15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

Automatic Extraction and Recognition of Shoe Logos with a Wide Variety of Appearance

Kazunori Aoki, Wataru Ohyama and Tetsushi Wakabayashi
Graduate School of Engineering, Mie University
Tsu-shi, Mie, Japan
{aoki, ohyama}@hi.info.mie-u.ac.jp

Abstract

A logo is a symbolic presentation that is designed not only to identify a product manufacturer but also to attract the attention of shoppers. Shoe logos are a challenging subject for automatic extraction and recognition using image analysis techniques because they have characteristics that distinguish them from those of other products, that is, there is much variation in the appearance of shoe logos. In this paper, we propose an automatic extraction and recognition method for shoe logos with a wide variety of appearance using a limited number training samples. The proposed method employs maximally stable extremal regions (MSERs) for the initial region extraction, an iterative algorithm for region grouping, and gradient features and a support vector machine for logo recognition. The results of performance evaluation experiments using a logo dataset that consists of a wide variety of appearance show that the proposed method achieves promising performance for both logo extraction and recognition.

1 Introduction

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

The 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. Figure 1 shows examples of shoe logos captured by standard still cameras. Figures 1(a) and (b), which belong to the same company, have the same shape but different colors. Figures 1(c) and (d) are examples of logos consisting of multiple components. Figures 1(e) and (f) show the most common appearance variations of shoe-logo images, i.e., rotation, occlusion, and perspective distortion. While the automatic extraction and recognition technique must handle these problems properly, because the shoes themselves are usually worn on feet, the extraction and recognition of shoe logos are expected to contribute to person identification and tracking.

The extraction and recognition of logos in images is a topic that has attracted the attention of researchers. Several studies on the automatic extraction of logos have been reported.

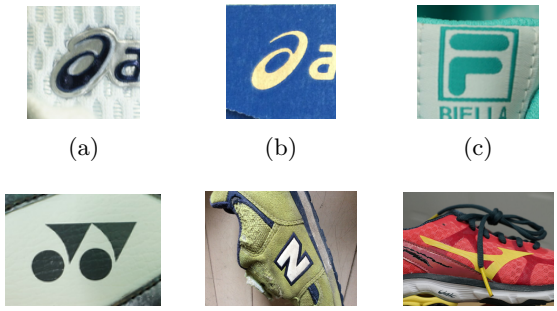
Farajzadeh [1] proposed an exemplar-based method for logo or trademark recognition. This approach extracts logos using new samples that are synthesized from logo images with different tilts and rotations and recognizes them using a linear support vector machine (SVM). This technique has the disadvantage in that there is a significant tradeoff between recognition accuracy and false detection. When the number of synthesized samples are increased to improve recognition accuracy, the number of false detections drastically increases. Chu et al. [2] proposed a method using visual patterns. This approach first extracts scale-invariant feature transform (SIFT) features [3] from both test images and a logo image. Features with high similarity in both the test images and logo image are found using locality sensitive hashing [4]. The main purpose of their method is to improve computational efficiency by eliminating outliers in a test image obtained from an exhaustive sliding windows search. However, their method obtains this high computational efficiency at the sacrifice of extraction accuracy. They reported that the method obtains only 19.0% recall and 30.0% precision.

In the field of generic object detection and recognition from images, deep neural network architectures have been employed because of their promising performance and high adaptivity. Girshick et al. [5] proposed a method called regions with convolutional neural network (R-CNN) for the generic object recognition problem. R-CNN is an approach that combines selective search [6] and CNN. The method extracts initial regions using efficient graph-based image segmentation [7] and iteratively groups regions using similarity calculated from appearance features in the regions. While R-CNN achieved high performance score on PASCAL2010 [8], the method requires a large-scale, accurately annotated dataset for training, which is quite difficult for shoe logos, which have a wide variety of appearance.

In this paper, we propose an automatic extraction and recognition method for shoe logos with a wide variety of appearance using a limited number of training samples. The proposed method employs maximally stable extremal regions (MSERs) [10] for the initial region extraction, an iterative algorithm for region grouping and gradient features, and an SVM for logo recognition.

2 Proposed method

Figure 2 shows the outline of the proposed method. The proposed method inputs one still color (RGB) im-



(a) (b) (c) (d) (e) (f)

Figure 1: Examples of shoe logos[11]: (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.

age and 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 the shoe logos and recognition of the extracted logo regions. The following sections describe each stage in detail.

2.1 Extraction of shoe logos

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

2.1.1 Initial region extraction

The initial region extraction for shoe logos employs MSERs [10] for each color plane of the input RGB image. The MSER method is a region segmentation based on pixel values in a grayscale image. To employ MSER sufficiently for the input shoe logo image, preprocessing consisting of image smoothing and histogram equalization is performed.

First, we perform image smoothing to reduce noise. Because it is necessary to preserve region edges for extracting sharp-shape regions such as logos, we employ a bilateral filter to take advantage of its noise reduction and edge preserving properties. Our early trials and investigation suggested that image smoothing using median or Gaussian filters is not suitable for segmenting regions for logo extraction because these filters remove edges as well as noise. The parameters for the bilateral filter are adjusted so that they extract the smallest logo in the reference image.

Second, the grayscale histogram equalization reduces the effects of illumination variation.

Finally, the smoothed and equalized image is separated into three color-plane images based on the RGB value of the pixels and the MSER segmentation algorithm is applied to each color-plane image. This color-plane separation enables the extraction of shoe logos with multiple colors. Regions extracted from the three color-plane images are then concatenated to construct the initial candidate regions. The parameters for the MSER algorithm were also determined by preliminary investigation using reference images such that the smallest logo is extracted correctly.

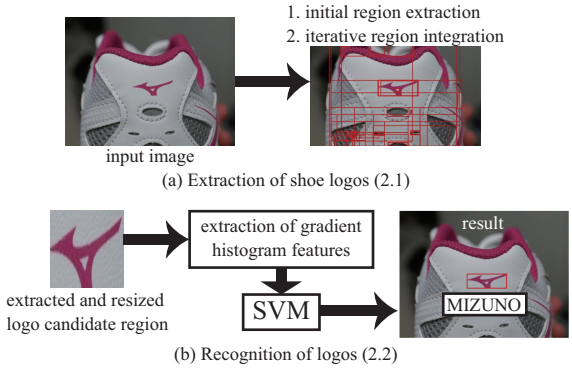


Figure 2: Overview of the proposed method

2.1.2 Iterative region integration

The above extraction process basically extracts single connected components (CCs) as initial regions. Because some logos contain multiple CCs, we perform region integration using histogram intersection to extract these multiple CCs as one integrated candidate region. The region integration works iteratively as follows.

The input and output of the region integration are a set of extracted initial regions $R^0 = \{r_1, \cdots, r_n\}$ and a set of integrated regions R^C, respectively.

1. Initialize R^C as R^0:

$$R^C = R^0. \tag{1}$$

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 histogram intersection $H(r_i, r_j)$:

$$(r_i, r_j) = \arg\max H(r_i, r_j). \tag{2}$$

The histogram intersection $H(r_i, r_j)$ is calculated by

$$H(r_i, r_j) = \sum_{l=0}^{L-1} \min(h_l^{(i)}, h_l^{(j)}), \tag{3}$$

where $h_l^{(i)}$ and $h_l^{(j)}$ denote the l-th value of histograms obtained in regions r_i and r_j, respectively.

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 is less than one.

5. Region R^C is the final integrated region result.

We obtain the color histogram in (3) from each RGB plane with 16 bins. For example, when regions are detected in the R-plane image, we obtain a color histogram using the R value.

2.2 Recognition of logos

The extracted candidate regions are evaluated and classified in the recognition stage. We employ the grayscale gradient histogram features [9] and an SVM in the recognition stage.

We first clip the input image using the detected logo candidate regions and convert them to a fixed size (50×50 pixels). The gradient histogram features are extracted from the grayscale of the original image using the process in [9] with 5×5 spatial sub-blocks and 16 quantization directions. The dimensionality of extracted feature vectors is 400. These parameters are determined by our initial investigation with the training dataset, where we maximized the recognition performance.

An SVM with a radial basis function kernel is trained as multiple-class classifier and used for logo detection and classification. The training scheme of SVM is described in the following section.

3 Evaluation Experiments

We conducted experiments to evaluate the effectiveness of the proposed method.

3.1 Dataset

We used the Pattern Recognition and Media Understanding (PRMU) shoe logo dataset [11]. 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 appearance are contained in the dataset. Samples of the eight classes in the dataset are shown in Figure 3. As shown in the figure, some images were captured under uncontrolled conditions, and the images contain blur, rotation, occlusion, and perspective distortion.

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.

For logo verification, in which the extracted candidate region is identified as containing or not containing an actual logo, negative samples are necessary for SVM training. The negative samples were generated from detected regions that do not overlap the annotated logo regions.

3.2 Evaluation

We evaluated the overall performance of the method using recall R, precision P, and F-measure F, are defined, respectively, by

$$R = \frac{1}{N_{\mathrm{GT}}} \sum_{i=1}^{N_{\mathrm{GT}}} \delta(S_c^{(i)}, S_o^{(i)}) O(S_c^{(i)}, S_o^{(i)}), \quad (4)$$

$$P = \frac{1}{N_{\mathrm{DET}}} \sum_{j=1}^{N_{\mathrm{DET}}} \delta(S_c^{(j)}, S_o^{(j)}) O(S_c^{(j)}, S_o^{(j)}), \quad (5)$$

$$F = \frac{2PR}{P+R}. \quad (6)$$

Figure 3: Examples of shoe logo of eight brands contained in PRMU shoe logo dataset. The brand names consist of ASICS, FILA, Le Coq Sportif, Syunsoku , Mizuno, New Balance, Under Armour, and Yonex.

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

$$ABO = \frac{1}{N_{\mathrm{GT}}} \sum_{k=1}^{N_{\mathrm{GT}}} O(S_c^{(k)}, S_o^{(k)}), \quad (7)$$

where the overlap rate between regions S_c and S_o is determined by

$$O(S_c, S_o) = \frac{A(S_c \cap S_o)}{A(S_c) + A(S_o) - A(S_c \cap S_o)} \times 100. \quad (8)$$

For the above calculation, N_{GT} and N_{DET} denote the number of ground truth regions and detected regions, respectively. Further, $S_o^{(i)}$ in the recall calculation in (4) 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

$$\delta(S_c, S_o) \begin{cases} 1 \text{ (class labels of } S_c \text{ and } S_o \text{ are same)}, \\ 0 \text{ (otherwise)}. \end{cases} \quad (9)$$

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

To compare 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 50×50 RGB image through two convolutional layers and two fully-connected layers. We then obtain a 400-dimensional feature vector using the CNN.

4 Results and Discussion

Table 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 three color-plane images outperforms the naive MSER,

(a) (b) (c) (d)

Figure 4: Examples of extracted and recognized logos

Table 1: Comparison of logo extraction performance

method	ABO (%)
naive MSER	49.36
initial region extraction (2.1.1)	59.26
proposed (2.1)	64.83
Uijlings et al.[6]	44.46

Table 2: Quantitative evaluation of the extraction and recognition performance of the proposed method

criterion	mean	std	max	min
Recall	19.10	0.53	20.30	18.25
Precision	64.11	1.74	66.51	60.28
F-measure	29.42	0.68	30.89	28.51

Table 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

an 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 [6]. 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 suggest that logos containing multiple CCs are successfully reconstructed by the region integration approach.

Figure 4 shows examples of shoe logos extracted by the proposed method. While these logos contain a wide variety of sizes and rotations, the proposed method successfully recognizes these logos. Note that a number of false-positive regions obtained by the region extraction stage are also successfully eliminated by introducing a negative class into classifier.

Tables 2 and 3 show a quantitative evaluation and comparison of the extraction and recognition performance. We conducted a three-fold cross validation ten times and show the mean, standard deviation, maximum, and minimum value of each criterion in the tables. These results indicate that the proposed method outperforms R-CNN when the training dataset is small.

5 Conclusion

In this paper, we proposed an approach combining shoe logo extraction and recognition. Our approach achieves an F-measure of 29.10 %. Shoe logo extraction employing MSERs works effectively for logos consisting of different colors and sizes. Logo-mark recognition using gradient histogram features and an SVM work for both the recognition and false-positive elimination of logos.

Further study and investigation of the performance of the proposed method on larger dataset is a remaining research topic.

References

[1] N. Farajzadeh.: "Exemplar-based logo and trademark recognition," *Machine Vision and Applications*, vol.26, Issue 6, pp.791-805, 2010.

[2] 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.

[3] D.G. Lowe.: "Object recognition from local scale-invariant features" *International Conference Computer Vision*, pp.1150-1157, 1999.

[4] M. Datar, N. Immorlica, P. Indyk, and V. Mirroknu.: "Locality-sennsitive hashing scheme based on P-stable distributions" *Annual Symposium on Computational Geometry*, pp.253-262, 2004.

[5] 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] 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.

[7] P. Felzenwalb, D. Huttenlocher.: "Efficient Graph-Based Image Segmentation" *International Journal of Computer Vision*, Vol. 59, No. 2, pp. 167-181, 2004.

[8] 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.

[9] 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.

[10] J. Matas, O. Chum, and T. Pajdla: "Robust wide baseline stereo from maximally stable extremal regions." *British Machine Vision Conference*, pp.384-396, 2002.

[11] IEICE-PRMU shoe logo dataset: `https://sites.google.com/site/alcon2015prmu/prmu-shoelogo-dataset`