2011 International Conference on Document Analysis and Recognition

A Keypoint-Based Approach Toward Scenery Character Detection

Seiichi Uchida, Yuki Shigeyoshi, Yasuhiro Kunishige, and Feng Yaokai
Kyushu University, Fukuoka, 819-0395, Japan

***Abstract*—This paper proposes a new approach toward scenery character detection. This is a keypoint-based approach where local features and a saliency map are fully utilized. Local features, such as SIFT and SURF, have been commonly used for computer vision and object pattern recognition problems; however, they have been rarely employed in character recognition and detection problems. Local feature, however, is similar to directional features, which have been employed in character recognition applications. In addition, local feature can detect corners and thus it is suitable for detecting characters, which are generally comprised of many corners. For evaluating the performance of the local feature, an experimental result was done and its results showed that SURF, i.e., a simple gradient feature, can detect about 70% of characters in scenery images. Then the saliency map was employed as an additional feature to the local feature. This trial is based on the expectation that scenery characters are generally printed to be salient and thus higher salient area will have a higher probability to be a character area. An experimental result showed that this expectation was reasonable and we can have better discrimination accuracy with the saliency map.**

***Keywords*-character localization, camera-based character recognition, scenery image, local feature, saliency map**

I. Introduction

As widely known, character detection in natural scene images is one of the most difficult problems for computers. For this challenging problem, several promising trials have been made [1]–[4] and even some commercial services, such as Evernote[1], Google Goggles[2], and Word Lens[3], are available nowadays. We, however, can say that there is still room for improvement. In fact, Evernote and Google Goggles employ character detection (and recognition) for image retrieval and thus they can allow many false detections. Word Lens can detect and recognize characters under regulated conditions.

The contribution of this paper is to propose a new approach toward the scenery character detection problem. The main idea is the utilization of *local feature* and *visual saliency*. Although both of those techniques have been utilized in computer vision problems, they have rarely utilized in the scenery character detection problem. Since our problem can be considered as a kind of computer vision problems in real environment, we can naively expect the usefulness of those techniques. Of course, in more theoretical viewpoints, we can expect their usefulness, as will be emphasized throughout this paper.

"Local feature" is generally comprised of two functions; detection and description of *keypoints*. SIFT and SURF [5] are typical local features. Although local features have not utilized for character detection and recognition (except for a very limited number of trials [6]–[9]), the properties of local features seem to be useful for the scenery character detection problem. First, local feature often can detects keypoints around corners. Since every character generally contains many corners, local feature will not miss character regions. Second, the described feature is similar to quantized directional features, which have been utilized in OCR. This indicates that local feature can represent character shapes sufficiently. Third, local feature is robust against deformations, such as partial occlusion, rotation, and scaling. Fourth, there is a possibility to realize "segmentation-free" character detection by using the set of local features detected as a (part of) character. In fact, the proposed detection technique employs neither binarization process nor other segmentation processes.

"Visual saliency map" [10] is a technique to convert a bitmap image to a grayscale image whose intensity value is relative to the saliency of the pixel. Simply speaking, the saliency is defined as the degree of differences between the target pixel and its surroundings in brightness, edge direction, and color. The property of the saliency will to be useful for the scenery character detection problem. This is because most characters in natural scene images are prepared for showing some message and therefore should be appealing, i.e., salient to humans. This indicates that the saliency can be used as a prior for character detection; higher saliency becomes, higher the probability of existing some character becomes.

In the proposed method, a keypoint-based discrimination between character and non-character will be done. That is, (i) keypoints are detected as character candidates, then (ii) the local area around each keypoint is described as a feature vector, and finally (iii) the discrimination is done by a classifier trained by AdaBoost. The visual saliency is incorporated into the feature vector as an element or used in another approach. In the following sections, those procedures will be detailed.

Note that the proposed method is based on scattered keypoints and thus provides a distribution of character candidate keypoints on the scenery image. In other words, it does not provide the exact location of individual characters. We, however, will see that this rough distribution will be a substantial clue to have a final character detection result. The more significant contribution of this paper is to show that just a simple and local 128-dimensional directional feature has a potential of discriminating characters from

[1] http://www.evernote.com
[2] http://www.google.com/mobile/goggles
[3] http://questvisual.com/

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.168

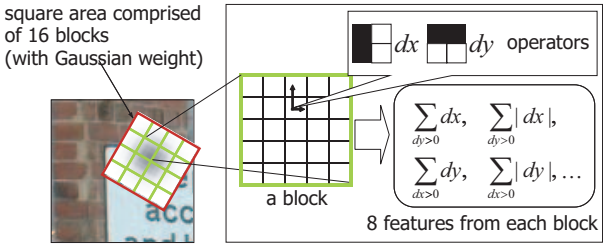


Figure 1. Local feature description by SURF. [5]

clutter background and its performance is enhanced by incorporating visual saliency.

II. Keypoint-based discrimination between character and non-character

A. Detection and description of keypoint by SURF

In this paper, we use SURF [5] as local feature. In SURF, keypoints are first detected as local maxima in a scale-space of approximated Hessian filter response. Then, a rotated square region is determined around each keypoint and described as a 128-dimensional feature vector. Since the rotation angle and the size of the region are determined adaptively and automatically, the resulting vector becomes invariant to not only rotation but also scale.

The elements of the feature vector described by SURF represent local x-y gradients within the square region. Specifically, as shown in Fig. 1, the square region is divided into $4 \times 4 = 16$ blocks and at each block, eight gradient features ($\Sigma_{y \geq 0} dx$, $\Sigma_{y<0} dx$, $\Sigma_{y \geq 0} |dx|$, $\Sigma_{y<0} |dx|$, $\Sigma_{x \geq 0} dy$, $\Sigma_{x<0} dy$, $\Sigma_{x \geq 0} |dy|$, and $\Sigma_{x<0} |dy|$) are calculated.

As noted in Section I, local feature has suitable properties for scenery character detection. The keypoints are detected at the pixels with a larger Hessian value and thus can capture the corners of various characters. Figure 2 (a)-(d) show an scenery image, the character region in the image, the detected keypoints, and the keypoints detected on the character region. Many keypoints were detected successfully around every character region. Accordingly, if we make a successful discrimination later, we can grasp the accurate and dense distribution of all characters on the scenery images.

Again, it is important to note that the SURF feature vector is very similar to quantized directional features, such as weighted direction code histogram [11], which has been utilized in character recognition. This fact also supports that SURF is suitable for scenery character detection; this is because the fact proves that SURF as well as the quantized directional features can represent the character shape in an appropriate manner[4].

[4]It is well-known that the human visual perception system captures quantized local directions in its early step, called primary visual cortex or V1. Since both of detection and recognition is, of course, done through the system, it seems reasonable that directional feature is generally effective for detection as well as recognition.

B. Discrimination by AdaBoost

At each keypoint, discrimination between character and non-character is done by a classifier trained by AdaBoost. The classifier is a weighted combination of K "weak learners". Each weak learner makes the discrimination by a simple thresholding operation on the value of one selected from the 128 SURF feature vector elements. The element selection is automatically done; this fact indicates that, from the trained classifier, we can know which local gradients are important for the character detection problem among 128 SURF feature vector elements. In the following experiment, K was fixed at 256; this means that important elements are selected several times.

III. Saliency Map as a prior

A. Visual Saliency

The saliency map [10] simulates psychological saliency in human visual perception by evaluating the difference between the target pixel and its surroundings in brightness, edge direction, and color. Figure 2(e) shows the saliency maps for the scenery images (a).

The saliency map is useful as a good prior for the character detection problem. As noted in Section I, we can expect that higher saliency becomes, higher the character existence probability becomes. This expectation comes from the assumption that most scenery characters are designed to be read by humans. Signboards are good examples to validate the assumption. Another example is the fact that humans never write black characters on a black background — to read characters, we need some contrast, i.e., saliency.

Figure 2 (f) shows another version of the saliency map where color saliency is not evaluated. Practically, this version might be more effective than the original version because there are many black (white) characters on white (black) backgrounds.

B. Three Methods of Utilizing Saliency

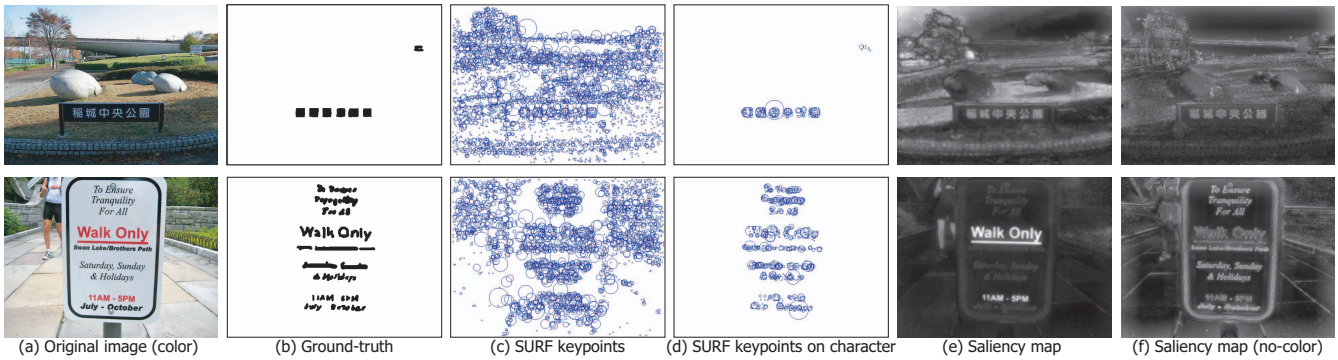
We can consider the following three methods of combining saliency with SURF feature vector.

- SURF+Saliency1: SURF feature vector+ the value of the saliency map at the keypoint, i.e., the center of the square region describing the keypoint. (129-dim.)
- SURF+Saliency2: SURF feature vector+ the mean value of the saliency map within the square region. (129-dim.)
- SURF+Saliency3: SURF feature vector+ + another SURF feature vector described by using the saliency map like a bitmap image at the same square region. (256-dim.)

IV. Experimental result

A. Dataset

A scenery image dataset were prepared for our experiments. Using Google Image Search™, top 300 photo images



(a) Original image (color) (b) Ground-truth (c) SURF keypoints (d) SURF keypoints on character (e) Saliency map (f) Saliency map (no-color)

Figure 2. Example of SURF local feature and visual saliency map.

(each of which contains some characters and has around 640×480) were first collected. The keywords used in the search were "park" and "sign." Those 300 images were then decomposed into a training dataset (150 images) and a test dataset (150 images). For each image, a ground-truth (i.e., character and non-character labels) is attached at each pixel manually. Note that small characters have ambiguous boundary and thus their ground-truth became inevitably rough (like a bounding box).

B. Quantitative Evaluation

The accuracies of discrimination between character and non-character were listed in Table I. Here, the accuracy for the character class was calculated as the ratio between the number of all the keypoints from the character region and the number of the keypoints correctly classified into the character class. The accuracy for the non-character class was calculated in the similar way[5].

The facts shown by Table I are summarized as follows.

- Surprisingly, 67% discrimination accuracy was achieved at the character region by the SURF feature vector, which is just a simple local gradient feature. Considering the fact that many SURF keypoints are extracted in the character region (Fig. 2(d)), this accuracy shows that the proposed method can provide a reasonable distribution of character candidate points, which will be useful for the final task, that is, detection of individual or consecutive characters.
- Similarly, the proposed method could achieve about 75% accuracy in non-character regition; it is also an unexpectedly high accuracy if we consider huge variation of the non-character region.
- Saliency was clearly effective for achieving better discrimination. It is also shown that, for our problem, the saliency without color was better than the original saliency; this is expected because there are many black or white characters even in scene.
- Among the three methods of utilizing saliency (Section III-B), "SURF+Saliency 3" could achieve the best performance in both of character and non-character regions. Especially, the discrimination accuracy in the character region was improved largely (67%→74%). The contrast of the saliency map was large around the character region and SURF could capture this contrast.
- An important contribution is that this result proves quantitatively that scenery characters are often salient. Figure 4 shows the distribution of saliency values on keypoints. This distribution proves that keypoints in the character region have higher saliency values (especially, non-color saliency values) than those in the non-character region.

It is noteworthy that if we can expect the situation that the above accuracy is uniform over the entire image, we can easily convert the result into more accurate one. Specifically, if we change the discrimination result of each keypoint by taking "majority voting" among its neighboring keypoints, we can remove "outliers" and then have a smoother and more reliable result.

C. Qualitative Evaluation

Figure 3 shows several discrimination results. In Fig. 3 (a)-(e), many correct discrimination results (red circle) are found *densely* around character regions. These results also verify that the proposed method is promising for realization of scenery character detection.

A closer inspection will show the effect of the saliency. For example, the keypoints discriminated wrongly as the characters around the upper-left part of Fig. 3 (e) were correctly discriminated by using the saliency. In fact, the saliency of this part (several pillars in front of grass) was low and thus gives a lower character existence probability.

Figure 3 (f) and (g) are examples with lower discrimination accuracies. In (f), pillars of the building were wrongly discriminated as characters. As discussed later, the trained classifier discriminates characters from non-characters by

[5]The evaluation by *recall* and *precision*, which were often employed in many detection problems, were difficult for our case. This is because the number of the detected keypoints is different from the number of characters in the image and therefore it was difficult to evaluate the exact number of false negatives and false positives.

Figure 3. Discrimination results. (Better viewed in color.) A thick circle (red) and a thin circle (green) are keypoints discriminated into character and non-character classes, respectively.

Table I
ACCURACIES OF DISCRIMINATION BETWEEN CHARACTER AND NON-CHARACTER (%).

feature	char	nonchar
SURF	66.89	74.85
SURF+Saliency 1	67.20	74.10
(no-color)	72.79	75.58
SURF+Saliency 2	66.71	75.17
(no-color)	67.42	74.39
SURF+Saliency 3	71.20	75.12
(no-color)	74.16	77.79

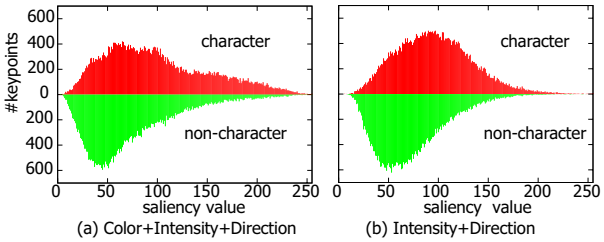


Figure 4. Distribution of saliency values on keypoints.

using linear structures around corners. This result indicates the limitation of the discrimination by using the simple gradient feature. For eliminating those errors, a higher-level information will be necessary. In (g), characters were too small and thus SURF could not describe their features sufficiently.

D. SURF Feature Elements Selected

Figure 5 shows which SURF feature elements were selected by the top 50 weak learners by AdaBoost training. Since the accuracy by the resulting classifier was almost saturated by the 50 weak learners, it is possible to say that the selected elements can describe the essential shapes of scenery characters. In the figure, a circle indicates that at least one weak classifier has selected the corresponding feature element. More weak learners have selected the same element, larger a circle becomes. Note that for simplicity, a pair of elements, such as $\Sigma_{y\geq 0}dx$ and $\Sigma_{y>0}dx$, are not distinguished in this figure.

From Fig. 5, it is proved that the element was selected not randomly but according to a specific tendency. Specifically, many weak learners have selected the $\sum|dx|$ elements of the upper middle and lower middle areas. This fact indicates that the horizontal changes (i.e., $\sum|dx|$) in those areas are important clues for the discrimination. (Note that since the direction of a SURF keypoint is determined adaptively and thus the word "horizontal" does not mean some absolute direction but means a relative direction to the dominant gradient direction.) Recalling the fact that SURF keypoints are often detected around corners, this result indicates that the classifier discriminates a keypoint into the character class if it exists around a corner and has a specific change around its upper middle and lower middle areas.

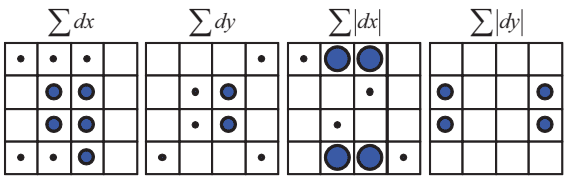


Figure 5. SURF feature elements selected by AdaBoost. Note that SURF feature elements are derived at each of 16 blocks.

V. CONCLUSION

The most important contribution of this paper is that it was experimentally proved that just a very simple local gradient feature given by SURF has a potential of not only detecting character regions densely but also discriminating the character region from the non-character region with around 70% accuracy. Considering the complexity of characters and non-characters (i.e., all the scenery components other than characters), it is possible to say that this is a reasonably high accuracy. It has also been proved that this accuracy is improved to around 75% by using visual saliency. This improvement also proved an interesting fact quantitatively that characters are often salient in scene.

Toward for our final goal, that is, detection of individual or consecutive characters, keypoints discriminated as the character class should be gathered to form character regions. A simple clustering may do it, if the current discrimination accuracy is improved by local majority voting.

ACKNOWLEDGMENT

This research was partially supported by JST, CREST.

REFERENCES

[1] Xiangrong Chen and Alan L. Yuille, "Detecting and Reading Text in Natural Scenes," CVPR, 2004.

[2] L. Xu, H. Nagayoshi, and H. Sako, "Kanji Character Detection from Complex Real Scene Images Based on Character Properties," DAS, 2008.

[3] H. Goto, "Redefining the DCT-Based Feature for Scene Text Detection," IJDAR, 2008.

[4] M. Iwamura, T. Tsuji, and K. Kise, "Memory-Based Recognition of Camera-Captured Characters," DAS, 2010.

[5] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," ECCV, 2006 .

[6] M. Diem and R. Sablatnig, "Recognition of Degraded Handwritten Characters Using Local Features," ICDAR, 2009.

[7] M. Diem and R. Sablatnig, "Are Characters Objects?," ICFHR, 2010.

[8] S. Uchida and M. Liwicki, "Part-Based Recognition of Handwritten Characters," ICFHR, 2010.

[9] P. Sankar, C. V. Jawahar, and R. Manmatha, "Nearest Neighbor Based Collection OCR," DAS, 2010.

[10] L. Itti, C. Koch, E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis," PAMI, 20(11), 1998.

[11] F. Kimura, T. Wakabayashi, S. Tsuruoka, and Y. Miyake, "Improvement of Handwritten Japanese Character Recognition Using Weighted Direction Code Histogram," Pattern Recognition, 30(8), 1997.