

DESCRIPTION

Invention Title: METHOD AND APPARATUS FOR DETECTING IMAGE ABNORMALITY

5

Technical Field

[1] The present disclosure relates to a method and an apparatus for detecting an image abnormality by detecting an abnormal sign included in an image.

10

[2]

Background Art

15

[3] In recent years, an image abnormality detection for determining abnormal data included in images without direct human observation has been applied and used in various fields. For example, in the manufacturing field, abnormal behavior and irregular work can be detected in the production process and, in the medical field, the patient's condition can be monitored to automatically send an alert to the medical staff when an accident occurs. In addition,

20

in the public safety field, acts such as illegal activities or accidents in images can be preemptively recognized in order to protect lives and property.

25

[4] In order for such an image abnormality detection to be performed smoothly, it is necessary to secure each of normal data and abnormal data in a balanced manner. However,

as an event corresponding to the abnormal data rarely occurs in daily life, the abnormal data is insufficient, resulting in a problem of making it difficult to smoothly perform the image abnormality detection because of an imbalance between the normal data and the abnormal data. In addition, there is also a problem of making it difficult to smoothly perform the image abnormality detection because the criteria for distinguishing the normal data from the abnormal data are not clear.

[5] In order to address the problem in the related art, the image abnormality detection is performed based on information about visual features included in the image, depending on the level of supervision by an image observer. However, it is difficult to clearly determine the semantic meaning included in the images based on only the information about the visual features, thereby frequently degrading the accuracy of the image abnormality detection.

[6]

[7] The matters described above as the background technology are only intended to enhance an understanding of the background of the present disclosure, and should not be construed as acknowledging that it corresponds to the prior art already known to those skilled in the art.

[8]

Disclosure

Technical Problem

[9] The present disclosure is proposed to solve such problems, and is to provide a method and an apparatus for detecting an image abnormality by detecting an abnormal sign included in an image by using not only a visual feature but also a text feature included in the image.

[10]

[11] Technical tasks to be achieved by the present disclosure are not limited to the technical tasks mentioned above, and other technical tasks not mentioned may be clearly understood by those skilled in the art to which the present disclosure belongs from the following description.

[12]

Technical Solution

[13] A method for detecting an image abnormality according to the present disclosure for achieving the objective described above includes collecting at least one piece of image information to divide the at least one piece of image information into a plurality of snippets, extracting visual feature information and text feature information for each of the plurality of divided snippets, generating synthesis information by synthesizing the extracted visual feature information and text feature information, and detecting an abnormal sign of the at least one piece of image information based on the generated synthesis information.

[14] For example, the extracting step may include performing a pre-processing, in which a frame is cropped according to a predetermined number of cropping times for each of the plurality of divided snippets, and extracting the visual feature information for each of the plurality of snippets in which the pre-processing is performed.

[15] For example, the extracting step may include generating a caption included in each of the plurality of divided snippets, and extracting the text feature information by performing sentence embedding based on the generated caption.

[16] For example, the extracting step may include performing multi-scale correction on the visual feature information and the text feature information for each of the plurality of snippets, and extracting the visual feature information and the text feature information in which the correction is performed.

[17] For example, the generating step may include generating the synthesis information by using at least one method of a plurality of synthesizing methods for the extracted visual feature information and text feature information.

[18] For example, the detecting step may include determining a first loss value corresponding to a total loss value of the at least one piece of image information and a second loss value corresponding to a training value for detecting the abnormal sign of the at least one piece of image information

based on the generated synthesis information, generating a loss function based on the determined first loss value and second loss value, and detecting the abnormal sign of the at least one piece of image information based on the generated loss function.

[19] For example, the determining each of the first loss value and the second loss value may include determining a feature size corresponding to each of the plurality of snippets based on the generated synthesis information, determining an average feature size of the feature sizes corresponding to a predetermined reference number in size order among the determined feature sizes, and determining the first loss value based on the determined average feature size.

[20] For example, the determining each of the first loss value and the second loss value may include determining a feature size corresponding to each of the plurality of snippets based on the generated synthesis information, determining a frame-level prediction value of each of the plurality of snippets based on the determined feature size, determining an abnormality score of the at least one piece of image information based on the frame-level prediction value corresponding to the feature size corresponding to a predetermined reference number in size order among the feature sizes, and determining the second loss value based on the determined abnormality score.

[21] For example, when detecting the abnormal sign after the detecting step, outputting an abnormality score of the at least one piece of image information and a caption included in the at least one piece of image information may be further included.

[22] For example, the outputting step may include outputting a warning alarm when the abnormality score of the at least one piece of image information exceeds a predetermined reference score.

[23]

[24] In addition, an apparatus for detecting an image abnormality according to the present disclosure for achieving the objective includes an image collection unit for collecting at least one piece of image information to divide the at least one piece of image information into a plurality of snippets, an information extraction unit for extracting visual feature information and text feature information for each of the plurality of snippets divided by the image collection unit, and a detection control unit for generating synthesis information by synthesizing the extracted visual feature information and text feature information and for detecting an abnormal sign of the at least one piece of image information based on the generated synthesis information.

[25] For example, the information extraction unit may perform a pre-processing, in which a frame is cropped according to a predetermined number of cropping times for each of the plurality of divided snippets, and extract the visual feature information for each of the plurality of snippets in which the pre-processing is performed.

[26] For example, the information extraction unit may generate a caption included in each of the plurality of divided snippets, and extract the text feature information by performing sentence embedding based on the generated caption.

[27] For example, the information extraction unit may perform multi-scale correction on the visual feature information and the text feature information for each of the plurality of snippets, and extract the visual feature information and text feature information in which the correction is performed.

[28] For example, the detection control unit may generate the synthesis information by using at least one method of a plurality of synthesizing methods for the extracted visual feature information and text feature information.

[29] For example, the detection control unit may determine a first loss value corresponding to a total loss value of the at least one piece of image information and a second loss value corresponding to a training value for detecting the abnormal sign of the at least one piece of image information

based on the generated synthesis information, generate a loss function based on the determined first loss value and second loss value, and detect the abnormal sign of the at least one piece of image information based on the generated loss function.

[30] For example, the detection control unit may determine a feature size corresponding to each of the plurality of snippets based on the generated synthesis information, determine an average feature size of the feature sizes corresponding to a predetermined reference number in size order among the determined feature sizes, and determine the first loss value based on the determined average feature size.

[31] For example, the detection control unit may determine a feature size corresponding to each of the plurality of snippets based on the generated synthesis information, determine a frame-level prediction value of each of the plurality of snippets based on the determined feature size, determine an abnormality score of the at least one piece of image information based on the frame-level prediction value corresponding to the feature size corresponding to a predetermined reference number in size order among the feature sizes, and determine the second loss value based on the determined abnormality score.

[32] For example, when detecting the abnormal sign, the detection control unit may output an abnormality score of the at least one piece of image information and a caption included in the at least one piece of image information.

5 [33] For example, the detection control unit may output a warning alarm when the abnormality score of the at least one piece of image information exceeds a predetermined reference score.

[34]

10 **Advantageous Effects**

[35] As described above, a method and an apparatus for detecting an image abnormality of the present disclosure can improve the accuracy in detecting the image abnormality by detecting abnormal signs included in an image based on a visual feature and a text feature extracted from the image.

15 [36] Also, it is possible to enhance the robustness of a method and an apparatus for detecting an image abnormality by performing multi-scale correction for each of a visual feature and a text feature.

20 [37] In addition, by outputting an abnormality score for an image and a caption included in the image, a user who uses a method and an apparatus for detecting an image abnormality according to the present disclosure can respond efficiently to abnormal signs included in the image.

[38] Furthermore, it is possible to capture a semantic meaning of abnormal signs included in an image by using not only a visual feature but also a text feature, which can be applied to various fields.

5 [39]

[40] The effects obtainable from the present disclosure are not limited to the effects mentioned above, and other effects not mentioned will be clearly understood by those skilled in the art to which the present disclosure belongs from the description below.

10 [41]

Description of Drawings

[42] FIGS. 1 to 4 are block diagrams illustrating a configuration of an apparatus for detecting an image abnormality according to an exemplary embodiment of the present disclosure.

15 [43] FIG. 5 is a flowchart illustrating a method for detecting an image abnormality according to an exemplary embodiment of the present disclosure.

20 [44]

Mode for Disclosure

[45] In describing an exemplary embodiment disclosed in the present specification, the detailed description thereof may be omitted when it is determined that a detailed description of the related known technology may obscure the gist of the

exemplary embodiment disclosed in the present specification. In addition, the accompanying drawings may be only intended to facilitate the understanding of the exemplary embodiment disclosed in the present specification, and the technical
5 idea disclosed in the present specification may not be limited by the accompanying drawings and should be understood to include all modifications, equivalents, or substitutes included in the spirit and technical scope of the present disclosure.

10 **[46]** Terms including ordinal numbers, such as a first and second, may be used to describe various components, but the components may not be limited by the terms. The terms may be used only for the purpose of distinguishing one component from another component.

15 **[47]** When it is mentioned that a component is "connected" or "linked" to another component, it should be understood that the component may be directly connected or linked to the other component, but another component may be interposed therebetween. On the other hand, when it is mentioned that a
20 component is "directly connected" or "directly linked" to another component, it should be understood that no other component exists in the middle.

[48] Singular expressions may include plural expressions unless the context clearly indicates otherwise.

[49] In the present specification, terms such as "include" or "have" may be intended to specify the existence of features, numbers, steps, operations, components, parts, or combinations thereof described in the specification, which should be understood not to preclude the existence or addition of one or more other features, numbers, steps, operations, components, parts, or combinations thereof.

[50] Hereinafter, exemplary embodiments disclosed in the present specification will be described in detail with reference to the accompanying drawings, but identical or similar components will be assigned the same reference numbers regardless of the drawing codes, and redundant descriptions thereof will be omitted.

[51] An apparatus for detecting an image abnormality according to an exemplary embodiment of the present disclosure will be described with reference to FIGS. 1 to 4.

[52]

[53] FIGS. 1 to 4 are block diagrams illustrating a configuration of an apparatus for detecting an image abnormality according to an exemplary embodiment of the present disclosure.

[54] Referring to FIG. 1, the apparatus for detecting an image abnormality according to an exemplary embodiment of the present disclosure may include an image collection unit 100, an information extraction unit 200, and a detection control

unit 300. FIG. 1 mainly shows components related to an exemplary embodiment of the present disclosure, and it is obvious that fewer or more components may be included in implementing an apparatus for detecting an actual image abnormality.

[55] Meanwhile, each component 100, 200, 300 of the apparatus for detecting an image abnormality according to an exemplary embodiment of the present disclosure may include a communication device for communicating with other components, a memory for storing an operating system, logic commands and input/output information, and at least one processor for performing determination, calculation, and decision necessary for controlling a function in charge.

[56] Hereinafter, each component will be described with reference to FIGS. 2 to 4.

[57] First, the image collection unit 100 according to an exemplary embodiment of the present disclosure will be described with reference to FIG. 2.

[58] Referring to FIG. 2, the image collection unit 100 may collect at least one piece of image information and divide the at least one piece of image information into a plurality of snippets. For example, the image collection unit 100 may collect at least one piece of image information captured by an external capturing device configured to capture an image in real time as input information. Also, the image

collection unit 100 may divide the at least one piece of image information into the plurality of snippets, and the at least one piece of image information may be divided into the plurality of snippets (e.g., SN_1 , SN_2 , ..., SN_T) as many as a predetermined number (e.g., T).

[59] In addition, the image collection unit 100 may divide at least one piece of image information into the plurality of snippets, and manage the snippets by grouping snippets corresponding to the at least one piece of image information. However, this may be illustrative and may not be necessarily limited thereto. For example, the image collection unit 100 may be configured to collect only image information captured by the external capturing device, and accordingly the function of dividing the image information into the plurality of snippets may be performed by a device separate from the image collection unit 100.

[60] The image collection unit 100 may transmit information about a plurality of divided snippets to the information extraction unit 200.

[61] Next, the information extraction unit 200 according to an exemplary embodiment of the present disclosure will be described with reference to FIG. 3.

[62]

[63] Referring to FIG. 3, the information extraction unit 200 may receive information about the plurality of divided

snippets, and, based on the information, extract visual feature information and text feature information for each of the plurality of snippets. To this end, the information extraction unit 200 may include a visual feature extraction unit 210 and a text feature extraction unit 220. Also, the visual feature extraction unit 210 may include a preprocessing unit 211 and a multi-scale correction unit 212, and the text feature extraction unit 220 may include a caption generation unit 221, a sentence embedding unit 222, and a multi-scale correction unit 223. FIG. 3 mainly shows components related to an exemplary embodiment of the present disclosure, and it is obvious that fewer or more components may be included in implementing the actual information extraction unit 200.

[64] Hereinafter, each component will be described.

[65] The visual feature extraction unit 210 may extract the visual feature information included in each of the plurality of snippets based on the information about the plurality of snippets provided by the image collection unit 100. For example, the visual feature extraction unit 210 may extract the visual feature information included in each of the plurality of snippets through ResNet-50, which is a pre-trained neural network, and the visual feature information may be information according to an Inflated 3D ConvNet (I3D)

method. However, this may be illustrative and may not be necessarily limited thereto.

[66] Also, the visual feature extraction unit 210 may perform a pre-processing process through the preprocessing unit 211 in order to increase the consistency of information when extracting visual feature information. For example, the preprocessing unit 211 may perform a pre-processing of cropping a frame according to a predetermined number of cropping times for each of the plurality of divided snippets. For example, when the predetermined number of cropping times is 5, the preprocessing unit 211 may enable each frame of the plurality of snippets to be cropped into four corner portions and a center portion.

[67] When the plurality of snippets is preprocessed through the preprocessing unit 211, the visual feature extraction unit 210 may extract the visual feature information for each of the plurality of snippets on which the pre-processing is performed. By performing the pre-processing of cropping the frame for each of the plurality of snippets, the amount of information to be processed may increase compared to information about the existing plurality of snippets and, by extracting the visual feature information through this, the consistency of information can be enhanced.

[68] The text feature extraction unit 220 may extract the text feature information included in each of the plurality of

snippets based on information about each of the plurality of snippets provided by the image collection unit 100. For example, the text feature extraction unit 220 may generate a caption included in each of the plurality of snippets through the caption generation unit 221. In this case, the caption generation unit 221 may generate captions for consecutive frames by using a sliding window strategy for each of the plurality of snippets. In addition, the caption may be for describing an event included in each of the plurality of snippets and, for example, for describing an action of a person, an arrangement state of an object, and the like included in each of the plurality of snippets. However, this may be illustrative and may not be necessarily limited thereto.

[69] The caption generation unit 221 may transmit information about the generated caption to the sentence embedding unit 222, and the sentence embedding unit 222 may perform sentence embedding based on the generated caption. For example, the sentence embedding unit 222 may perform the sentence embedding by using a drop-out noise method or based on a pair recorded in a natural language inference data set, which is one of natural language processing (NLP). However, this may be illustrative, and may not be necessarily limited thereto. The text feature extraction unit 220 may generate the caption for each of the plurality of snippets through

the caption generation unit 221 and the sentence embedding unit 222, and extract the text feature information by performing the sentence embedding.

[70] Meanwhile, the visual feature extraction unit 210 and the text feature extraction unit 220 may perform the multi-scale correction on the firstly extracted visual feature information and text feature information for each of the plurality of snippets before transmitting the visual feature information and the text feature information to the outside. To this end, the visual feature extraction unit 210 and the text feature extraction unit 220 may be provided with multi-scale correction units 212, 223, respectively.

[71] For example, the multi-scale correction units 212, 223 may perform the multi-scale correction on the visual feature information and the text feature information firstly extracted through a temporal network. The temporal network may include a multiple layer pyramid dilated convolution block and a non-local block. The pyramid dilated convolution for the temporal range may be used to learn representations of the plurality of snippets at multiple scales, and the non-local block may be used to learn the overall temporal dependence between the plurality of snippets. By performing the multi-scale correction on each of the visual feature information and the text feature information through the temporal network in the multi-scale correction unit 212, 223,

the temporal dependency may be well captured when synthesizing the visual feature information and the text feature information corrected and extracted in the visual feature extraction unit 210 and the text feature extraction unit 220.

[72] As described above, the information extraction unit 200 may extract the visual feature information through the visual feature extraction unit 210, extract the text feature information through the text feature extraction unit 220, and provide the extracted visual feature information and text feature information to the detection control unit 300.

[73] Next, the detection control unit 300 according to an exemplary embodiment of the present disclosure will be described with reference to FIG. 4.

[74]

[75] Referring to FIG. 4, the detection control unit 300 according to an exemplary embodiment of the present disclosure may generate synthesis information by synthesizing the visual feature information and the text feature information extracted in the information extraction unit 200 and may detect an abnormal sign of at least one image piece of information based on the generated synthesis information. To this end, the detection control unit 300 may include a synthesis unit 310, a determination unit 320, and an abnormality detection unit 330. FIG. 4 mainly shows

components related to an exemplary embodiment of the present disclosure, and it is obvious that fewer or more components may be included in implementing the actual detection control unit 300.

5 **[76]** Hereinafter, each component will be described.

10 **[77]** The synthesis unit 310 may generate the synthesis information by synthesizing the visual feature information and text feature information extracted and provided by the information extraction unit 200, and may generate the synthesis information by using at least one method of a plurality of synthesizing methods. There may be concatenation, addition, and product methods in the plurality of synthesizing methods, but this may be illustrative and may not be necessarily limited thereto.

15 **[78]** Also, when generating the synthesis information through the addition and product method, the synthesis unit 310 may add a fully connected layer in order to reduce the dimension of the visual feature information to the same dimension as that of the text feature information, and then generate the synthesis information for the visual feature information and the text feature information. However, this may be illustrative and may not be necessarily limited thereto.

20 **[79]** The determination unit 320 may determine an image loss value based on the synthesis information generated by the synthesis unit 310, and may determine a first loss value and

25

a second loss value in order to determine the image loss value. For example, the first loss value may be a value corresponding to a total loss value of at least one piece of image information collected by the image collection unit 100, and the second loss value may be a value corresponding to a training value for detecting an abnormal sign of at least one piece of image information.

[80] Specifically, the determination unit 320 may determine a feature size corresponding to each of the plurality of snippets based on the synthesis information generated by the synthesis unit 310. In this case, the determination unit 320 may determine the feature size by using the l_2 norm method, but this may be illustrative and may not be necessarily limited thereto.

[81] Also, the determination unit 320 may determine an average feature size of the feature sizes corresponding to the predetermined reference number in size order among the determined feature sizes. For example, when the predetermined reference number is k , the determination unit 320 may extract feature sizes corresponding to the upper as many as k in size order among the feature sizes corresponding to each of the plurality of snippets, based on which an average feature size is determined. The average feature size ($f_{FM}(v; k)$) may be determined through Equation 1 below. In particular, the average feature size may be

determined as a feature size corresponding to at least one piece of image information.

[82] [Equation 1]

$$f_{FM}(v; k) = \frac{1}{k} \sum_{X_i \in \text{top}K(v; k)} \|X_i\|_2$$

[83] The determination unit 320 may determine the first loss value (L_{fm}) corresponding to the total loss value of at least one piece of image information by using Equation 2 below based on the average feature size determined through Equation 1 above.

[84] [Equation 2]

$$L_{fm} = \begin{cases} \sum_{j=1}^{|v|} (c - f_{FM}(v_j; k)), & \text{if } y_j = 1 \\ \sum_{j=1}^{|v|} f_{FM}(v_j; k), & \text{if } y_j = 0 \end{cases}$$

[85] The first loss value (L_{fm}) may refer to the total loss value including both normal data and abnormal data in at least one piece of image information.

[86] In addition, the determination unit 320 may determine the feature size corresponding to each of the plurality of snippets based on the generated synthesis information, based on which a frame-level prediction value of each of the plurality of snippets is determined. For example, the determination unit 320 may train a separately provided

binary snippet classifier based on the feature sizes corresponding to the predetermined reference number in size order among the determined feature sizes. Also, the determination unit 320 may determine a snippet-level prediction value propagated to an individual frame included in each of the plurality of snippets based on the trained binary snippet classifier, and may determine a frame-level prediction value of each of the plurality of snippets based on the determined snippet-level prediction value.

[87] Also, the determination unit 320 may determine an abnormality score of at least one piece of image information based on the frame-level prediction value corresponding to the feature size corresponding to the predetermined reference number in size order among feature sizes corresponding to each of the plurality of snippets. It is assumed that the determination unit 320 determines the abnormality score for the entire of at least one piece of image information based on a value corresponding to the upper reference number in size order among the frame-level prediction value corresponding to each of the plurality of snippets. Accordingly, the determination unit 320 may determine the abnormality score ($f_s(v; k)$) for the at least one piece of image information through Equation 3 below.

[88] [Equation 3]

$$f_s(v; k) = \frac{1}{k} \sum_{X_i \in \text{top}K(v; k)} f_{\text{pred}}(X_i; \delta)$$

[89] At this time, $f_{\text{pred}}(X_i; \delta)$ may refer to the frame-level prediction value corresponding to the feature size corresponding to the predetermined reference number in size order among feature sizes corresponding to each of the plurality of snippets. A secondary objective of the present disclosure may be to facilitate the distinction between normal data and abnormal data by increasing the difference between the abnormality score of image information containing normal data and the abnormality score of image information containing abnormal data. To achieve this, it is possible to train the binary snippet classifier based on the feature size corresponding to the predetermined reference number in size order among the feature sizes corresponding to each of the plurality of snippets. Also, it is possible to increase the difference between the abnormality score of image information containing normal data and the abnormality score of image information containing abnormal data by determining the abnormality score of image information based on the frame-level prediction value determined by the trained binary snippet classifier.

[90] Thereafter, the determination unit 320 may determine the second loss value corresponding to the training value for detecting an abnormal sign of at least one piece of image information based on the abnormality score ($f_s(v; k)$) for at least one piece of image information determined through Equation 3 above. In this case, the determination unit 320 may determine the second loss value (L_{bce}) through the following Equation 4 formed based on a binary cross entropy.

[91] [Equation 4]

$$L_{bce} = -\frac{1}{|v|} \sum_{j=1}^{|v|} \left(y_j \log(f_s(v_j; k)) + (1 - y_j) \log(1 - f_s(v_j; k)) \right)$$

[92] Also, the determination unit 320 may generate a loss function based on the first loss value (L_{fm}) and the second loss value (L_{bce}) determined as described above. In this case, the determination unit 320 may generate a loss function (L) by Equation 5 below.

[93] [Equation 5]

$$L = \alpha L_{fm} + L_{bce}$$

[94] In Equation 5 above, α may be a parameter for adjusting the weight of the loss component and, in an exemplary embodiment of the present disclosure, may adjust the weight of the first loss value. That is, the loss function may be generated in the form of a linear function based on the parameter α , the first loss value (L_{fm}) and the second loss

value (L_{bee}) as shown in Equation 5 above. However, this may be illustrative and may not be necessarily limited thereto.

[95] The determination unit 320 may provide the generated loss function to the abnormality detection unit 330, and the abnormality detection unit 330 may detect an abnormal sign of at least one piece of image information based on the generated loss function. Also, when detecting the abnormal sign in the at least one piece of image information, the abnormality detection unit 330 may output information about the detected abnormal sign, such that an external user may recognize the abnormal sign.

[96] For example, when detecting an abnormal sign from at least one piece of image information, the abnormality detection unit 330 may enable outputting an abnormality score of at least one piece of image information and a caption included in the at least one piece of image information. In this case, the abnormality score of the at least one piece of image information may be an abnormality score for the plurality of snippets into which the at least one piece of image information is divided, and the caption included in the at least one piece of image information may be a caption for the plurality of snippets into which the at least one piece of image information is divided.

[97] In addition, when outputting the abnormality score of the at least one piece of image information and the caption

included in the at least one piece of image information described above, the abnormality detection unit 330 may also enable outputting a warning alarm together therewith. For example, the abnormality detection unit 330 may compare the abnormality score of at least one piece of image information determined when detecting an abnormal sign with the predetermined reference score. As a result of the comparison, when the abnormality score of at least one piece of image information exceeds the predetermined reference score, the abnormality detection unit 330 may generate a warning alarm. By generating a warning alarm and outputting the generated warning alarm, the abnormality detection unit 330 may enable an external user to recognize and respond to the abnormal sign. However, this may be illustrative and may not be necessarily limited thereto.

[98]

[99] Hereinafter, a method for detecting an image abnormality according to an exemplary embodiment of the present disclosure will be described with reference to FIG. 5 based on the configuration of the apparatus for detecting an image abnormality described above with reference to FIGS. 1 to 4. However, since the detailed description of each step of FIG. 5 has been described above with reference to FIGS. 1 to 4, it will be omitted below.

[100] FIG. 5 is a flowchart illustrating a method for detecting an image abnormality according to an exemplary embodiment of the present disclosure.

[101] Referring to FIG. 5, the image collection unit 100 may collect at least one piece of image information from an external capturing device (S510), and divide the collected at least one piece of image information into a plurality of snippets (S520).

[102] Also, the information extraction unit 200 may extract the visual feature information and the text feature information for each of the plurality of snippets based on the plurality of divided snippets (S530).

[103] Also, the detection control unit 300 may generate the synthesis information by synthesizing the visual feature information and the text feature information extracted from the information extraction unit 200 (S540), and may determine a first loss value corresponding to the total loss value of at least one piece of image information and a second loss value corresponding to the training value for detecting abnormal signs of at least one piece of image information based on the generated synthesis information (S550).

[104] The detection control unit 300 may generate a loss function based on the determined first loss value and the second loss value (S560), and may perform abnormality

detection on at least one piece of image information based on the generated loss function (S570).

[105] Thereafter, when detecting an abnormal sign from at least one piece of image information during the abnormality detection (Yes in S580), the detection control unit 300 may enable outputting an abnormality score of at least one piece of image information and a caption included in the at least one piece of image information (S590). In addition, when the abnormality score of at least one piece of image information exceeds a predetermined reference score, the detection control unit 300 may enable outputting a warning alarm together when outputting the abnormality score and the caption.

[106]

[107] As described above, a method and an apparatus for detecting an image abnormality of the present disclosure can improve the accuracy in detecting the image abnormality by detecting abnormal signs included in an image based on a visual feature and a text feature extracted from the image.

[108] Also, it is possible to enhance the robustness of a method and an apparatus for detecting an image abnormality by performing multi-scale correction for each of a visual feature and a text feature.

[109] In addition, by outputting an abnormality score for an image and a caption included in the image, a user who uses a

method and an apparatus for detecting an image abnormality according to the present disclosure can respond efficiently to abnormal signs included in the image.

[110] Furthermore, it is possible to capture a semantic meaning of abnormal signs included in an image by using not only a visual feature but also a text feature, which can be applied to various fields.

[111]

[112] Although shown and described with respect to a specific exemplary embodiment of the present disclosure, it will be apparent to those skilled in the art that the present disclosure may be modified and changed in various ways without departing from the technical idea of the present disclosure provided by the following claims.

[113] The present disclosure described above may be implemented as computer-readable code on a medium on which a program is recorded. The computer-readable medium may include all types of recording devices in which data readable by a computer system is stored. Examples of computer-readable medium may include hard disk drives (HDDs), solid state disks (SSDs), silicon disk drives (SDDs), ROMs, RAMs, CD-ROMs, magnetic tapes, floppy disks, optical data storage devices, and the like. Therefore, the above detailed description should not be interpreted in all respects and should be considered as illustrative. The scope of the present disclosure should be

determined by reasonable interpretation of the appended claims, and all changes within the equivalent scope of the present disclosure may be included in the scope of the present disclosure.

5 **[114]**

[115] Description of Reference Numerals

[116] 100: image collection unit

[117] 200: information extraction unit

[118] 300: detection control unit

10

CLAIMS

[Claim 1] A method for detecting an image abnormality, the method comprising:

5 collecting at least one piece of image information to divide the at least one piece of image information into a plurality of snippets;

 extracting visual feature information and text feature information for each of the plurality of divided snippets;

10 generating synthesis information by synthesizing the extracted visual feature information and text feature information; and

 detecting an abnormal sign of the at least one piece of image information based on the generated synthesis information.

15

[Claim 2] The method of claim 1, wherein the extracting step comprises:

 performing a pre-processing, in which a frame is cropped according to a predetermined number of cropping times for each of the plurality of divided snippets; and

20

 extracting the visual feature information for each of the plurality of snippets in which the pre-processing is performed.

[Claim 3] The method of claim 1, wherein the extracting step comprises:

25

generating a caption included in each of the plurality of divided snippets; and

extracting the text feature information by performing sentence embedding based on the generated caption.

5

[Claim 4] The method of claim 1, wherein the extracting step comprises:

performing multi-scale correction on the visual feature information and the text feature information for each of the plurality of snippets; and

10

extracting the visual feature information and the text feature information in which the correction is performed.

[Claim 5] The method of claim 1, wherein the generating step comprises:

15

generating the synthesis information by using at least one method of a plurality of synthesizing methods for the extracted visual feature information and text feature information.

[Claim 6] The method of claim 1, wherein the detecting step comprises:

20

determining a first loss value corresponding to a total loss value of the at least one piece of image information and a second loss value corresponding to a training value for detecting the abnormal sign of the at least one piece of image

25

information based on the generated synthesis information;

generating a loss function based on the determined first loss value and second loss value; and

detecting the abnormal sign of the at least one piece of image information based on the generated loss function.

[Claim 7] The method of claim 6, wherein the determining each of the first loss value and the second loss value comprises:

determining a feature size corresponding to each of the plurality of snippets based on the generated synthesis information;

determining an average feature size of the feature sizes corresponding to a predetermined reference number in size order among the determined feature sizes; and

determining the first loss value based on the determined average feature size.

[Claim 8] The method of claim 6, wherein the determining each of the first loss value and the second loss value comprises:

determining a feature size corresponding to each of the plurality of snippets based on the generated synthesis information;

determining a frame-level prediction value of each of the

plurality of snippets based on the determined feature size;

determining an abnormality score of the at least one piece of image information based on the frame-level prediction value corresponding to the feature size corresponding to a predetermined reference number in size order among the feature sizes; and

determining the second loss value based on the determined abnormality score.

[Claim 9] The method of claim 1, further comprising:

outputting an abnormality score of the at least one piece of image information and a caption included in the at least one piece of image information when detecting the abnormal sign after the detecting step.

[Claim 10] The method of claim 9, wherein the outputting step comprises:

outputting a warning alarm when the abnormality score of the at least one piece of image information exceeds a predetermined reference score.

[Claim 11] An apparatus for detecting an image abnormality, the apparatus comprising:

an image collection unit configured to collect at least one piece of image information to divide the at least one piece

of image information into a plurality of snippets;

an information extraction unit configured to extract visual feature information and text feature information for each of the plurality of snippets divided by the image collection unit; and

a detection control unit configured to generate synthesis information by synthesizing the extracted visual feature information and text feature information and for detecting an abnormal sign of the at least one piece of image information based on the generated synthesis information.

[Claim 12] The apparatus of claim 11, wherein the information extraction unit configured to perform a pre-processing, in which a frame is cropped according to a predetermined number of cropping times for each of the plurality of divided snippets, and to extract the visual feature information for each of the plurality of snippets in which the pre-processing is performed.

[Claim 13] The apparatus of claim 11, wherein the information extraction unit further configured to generate a caption included in each of the plurality of divided snippets, and extracts the text feature information by performing sentence embedding based on the generated caption.

[Claim 14] The apparatus of claim 11, wherein the information extraction unit further configured to perform multi-scale correction on the visual feature information and the text feature information for each of the plurality of snippets, and extracts the visual feature information and text feature information in which the correction is performed.

[Claim 15] The apparatus of claim 11, wherein the detection control unit further configured to generate the synthesis information by using at least one method of a plurality of synthesizing methods for the extracted visual feature information and text feature information.

[Claim 16] The apparatus of claim 11, wherein the detection control unit further configured to

determine a first loss value corresponding to a total loss value of the at least one piece of image information and a second loss value corresponding to a training value for detecting the abnormal sign of the at least one piece of image information based on the generated synthesis information,

generate a loss function based on the determined first loss value and second loss value, and

detect the abnormal sign of the at least one piece of image information based on the generated loss function.

[Claim 17] The apparatus of claim 16, wherein the detection control unit further configured to

determine a feature size corresponding to each of the plurality of snippets based on the generated synthesis information,

determine an average feature size of the feature sizes corresponding to a predetermined reference number in size order among the determined feature sizes, and

determine the first loss value based on the determined average feature size.

[Claim 18] The apparatus of claim 16, wherein the detection control unit further configured to

determine a feature size corresponding to each of the plurality of snippets based on the generated synthesis information,

determine a frame-level prediction value of each of the plurality of snippets based on the determined feature size,

determine an abnormality score of the at least one piece of image information based on the frame-level prediction value corresponding to the feature size corresponding to a predetermined reference number in size order among the feature sizes, and

determine the second loss value based on the determined abnormality score.

[Claim 19] The apparatus of claim 11, wherein in response that detecting the abnormal sign, the detection control unit configured to output an abnormality score of the at least one piece of image information and a caption included in the at least one piece of image information.

[Claim 20] The apparatus of claim 19, wherein the detection control unit further configured to output a warning alarm when the abnormality score of the at least one piece of image information exceeds a predetermined reference score.