important region estimation using image captioning

Taku Suzuki1, Daikuke Sato1, Yoshihiro Sugaya2, (Member, IEEE), Tomo Miyazaki1, (Member, IEEE), and Shinichiro Omachi1, (Senior Member, IEEE)

1Graduate School of Engineering, Tohoku University, Sendai 980-8579, Japan
2Faculty of Advanced Science and Technology, Ritsumeikan University, Otsu 520-2194, Japan
Corresponding author: Shinichiro Omachi (mach@ecei.tohoku.ac.jp)

This work was supported in part by the JSPS KAKENHI under Grant JP20H04201; and in part by the Yotta Informatics Project by MEXT, Japan.

ABSTRACT When storing images and videos on a limited storage device or transmitting them over a narrow-band network, an effective approach is to detect the necessary parts and process them preferentially. Visual saliency has often been used for this purpose, and many methods have been proposed to detect salient objects. However, a salient object is not necessarily the primary subject in an image. Determining the important regions in an image is not clear or easy to achieve because it generally depends on the context of the image. In this study, we propose a novel framework for detecting important image regions. We leverage an image-captioning technique because it interprets the context of an image when generating sentences. The proposed method determines those important regions that are closer to the level of human sensitivity by exploiting semantic information from the image captioning. To evaluate the effectiveness of the proposed method, we created a dataset that defines important regions within images based on experiments using subjective evaluation. Applying this dataset, we confirmed that the accuracy of the proposed approach was higher than that of conventional saliency-based object detection methods.

INDEX TERMS Important region estimation, image captioning, visual saliency, machine learning.

I. INTRODUCTION Images and videos are a means of information transmission that can be easily understood and are informative for humans. However, with the large numbers of daily generated images and videos, it is no longer practical to store all available images and videos in their original format, thereby requiring selection and reduction. It has been predicted that the total number of data captured, created, and replicated annually worldwide will reach 175 trillion gigabytes by 2025 [1], and according to statistics compiled by Cisco, video on the Internet accounts for over 80% of all traffic [2]. In addition, the increase in storage devices has not caught up with the increase in data, and only a small number of the data generated can be saved, a gap that continues to increase [3]. Under such a situation, it is extremely important to decide which data to store and for how long [4]. It is also necessary to select and reduce the number of data when using a narrow-band network or when communication is concentrated owing to a natural disaster, thereby creating large-scale congestion. In the case of a natural disaster, it is important to efficiently transmit images and videos to accurately convey the situation. Fortunately, in the case of images and videos, it is possible to leave only some of the data. For example, a video can be summarized by detecting key frames [5]. Using the region of interest (ROI) encoding introduced in JPEG2000 [6], images can be stored or transmitted efficiently by encoding the necessary parts with high definition and the remaining parts at a high compression rate [7], [8]. Visual saliency [9] has often been used for this purpose, and many different methods have been studied for saliency map estimation. However, a salient object is not always the primary subject of an image. In other words, there may be objects that, despite being visually salient, are not important to the user. In general, it may be difficult to determine which regions are important in an image because such
importance depends on the context of the image. It is useful to detect important regions and estimate their importance as a preprocessing step for intelligent image processing. In this study, we propose a novel framework for estimating important regions in an image. We leverage an image-captioning technique because it interprets the context of an image when generating sentences. The generated caption should contain important information regarding the image. The proposed method determines important regions by exploiting semantic information from the image-captioning module.

The main contributions of this study are summarized as follows.

- We propose a novel framework for estimating the important regions in an image. This is a challenging task because estimating important regions requires both the visual features and the context of the image to be considered. We propose a method that leverages image captioning techniques to take the context into account. To the best of the authors’ knowledge, this is the first attempt at estimating important regions in images through image captioning.
- We propose a method for precisely matching the words in the caption with the objects in the image by introducing cyclical training into the image captioning and considering the probability of the preceding word.
- We constructed datasets for an evaluation and parameter estimation that defines the importance of the objects in an image through subjective evaluation experiments and conducted an evaluation experiment on the proposed method. It was qualitatively and quantitatively confirmed that the accuracy of the proposed method was higher than that of saliency-based methods.

The remainder of this paper is organized as follows. First, we introduce related studies in Section II. The proposed method is described in detail in Section III. We describe the datasets used for the experiments in Section IV. In Section V, we present and discuss the experimental results. Finally, Section VI provides some concluding remarks regarding this research.

II. RELATED WORK

A. SALIENT OBJECT DETECTION

To the best of our knowledge, no studies have been explicitly aimed at estimating important image regions. As a similar research topic, we therefore introduce methods for salient object detection. Visual saliency is an index that quantifies the ease of gazing based on human visual characteristics. An image representing the estimated saliency for each pixel is called a saliency map. There are various applications for this purpose, one of which is salient object detection.

The method developed by Itti et al. [9] is one of the earliest and most well-known models for computing a saliency map. It calculates features such as the brightness, hue, and edge directional components obtained from the image over different scales and generates a saliency map by combining them. Derived from this method, many methods have been proposed for detecting salient objects using features obtained from an image. Zhang et al. proposed a Boolean map-based saliency model (BMS) [10] that detects salient objects by extracting a connected component from the binary image obtained by randomly thresholding each color channel of the image and averaging the results. Shen et al. [11] took an approach in which, in addition to traditional low-level features such as color information, high-level features are provided as prior knowledge indicating that salient objects tend to be located at the center of an image and that humans tend to pay attention to faces.

Along with its developed technology, methods using machine learning have also been proposed. Hou et al. [12], [13] generated saliency maps from multi-level and multi-scale features obtained through a convolutional neural network (CNN) using a network that adds short connections to the holistic nested edge detector (HED) [14], which is used to detect edges. Islam et al. [15] adopted an encoder/decoder model in which the first half is used for feature extraction during down-sampling, and the second half is used for feature integration during up-sampling. To improve the accuracy of the decoder, the map is refined in a step-by-step manner. Zhao and Wu [16] extracted high-level contextual features and low-level spatial structural features, and then combined them into a pyramid feature attention network to effectively utilize multiscale features. Pang et al. [17] dealt with multiscale features more effectively by applying an aggregate interaction module (AIM), which combines features from neighboring levels, and a self-interaction module (SIM), which makes effective use of intralayer features obtained through an AIM. Chen et al. [18] proposed a network structure that integrates salient features for object detection.

Methods focusing on edges and boundaries have recently been proposed for accurately capturing the object shapes. Wei et al. [19] proposed a method for dividing binary labels into two types with continuous values, body and detail, to improve the accuracy, particularly around the edges. Zhao et al. [20] focused on the complementarity between salient edges and object information and proposed a network for modeling this complementary information. Han et al. [21] introduced an edge constraint term into the loss function to preserve the edge information of an object. Qin et al. [22] proposed a network integrating predicted and refined modules and a new loss function for boundary-aware salient object detection. Using guided filters for boundary preservation, Lad et al. [23] proposed a method for object detection based on the use of wavelet-based saliency maps.

All of these methods detect salient regions using image features. By contrast, the proposed method is novel in that it uses not only image features for detecting important regions, but also captions that express the context of the entire image.

B. IMAGE CaptionING

Image captioning generates short sentences that describe a given image. The accuracy of image captioning has improved
with the development of machine learning and has been actively studied in recent years. In 2015, Vinyals et al. [24] proposed a model for generating sentences by concatenating long short-term memory (LSTM) [25]. The model has an encoder/decoder structure that inputs the image features extracted by the CNN in the first stage to the LSTM in the second stage and then generates a sentence.

Attention-based methods such as casting support tracking [26], [27], [28] have been proposed and have become mainstream approaches. They improve the accuracy and visual comprehensibility by outputting the weight indicating where to focus in an image when generating words. Because attention can take the correspondence between the word in a caption and the image information that was used as a cue to outputting the word, our method exploits an attention-based captioning method. As an object detection method, Anderson et al. [27] used a Faster R-CNN [29] as the encoder and a combination of two LSTMs as the decoder. Yang et al.’s method [28] uses grid-based features obtained from a ResNet-152 [30] encoder. It also extracts global image features using image feature encoding (IFE) and then generates captions using a decoder based on a modified LSTM, called CaptionNet.

In addition, transformer [31]-based methods have become mainstream in the field of natural language processing and have also been introduced into the field of computer vision [32]. In image captioning, transformer-based methods have also been proposed and have shown a good performance [33], [34].

In this study, we used an attention-based method without applying a transformer, which has a simpler model and generates simpler captions than transformer-based models. This is because the main purpose of this study is to show the possibility of using image captioning to effectively estimate important regions, the main purpose of which is to detect subject and object words rather than generating a more natural and precise sentence.

C. VISUAL GROUNDING

Although the accuracy of captions in image captioning has been significantly improved through the introduction of an attention mechanism, the region in the image corresponding to a word is occasionally not properly attended, and a low explainability is a problem. Visual grounding improves the attention accuracy.

One feasible approach to this task is to directly learn attention using grounding supervision [35]. By contrast, Ma et al. [36] improved the accuracy of attention without the supervision of grounding by learning through a cyclic decoding, localization, and reconstruction processes without supervision. In the decoding process, the caption \( y^d = \{y_1^d, y_2^d, \ldots, y_T^d\} \) is generated by using CaptionNet. Subsequently, in the localization process, the localizer generates attention \( \beta_t \in \mathbb{R}^{1 \times K} \) from \( y_t^d \). In the reconstruction process, a caption \( y^r = \{y_1^r, y_2^r, \ldots, y_T^r\} \) is generated by replacing \( \alpha_t \) in the decoding process with \( \beta_t \). The localizer generates attention \( \beta_t \) from words only, and implicitly corrects the attention \( \alpha_t \) of the decoding process.

The captioning network generates captions using a beam search and occasionally generates a phrase including “of,” such as in “a group of people.” In this case, “group” is detected as the subject word, which does not correspond to a concrete object in the image. Therefore, the proposed method re-generates the caption by making the output probability of

III. METHODOLOGY

We propose a framework that exploits image captioning for important region estimation because captions are expected to be generated, reflecting the context of the image. When describing an image, the main objects in the image are usually the subject and object words of the sentence. Therefore, our framework estimates important objects by extracting subject and object words from the captions generated by the image-captioning method and identifying the regions that are focused on when generating those words.

An overview of the proposed framework is shown in Fig. 1. The proposed framework consists of five major modules: image feature extraction, caption generation, mask generation, subject and object word detection, and importance calculations. Details of each module are provided in the following subsections.

A. IMAGE FEATURE EXTRACTION

Given an input image, the objects in the image are detected, and their image features \( V \) are calculated. A Faster R-CNN [29] with ResNet-101-FPN as the backbone was applied for this purpose. After obtaining the features of the whole image, a region proposal network (RPN) generates some rectangular object regions. ROI pooling is applied to these regions to obtain the features \( V \in \mathbb{R}^{D \times K} \), where \( D \) and \( K \) denote the dimensionality of the feature and the number of objects, respectively. For the proposed method, we set \( K = 36 \) for each image.

B. CAPTION GENERATION

For caption generation, we used a method based on CaptionNet [28] proposed by Yang et al., incorporating cyclic training [36]. CaptionNet generates a caption \( C = \{y_1, y_2, \ldots, y_T\} \) from grid-based image features extracted through ResNet-152 [30], where \( y_t \) corresponds to a word. Here, \( \alpha_t \in \mathbb{R}^{1 \times K} \) is the attention corresponding to \( y_t \) and is calculated during the caption generation process.

Because our purpose is to estimate important regions, with the proposed method, we use region-based features instead of grid-based features because the attention obtained from the grid-based features does not necessarily correspond to each object. In addition, we incorporated cyclical training [36] proposed by Ma et al. to improve the accuracy of the attention. With their method, attention is updated in cyclical order of decoding, localization, and reconstruction processes without supervision. In the decoding process, the caption \( y^d = \{y_1^d, y_2^d, \ldots, y_T^d\} \) is generated using CaptionNet. Subsequently, in the localization process, the localizer generates attention \( \beta_t \in \mathbb{R}^{1 \times K} \) from \( y_t^d \). In the reconstruction process, a caption \( y^r = \{y_1^r, y_2^r, \ldots, y_T^r\} \) is generated by replacing \( \alpha_t \) in the decoding process with \( \beta_t \). The localizer generates attention \( \beta_t \) from words only, and implicitly corrects the attention \( \alpha_t \) of the decoding process.

The captioning network generates captions using a beam search and occasionally generates a phrase including “of,” such as in “a group of people.” In this case, “group” is detected as the subject word, which does not correspond to a concrete object in the image. Therefore, the proposed method re-generates the caption by making the output probability of
the preceding word (e.g., “group” from “a group of people”) extremely small if the first five words of the caption include “of.” This will produce a caption that does not contain “of” near the subject word.

C. MASK GENERATION

We detect pixel-level object regions as masks using a Mask R-CNN [38]. Masks \( M = \{M_1, M_2, \ldots, M_K\} \) corresponding to the image features \( V = \{V_1, V_2, \ldots, V_K\} \) are generated. The Mask R-CNN consists of a Faster R-CNN and a mask head that detects pixel-level object regions. Because the Faster R-CNN is used for image feature extraction, as described in III-A, by sharing the Faster R-CNN part of the Mask R-CNN with the image feature extraction module, masks that exactly correspond to the image features can be generated.

D. SUBJECT AND OBJECT WORDS DETECTION

The subject word in the caption usually indicates the primary object in the image and the object word indicates the object of the action. The proposed method detects the subject and object \( S = \{C_{\text{subj}}, C_{\text{obj}}\} \) through dependency parsing using the natural language processing tool, Stanza [39].

The algorithm used by the proposed method is shown in Fig. 2, where dependency parsing shows the modification relationship between the words, as indicated in Fig. 3. The subject \( C_{\text{subj}} \) is the word of the destination of the \( \text{nsubj} \) tag, which represents a subject noun, and is “man” in the example shown in Fig. 3(a). However, if \( \text{nsubj} \) is not detected because the sentence is incomplete, as shown in Fig. 3(b), the root of the dependency tree (the word not specified as the dependency destination of all the other words) is used as the subject. In addition, if the word (“man” in Fig. 3) that is determined through the above procedure to be the subject has a coordinating conjunction tag \( \text{conj} \), the destination of the tag (“woman”) is also regarded as the subject. Object \( C_{\text{obj}} \) is the word of the destination of the tag \( \text{obj} \) (“frisbee”).

E. IMPORTANCE CALCULATION

An importance map, which visualizes the important regions in an image according to their importance, is generated using a localizer \( \beta \in \mathbb{R}^{T \times K} \), subject and object \( S \), and masks \( M = \{M_1, M_2, \ldots, M_K\} \). Here, \( \beta \) is a matrix that shows the weight of each of \( K \) regions for each of \( T \) caption words.

As shown in Fig. 4, each mask is first filled with a value of attention \( \beta_{t,k} \) to obtain the maps \( \text{ImpMap}_{t,k}(x,y) \) for each word \( y_t \) of the subject and object. The attention map for the word \( y_t \) is then generated by referring to the maximum value of \( \text{ImpMap}_{t,k}(x,y)(k = 1, \ldots, K) \), and is normalized by the


Algorithm 1 Detect subject and object words in the caption

Require: caption
Ensure: sbj, obj
1: sbj ← []
2: sbj_tmp ← []
3: obj ← []
4: caption_tmp ← caption
5: while caption ≠ [] do
6:   word ← pop(caption)
7:   if word.deprel = “nsbj” then
8:     append(sbj, word)
9:   else if word.deprel = “obj” then
10:      append(obj, word)
11:   else
12:     if word.head = “root” then
13:        append(sbj_tmp, word)
14:   end if
15: end if
16: end while
17: if sbj = [] then
18:   sbj ← sbj_tmp
19: end if
20: caption ← caption_tmp
21: while caption ≠ [] do
22:   word ← pop(caption)
23:   if word.deprel = “conj” ∧ word.head ∈ sbj then
24:      append(sbj, word)
25:   end if
26: end while

FIGURE 2. Algorithm for subject and object word detection. deprel and head indicate the dependency tag and dependency destination, respectively.

maximum value to obtain $\text{ImpMap}_t(x, y)$.

$$\text{ImpMap}_{t,k}(x, y) = \begin{cases} \beta_{t,k} & \text{if } (x, y) \in M_k \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$\text{ImpMap}_t(x, y) = \max_k (\text{ImpMap}_{t,k}(x, y)) \quad (2)$$

After obtaining the normalized map $\text{ImpMap}_t(x, y)$ for all subjects and objects, the importance map is obtained by taking their maximum values as follows:

$$\text{ImpMap}(x, y) = \max_t (\text{ImpMap}_t(x, y)). \quad (3)$$

IV. DATASETS

A. DATASET OF IMPORTANT REGIONS IN IMAGES

To the best of the author’s knowledge, there are no datasets that define the important regions in an image. Therefore, we created a dataset to evaluate the proposed framework through subjective evaluation experiments.

Thirteen participants were asked to judge whether each region in the image was important. Twenty images were extracted from the test images in the Karpathy split [40] of the MS COCO dataset [41] and used in the experiments. These images have annotation masks of the objects, which we used as candidates for the important regions. When the research participants selected the important regions, they were restricted to within 30% of the entire image in their subjective evaluation. This is because we considered it important to narrow down the important areas to the minimum number necessary.

Finally, regions defined as important by more than half of the participants were defined as important regions.

B. DATASET USED FOR THRESHOLD SETTING

We also prepared another dataset to set the thresholds. Sixty images were extracted from the MS COCO test images that were not contained in the dataset described in Section IV-A. These images do not have any annotations, such as masks or captions; therefore, we used the masks obtained by the Mask R-CNN as candidates for the important regions. Because these masks are inaccurate and contain some errors, and because our method exploits the Mask R-CNN, it is inappropriate to use them to evaluate the methods. Therefore, we did not use this dataset for the evaluation but only for setting the thresholds.

The dataset was created in two stages. In the first stage, the research participants were asked to score each region corresponding to an object from four grades: 1 = very important (1.0), 2 = somewhat important (0.75), 3 = insignificant (0.25), and 4 = not important at all (0), where the numbers in parentheses indicate the scores at each grade. In the second stage, the regions were displayed in order of the highest score on average. The research participants were then asked to answer with a rank that divides important from non-important regions, and the threshold between important
Caption: "a man holding a hot dog in front of a crowd of people"

**TABLE 1.** Performance comparison with saliency-based methods. Bold indicates the best performance; and italics, the second best.

| Method   | Threshold | Recall | Precision | F-measure |
|----------|-----------|--------|-----------|-----------|
| Zhang [10]| 0.2       | 60.11  | 35.17     | 44.38     |
| Hou [13] | 0.7       | 76.17  | 67.97     | 71.84     |
| Zhao [16] | 0.7       | 71.87  | 83.23     | 77.15     |
| Pang [17] | 0.9       | 79.92  | 79.23     | 79.57     |
| Wei [19]  | 0.5       | 86.75  | 76.61     | 81.57     |
| Ours     | 0.4       | 81.74  | 89.00     | 85.22     |

**TABLE 2.** Processing time (s).

| Process                        | Time |
|--------------------------------|------|
| Image feature extraction       | 0.499|
| Mask generation                | 0.132|
| Caption generation             | 0.105|
| Subject and object words detection |      |
| Importance calculation         |      |

and non-important regions was determined based on the majority vote.

**V. EXPERIMENTS**

We used 113,287 images in the Karpathy split [40] of the MS COCO dataset [41] for training the image feature extraction, caption generation, and mask generation used by the proposed method. Because there has been no research aimed at explicitly detecting important regions, we compared our method with saliency-based approaches. With these methods, a saliency map is regarded as an importance map.

**A. QUANTITATIVE EVALUATION**

We conducted experiments using the important region dataset described in Section IV-A. We evaluated the methods by calculating the recall, precision, and F-measure using the pixel-level overlap between the estimated map binarized by the threshold and the ground-truth mask. We set the threshold for the binarization of each method as the value that maximizes the F-measure on the dataset for threshold setting described in Section IV-B. The results are presented in Table 1. As shown in the table, the proposed method achieved the highest F-measure scores, demonstrating the superiority of the proposed method.

The average processing time for each module is listed in Table 2. Image feature extraction and mask generation were applied simultaneously in one program; therefore, it was impossible to measure the processing time separately. The processing time of these processes was measured using a machine with four GeForce GTX 1080 GPUs, an Intel i7-6850K @3.6GHz CPU, and 128 GB of RAM. Subject and object word detection and importance calculations were conducted simultaneously in a single program. The processing time of these processes and caption generation were measured using a machine with a single GeForce RTX 3080 GPU, AMD Ryzen7 3700X @3.6GHz CPU, and 32 GB of RAM.

Fig. 5 displays the precision–recall curve. The proposed method has a larger change in precision when the threshold is
changed compared with the other methods. Examples of binarization of the importance map when changing the threshold are shown in Fig. 6. There are several people in this image, and it can be seen that the proposed method ranks the person near the center of the image and facing the camera as the most important. Thus, because the proposed method can rank each object based on its importance, the number of important regions can be flexibly changed by changing the threshold value. However, because saliency-based methods such as that developed by Wei et al. use a binary mask as the ground truth, they tend to output a saliency map with values extremely close to 0 or 1. Therefore, even if the threshold is changed, the binarized result does not significantly change.

### B. QUALITATIVE EVALUATIONS

Fig. 7 shows the results of the proposed method and saliency-based methods. Table 3 lists the captions of the images obtained using the proposed method.

Zhang et al.’s method [10] failed to properly extract the object regions. This is because this method creates a saliency map that emphasizes human visual characteristics, and it is considered to react too strongly to regions where the brightness is high owing to light reflection. The method proposed by Hou et al. [13] showed a similar tendency. In particular, image2 and image4 are representative examples, and unimportant regions are detected. Zhao et al.’s method [16] emphasizes the main subject of the image, such as the person at the center of image1, and suppresses the background. However, there are images where only a part of an important object was emphasized, such as image4, and it failed in complex images such as image2. Although the methods developed by Pang et al. [17] and Wei et al. [19] accurately identified the important regions in image1 and image3, there were cases in which the important regions were not properly narrowed down, such as image2 and image4.

The proposed method accurately detected important regions, not only in simple images with few objects such as image3 and image4, but also in complex images with many people such as image1. This is because the caption generation network was able to determine which object should be the subject, and the localizer was able to select the object to be focused on based on the direction in which the object was facing and what the object was holding, even in images with multiple objects (people) of the same class, such as image1.

**TABLE 3. Captions generated by the proposed method. Bold indicates a word detected as the subject and italics indicate a word detected as an object.**

| Image   | Caption                                                                 |
|---------|-------------------------------------------------------------------------|
| image1  | a man holding a hot dog in front of a crowd of people.                  |
| image2  | a woman holding an umbrella while standing next to a man in a blue shirt.|
| image3  | a man holding a baseball bat in a batting cage.                         |
| image4  | a woman standing in front of a table filled with bananas and other items.|
| image5  | a man is standing next to a herd of cattle.                             |
FIGURE 7. Visual comparison of different methods. The upper part of each row block shows the importance map (or saliency map in saliency-based methods), and the lower part represents the image obtained by binarizing the map with the threshold shown in Table 1.

| Table 4 | Comparison of different attention generation methods with or without cyclical learning. "Decoder" in the attention column uses $\alpha$, and "Localizer" uses $\beta$, as described in Section III-B. Bold indicates the best result; and italics, the second best. |
|---|---|---|---|---|
| Attention | Cyclical | Recall | Precision | F-measure |
| Decoder | 65.09 | 82.91 | 72.93 |
| Decoder | ✓ | 73.67 | 82.93 | 78.03 |
| Localizer | ✓ | 81.74 | 89.00 | 85.22 |

The proposed method detects different regions from the ground-truth in image 5. In this case, the man in the upper-left corner was detected as an important object because the caption generation network determined that the word corresponding to the man was the subject word. However, from another perspective, the proposed method can be interpreted as making us aware of human existence, which can be overlooked. In image 5, none of the methods detected regions similar to the ground truth. In summary, unlike saliency-based methods, the proposed method can detect important regions rather than visually conspicuous regions by considering the context of the image.

C. ABOLATION STUDY

With the proposed method, cyclical training [36] was exploited and a localizer was used to generate attention. We verified how they contribute to the accuracy of the
important region estimation. Table 4 shows the quantitative evaluation when the threshold is 0.4.

For the F-measure, the highest result was achieved by introducing cyclical training and using the attention generated by the localization process. Fig. 8 shows the importance maps obtained by these methods. Because the proposed method outputs an importance map from only words, it is able to generate an importance map that is close to human sensibility and is considered to be effective in discriminating important regions.

VI. CONCLUSION

In this study, we proposed a novel framework for estimating important regions in an image. The proposed method acquires semantic information from images using image captioning and estimates important regions using those regions corresponding to the subject and object words of the generated caption. By exploiting a localizer for cyclical training, it was confirmed experimentally that the proposed method can estimate important regions closer to the level of human sensitivity. In addition, to evaluate the effectiveness of the proposed method, we created a dataset that defines important image regions. By comparing the proposed method with conventional saliency-based approaches, we confirmed that the proposed method can estimate important regions more appropriately than the conventional methods based on both quantitative and qualitative evaluations.

Future studies will include developing a method to eliminate captioning failures by incorporating features of the entire image, which considers the relative size of the objects, not just local features obtained through image feature extraction, and a method for improving the accuracy of the attention.

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, “The digitization of the world from edge to core,” Int. Data Corp., Needham, MA, USA, White Paper US44413318, 2018.

[2] Cisco Visual Networking Index: Forecast and Trends, 2017–2022 White Paper, Cisco Syst., San Jose, CA, USA, 2019.

[3] S. Shioiri, Y. Sato, Y. Horaguchi, H. Muraoka, and M. Nihei, “Quali-informatics in the society with yotta scale data,” in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), May 2021, pp. 1–4.

[4] V. Snił, “Data world: Racing toward yotta,” IEEE Spectr., vol. 56, no. 7, p. 20, Jul. 2019.

[5] O. Issa and T. Shanableh, “CNN and HEVC video coding features for static video summarization,” IEEE Access, vol. 10, pp. 72080–72091, 2022.

[6] A. Skodras, C. Christopoulos, and T. Ebrahimi, “The JPEG 2000 still image compression standard,” IEEE Signal Process. Mag., vol. 18, no. 5, pp. 36–58, Sep. 2001.

[7] H. Alers, J. Redi, H. Liu, and I. Heyndericks, “Studying the effect of optimizing image quality in salient regions at the expense of background content,” J. Electron. Imag., vol. 22, no. 4, Nov. 2013, Art. no. 043012.

[8] V. Hous, F. Hahn, O. Wiedemann, S.-H. Jung, and D. Sape, “Saliency-driven image coding improves overall perceived JPEG quality,” in Proc. Picture Coding Symp. (PCS), Dec. 2016, pp. 1–5.

[9] L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[10] J. Zhang and S. Sclaroff, “Saliency detection: A Boolean map approach,” in Proc. IEEE Int. Conf. Comput. Vis., vol. 12, Dec. 2013, pp. 153–160.

[11] X. Shen and Y. Wu, “A unified approach to salient object detection via low rank matrix recovery,” in Proc. CVPR, Jul. 2012, pp. 853–860.

[12] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and H. T. Philip, “Deeply supervised salient object detection with short connections,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2017, pp. 5300–5309.

[13] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, “Deeply supervised salient object detection with short connections,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 4, pp. 815–828, Apr. 2019.

[14] S. Xie and Z. Tu, “Holistically-nested edge detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1395–1403.

[15] M. A. Islam, M. Kalash, and N. D. B. Bruce, “Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7142–7150.

[16] T. Zhao and X. Wu, “Pyramid feature attention network for saliency detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3085–3094.

[17] Y. Pang, X. Zhao, L. Zhang, and H. Lu, “Multi-scale interactive network for salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 9413–9422.

[18] Q. Chen, T. Liu, Y. Shang, Z. Shao, and H. Ding, “Salient object detection: Integrate salient features in the deep learning framework,” IEEE Access, vol. 7, pp. 152483–152492, 2019.
