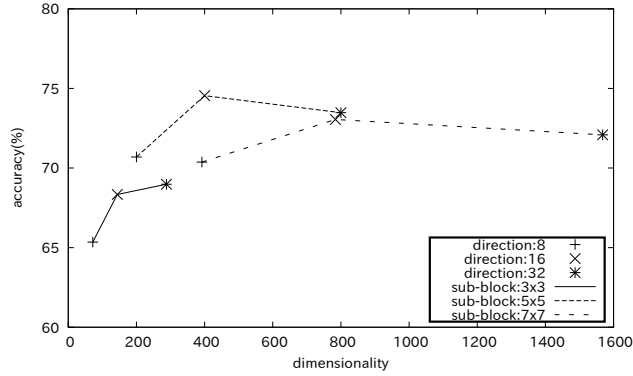


Fig. A.4: Result of initial investigation in logo classification: classification accuracy (%) vs dimensionality of gradient features. The dimensionality of a feature vector is determined by the product of the number of sub-blocks and quantized directions.

## A.3 SVM (Support Vector Machine)

An support vector machine (SVM) is supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [23]. An SVM constructs a hyperplane or set of hyperplanes in a high- or infinite-dimensional space, where hyperplane has the largest distance to the nearest training-data point of any class.

We are given a training dataset of  $n$  points the from

$$(\vec{x}_1, y_1), \dots, (\vec{x}_n, y_n) \quad (\text{A.10})$$

where the  $y_i$  are either 1 or -1, each indicating the class to which the point  $\vec{x}_i$  belongs. Each  $\vec{x}_i$  is a  $p$ -dimensional real vector.

A discriminant function of SVM is defined by the following:

$$g(\vec{x}) = \sum_{i=1}^n w_i x_i + b \quad (\text{A.11})$$

where  $\vec{w}$  and  $b$  denote the parameter of weight vector and the parameter of bias term, respectively.

Figure A.5 shows an example of an SVM trained with samples from two classes. An SVM find the “maximum-margin hyperplane” that divides the group of points  $\vec{x}$  for which  $g(\vec{x}) = 1$  from the group of points for which  $g(\vec{x}) = -1$ , which is defined so that the distance between the hyperplane and the nearest point  $\vec{x}_i$  from either group is maximized. Hyperplane is expressed when  $g(\vec{x}) = 0$ . In the SVM-C, Multiple SVM is used because this research classification multi-class.

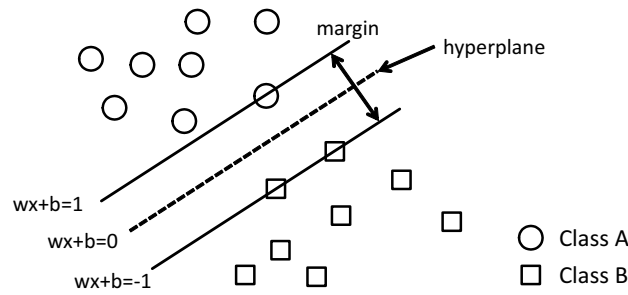


Fig. A.5: The example of an SVM trained with samples from two classes

### A.3.1 Kernel trick

There is a nonlinear classification in which distribution areas of classes of each other overlap. The kernel trick avoids the mapping that is needed to get linear learning algorithms in order to learn a nonlinear function or decision boundary. The kernel trick allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space by transforming a space into high dimensional feature space.

### A.3.2 An SVM scaling

The feature vector convert linearly before the image inputs to SVM. The method convert in range -1 to 1 linearly.

A test set is converted linearly to use the parameter which training set is converted linearly.

## AppendixB

# Individual strategies for extraction method

In the present paper, we propose two diversification strategies to obtain good quality proposed method: varying the color space and varying the similarity measures. AppendixB investigates the influence of each strategy.

### B.1 A strategy which varies color spaces

We examine the variations in the color space as setting we experiment in condition of only initial region extraction. Table B.1 shows comparison of color spaces, ranging from an ABO of 54.67 with 112 locations for  $I$  color space to an ABO of 68.23 with 378 for  $HSV$  color space. A  $HSV$  color space has a particularly good ABO score of 68.23 using 378 boxes. These results that initial region extraction method extracts logos efficiently using  $HSV$  color space than other color spaces. Moreover,  $HSV$  color space is found to be more similar to the way the human eyes perceive color [25].

### B.2 A strategy which combines similarity measures

We examine the combination of similarity measures as setting we use  $HSV$  color space with good ABO score. Table B.2 shows the combination of similarity measures.  $C$ ,  $S$  and  $D$

Table. B.1: Comparison of color spaces

color space	ABO (%)	# box
<i>HSV</i>	68.23	378
<i>Lab</i>	67.90	374
<i>YCbCr</i>	67.80	315
<i>RGB</i>	63.60	263
<i>I</i>	54.67	112

mean the (C)olor measure, (S)ize measure and (D)istance measure in the table, respectively. The only  $S$  similarity measure overdetect candidate regions because this similarity measure encourages small regions to integrate early. For the combination of other similarity measures, other similarity measures suppress over detection of  $S$  measure. An ABO of  $C + D + S$  similarity measure is higher than an ABO of  $D + S$  measure in spite of extracted regions of  $C + D + S$  and  $D + S$  similarity measures are same. The  $C + D + S$  similarity measure has a particularly good ABO score of 77.04 using 645 boxes. These results that our extraction method extracts logos efficiently using  $C + D + S$  similarity measure. From the resulting ordering we create extraction of shoe logos method using combination of  $HSV$  color space and  $C + D + S$  similarity measure.

Table. B.2: The combination of similarity measures

Similarities	ABO (%)	# box
$C$	74.09	578
$D$	74.19	593
$S$	72.43	740
$C + D$	74.94	603
$C + S$	74.64	685
$D + S$	76.90	645
$C + D + S$	77.04	645

## AppendixC

# Programs file location

All this research's program locate following directory in the human interface laboratory's server.

- /net/xserve0/users/aoki

The structure and resume of this research are shown in the following:

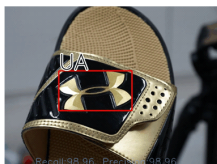
```
aoki/research/
|
|-detection
|   # Programs which extract logos using MSER and selective search
|-c_validation
|   |   # Programs which conduct a three fold cross validation
|   |-r-cnn
|       # comparison method (regions with convolution neural networks)
|-img
    # IEICE-PRMU shoe logo dataset which is used in this research
```

# AppendixD

## Briefing paper

We used following paper in the master thesis presentation.

### Extraction and Recognition of Shoe Logos with a Wide Variety of Appearance using Two-Stage Classifiers



三重大学大学院  
工学研究科情報工学専攻  
ヒューマンインターフェース研究室  
416M501 青木 一憲

1

### 研究背景

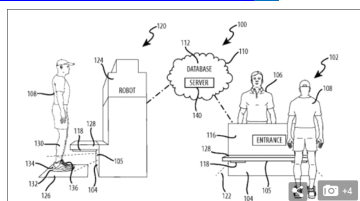
- ロゴマーク
    - 製造企業を明確化する.
    - 市場において人々の注目を集める.
    - ブランドイメージを的確に表現する.
  - 着用者、使用者を特徴付ける.
- ↓ 画像解析・認識と組み合わせて
- 人物のトラッキング、人物識別に利用可能



2

### 例えば...

- [ディズニーが来園者の「靴」をスキャンして移動先を把握するシステムの特許を取得\[1\]](http://www.daily-mail.co.uk/sciencetech/article-3709713/Now-Mickey-track-Manolags-Disney-patents-track-Magic-Kingdom-visitors-footwear.html)



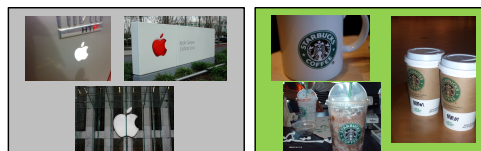
Using cameras and sensors, the technology would scan people's footwear upon arrival and follow them as they move throughout the grounds to see what rides, shows or stores piqued their interest. Visitors would first have their foot scanned at the entrance in order for the system to gather their information.

[1] <http://www.daily-mail.co.uk/sciencetech/article-3709713/Now-Mickey-track-Manolags-Disney-patents-track-Magic-Kingdom-visitors-footwear.html>

3

### ロゴマーク認識の課題

- 現実世界で対象物を撮影すると対象物の外観のアフィン変換や非剛体変換が発生する.



4

## 関連研究

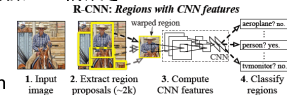
- “Exemplar-based logo and trademark recognition” [1] (exemplarベース, N. Farajzadeh, 2010)
  - 合成画像が増えるにつれ認識精度が向上
  - 認識精度と誤検出にトレードオフの関係がある欠点.
- “Logo recognition and localization in real-world images by using visual patterns” [2] (Modelベース, W. Chu, 2012)
  - 実行時間が高速
  - Recall = 19.0%, Precision = 30.0%
- “Bundle min-hashing” [3][4] (特徴記述子ベース, S. Romberg, 2013)
  - Data augmentation により認識率向上

[1] N. Farajzadeh, “Exemplar-based logo and trademark recognition,” *Pattern Recognition Letters*, vol. 31, no. 24, pp. 2485–2490, 2010.  
 [2] W. Chu, T. Li, and L. Wang, “Logo recognition and localization in real-world images by using visual patterns,” *Pattern Recognition*, vol. 45, no. 10, pp. 2485–2490, 2012.  
 [3] S. Romberg, “Bundle min-hashing,” *Pattern Recognition Letters*, vol. 34, no. 24, pp. 2485–2490, 2013.  
 [4] S. Romberg, “Bundle min-hashing,” *Pattern Recognition Letters*, vol. 34, no. 24, pp. 2485–2490, 2013.

## ディープラーニング

- Girshickら[4]: “Regions with Convolutional Neural Network (R-CNN)” 2012.

– Selective Searchを用いた抽出と CNNを用いた認識とで構成される.



- Selective Search
  - 初期領域を抽出し、反復領域統合する.
- PASCAL VOC 2010において高いスコアを得る.
- 学習に大規模なデータセットが必要.

[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, 2014.

6

## 靴ロゴマークはさらに難しい



7

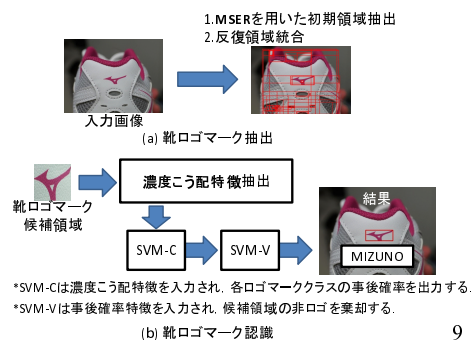
## 本研究の内容

- クラス内変動の大きい靴ロゴマークの抽出・認識手法
- 靴ロゴマーク抽出
  - MSERを用いた初期領域抽出と反復領域統合
  - [15th IAPR International Conference on Machine Vision Applications \(MVA 2017\)](#)【査読有】にて発表
- 靴ロゴマーク認識
  - 2段階分類器を用いた認識
  - [IEICE Transactions on Information and Systems](#)【採録決定・2018年5月刊行予定】

[5] K. Shino and W. Ohyama, “IEICE-PRMU shoe logo dataset,” <https://sites.google.com/ieice/ieice-prmu-shoe-logo-dataset>

8

## 提案手法の概要



9

## IEICE-PRMU shoe logo dataset

- IEICE pattern recognition and media understanding (IEICE-PRMU) shoe logo dataset[5]
  - 8クラス、935個のロゴマーク

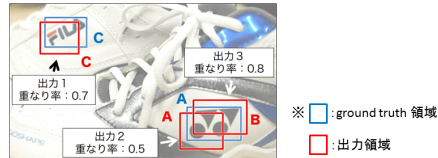


[5] K. Shino and W. Ohyama, “IEICE-PRMU shoe logo dataset,” <https://sites.google.com/ieice/ieice-prmu-shoe-logo-dataset>

10



## 評価方法



抽出・認識の評価

$$\text{再現率} = \frac{1 \times 70(\text{重なり率}0.7) + 0 \times 80(\text{重なり率}0.8)}{2(\text{ground truth 領域数})} = 35$$

$$\text{適合率} = \frac{1 \times 70(\text{重なり率}0.7) + 1 \times 50(\text{重なり率}0.5) + 0 \times 80(\text{重なり率}0.8)}{3(\text{出力領域数})} = 40$$

抽出のみの評価

$$\text{Average Best Overlap(ABO)} = \frac{70(\text{重なり率}0.7) + 80(\text{重なり率}0.8)}{2(\text{ground truth 領域数})} = 75$$

## 実験結果

抽出のみ評価

抽出と認識の評価			手法		ABO(%)
手法	再現率	適合率			
提案手法	30.32	59.53	naive MSER(グレー)	54.67	
R-CNN[3]	7.25	46.62	初期領域抽出(HSV)	68.23	
			提案手法	77.04	
			Selective Search[2]	44.46	

提案手法の抽出・認識性能はR-CNNを上回る。

見え方の多様性に対して訓練データセットが小さい実験において提案手法はより効果的である。

※データセットを3分割交差検証で10回実験。

12

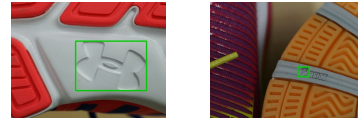
## 抽出・認識結果の例



13

## 失敗画像例

- ロゴマークと背景との色の差が小さい



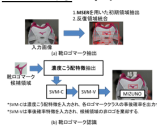
- 提案手法はアフィン変換には頑健だが、非剛体変形に対する認識精度が低い。

※  : ground truth 領域  : 出力領域

14

## まとめ

- クラス内変動の大きい靴ロゴマークの抽出・認識手法
- 靴ロゴマーク抽出
  - MSERを用いた初期領域抽出と反復領域統合
  - Selective searchのABO値に比べて32.58%上回った。
- 靴ロゴマーク認識
  - 2段階分類器を用いた認識
  - R-CNNの再現率に比べ23.07%, 適合率において12.91%上回った。
- 見え方の多様性に対して訓練データセットが小さい場合に提案手法はより効果的である。



15

# Acknowledgements

Profound thanks goto Professor Tetsushi Wakabayashi whose become kind and advice to this research, and Assistant Professor Wataru Ohyama whose comments and suggestions for this research. Without Assistant Professor Wataru Ohyama, I could not accomplish anything. Special thanks goto Honorary Professor Yasuji Miyake whose participate in the discussion while he are busy and advice valuable. I woud like to thank Miyuki Tanaka and Sachiko Nakatsuka of clerks, seniors of our laboratory whose advice this research and student of the same period whose improve themselves by competing with each other. Finally, I would also like to express my gratitude to my family for their moral support and warm encouragements.

# Reference

- [1] A. Joly and O. Buisson: “Logo retrieval with a contrario visual query expansion.” *ACM International Conference on Multimedia*, pp.581-584, 2009.
- [2] Y. Kalantidis, L. G. Pueyo, M. Trevisiol, R. van Zwol, and Y. Avrithis: “Scalable triangulation-based logo recognition.” *ACM International Conference on Multimedia Retrieval*, pp.20, 2011.
- [3] S. C. Hoi, X. Wu, H. Liu, Y. Wu, H. Wang, H. Xue, and Q. Wu: “Logo-net: Large-scale deep logo detection and brand recognition with deep region-based convolutional networks” *arXiv preprint arXiv:1511.02462*, 2015.
- [4] J. Matas, O. Chum, and T. Pajdla: “Robust wide baseline stereo from maximally stable extremal regions.” *British Machine Vision Conference*, pp.384-396, 2002.
- [5] K. Shirai and W. Ohyama: IEICE-PRMU shoe logo dataset:  
<https://sites.google.com/site/alcon2015prmu/prmu-shoelogo-dataset>
- [6] D. Deguchi, Y. Kameda, I. Kitahara, K. Kondo, A. Shimada and S. Hiura: “Looking back on past PRMU Algorithm Contests” *IEICE PRMU2012-45*, vol.112, no.197, pp.143-147, 2012.
- [7] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V.R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida and E. Valveny: “ICDAR 2015 competition on Robust Reading,” *2015 13th ICDAR*, pp. 1156-1160, 2015.
- [8] N. Farajzadeh.: “Exemplar-based logo and trademark recognition,” *Machine Vision and Applications*, vol.26, Issue 6, pp.791-805, 2010.
- [9] W. Chu, T. Lin.: “Logo recognition and localization in real-world images by using visual patterns” *IEEE International Conference on Acoustic, Speech and Signal Processing*, pp.973-976, 2012.
- [10] C. H. Lampert, M. B. Blaschko and T. Hofmann.: “Beyond sliding windows: Object localization by efficient subwindow search” *2008 IEEE Conference on Computer Vision and Pattern Recognition*, , pp.1-8, 2008.
- [11] D.G. Lowe.: “Object recognition from local scale-invariant features” *International Conference Computer Vision*, pp.1150-1157, 1999.
- [12] M. Datar, N. Immorlica, P. Indyk, and V. Mirroknu.: “Locality-sensitive hashing scheme based on P-stable distributions” *Annual Symposium on Computational Geometry*, pp.253-262, 2004.
- [13] K. Fukunaga and L. D. Hostetler.: “The estimation of the gradient of a density function, with applications in pattern recognition” *IEEE Transactions on information Theory*, Vol. 21, No.1, pp.32-40, 1975.
- [14] S. Romberg and L. Rainer: “Bundle min-hashing” *International Journal of Multimedia Information Retrieval*, vol.2, No.4, pp. 243-259, 2013

- [15] S. Romberg and L. Rainer: "Bundle min-hashing for logo recognition" *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pp.113-120, 2013
- [16] X. Shen, Z. Lin, J. Brandt and Y. Wu: "Spatially-constrained similarity measure for large-scale object retrieval" *IEEE transactions on pattern analysis and machine intelligence*, Vol.36, No.6, pp.1229-1241, 2014
- [17] R. Girshick, J. Donahue, T. Darrell, and J. Malik.: "Rich feature hierarchies for accurate object detection and semantic segmentation" *IEEE conference on Computer Vision and Pattern Recognition*, pp. 580-587, 2014.
- [18] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders.: "Selective search for object recognition" *International Journal of Computer Vision*, Vol. 104, No. 2, pp. 154-171, 2013.
- [19] P. Felzenwalb, D. Huttenlocher.: "Efficient Graph-Based Image Segmentation" *International Journal of Computer Vision*, Vol. 59, No. 2, pp. 167-181, 2004.
- [20] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) Challenge *International Journal of Computer Vision*, Vol. 88, No.2, pp. 303-338, 2010.
- [21] H. Su, X. Zhu and S. Gong: "Deep Learning Logo Detection with Data Expansion by Synthesising Context" *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 530-539, 2017
- [22] R. Girshick: "Fast R-CNN" *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 1440-1448, 2015.
- [23] K. Tsuda: "Overview of Support Vector Machine" *The journal of the IEICE*, 83.6, pp.460-466, 2000.
- [24] M. Shi, Y. Fujisawa, T. Wakabayashi, and F. Kimura: "Handwritten numeral recognition using gradient and curvature of gray scale image" *Pattern Recognition*, vol.35, Issue 10, pp.2051-2059, 2002.
- [25] V. Jumb, M. Sohani, A. Shrivastava: "Color Image Segmentation Using K-Means Clustering and Otsu's Adaptive Thresholding" *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, ISSN.2278-3075, vol.3, Issue 9, pp.72-76, 2014.