United States Patent Application Publication	20250259432
Kind Code	A1
Publication Date	August 14, 2025
Inventor(s)	BEYE; Florian et al.

VIDEO PROCESSING SYSTEM, VIDEO PROCESSING APPARATUS, AND VIDEO PROCESSING METHOD

Abstract

An object is to provide a video processing system, a video processing apparatus, and a video processing method that can be expected to improve recognition accuracy of an object in a video. A video processing system includes video acquisition means, time difference information acquisition mean, and recognition means. The video acquisition means acquires an input video. The time difference information acquisition means acquires first time difference information between the frames of the input video. The recognition means inputs the input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference information between frames of the training video and recognizing an object in the input video.

Inventors:	**BEYE; Florian (Tokyo, JP), IWAI; Takanori (Tokyo, JP), NIHEI; Koichi (Tokyo, JP), ITSUMI; Hayato (Tokyo, JP), TAKAHASHI; Katsuhiko (Tokyo, JP), BABAZAKI; Yasunori (Tokyo, JP), ANDO; Ryuhei (Tokyo, JP), PIAO; Jun (Tokyo, JP)**
Applicant:	**NEC Corporation** (Minato-ku, Tokyo, JP)
Family ID:	**90274577**
Assignee:	**NEC Corporation (Minato-ku, Tokyo, JP)**
Appl. No.:	**18/857236**
Filed (or PCT Filed):	**September 15, 2022**
PCT No.:	**PCT/JP2022/034510**

Publication Classification

Int. Cl.: **G06V10/82** (20220101); **G06T7/20** (20170101)

U.S. Cl.:

CPC **G06V10/82** (20220101); **G06T7/20** (20130101);

Background/Summary

TECHNICAL FIELD

[0001] The present disclosure relates to a video processing system, a video processing apparatus, and a video processing method.

BACKGROUND ART

[0002] Techniques in which videos obtained by imaging edge side terminals are transmitted to center side servers, and the center side servers recognize objects in the videos by AI engines have been developed. For example, the servers recognize types of work performed by workers. Here, the terminals on the edge side change frame rates of the videos by frame filtering or the like for efficient use of calculation resources and efficient use of network bands.

[0003] As a related technique, Patent Literature 1 discloses a technique of performing video scene recognition from time-series frames extracted from a video using a deep learning algorithm such as a recurrent neural network (RNN).

CITATION LIST

Patent Literature

[0004] Patent Literature 1: Japanese Unexamined Patent Application Publication No. 2018-005638

SUMMARY OF INVENTION

Technical Problem

[0005] In the technique of Patent Literature 1 or the like, since a server on a center side does not support a change in a frame rate of a video, object recognition supporting the change in the frame rate of the video cannot be performed, and thus there is room for improvement in recognition accuracy of an object in the video.

[0006] In view of such a problem, an object of the present disclosure is to provide a video processing system, a video processing apparatus, and a video processing method that can be expected to improve recognition accuracy of an object in a video.

Solution to Problem

[0007] According to an aspect of the present disclosure, a video processing system includes: [0008] video acquisition means for acquiring an input video; [0009] time difference information acquisition means for acquiring first time difference information between frames of the input video; and [0010] recognition means for inputting the input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference information between frames of the training video and recognizing an object in the input video.

[0011] According to another aspect of the present disclosure, a video processing apparatus includes: [0012] video acquisition means for acquiring an input video; [0013] time difference information acquisition means for acquiring first time difference information between frames of the input video; and [0014] recognition means for inputting the input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference information between frames of the training video and recognizing an object in the input video.

[0015] According to still another aspect of the present disclosure, a video processing method

includes: by a computer, [0016] acquiring an input video; [0017] acquiring first time difference information between frames of the input video; and [0018] inputting the input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference information between frames of the training video and recognizing an object in the input video.

Advantageous Effects of Invention

[0019] According to the present disclosure, it is possible to provide a video processing system, a video processing apparatus, and a video processing method that can be expected to improve recognition accuracy of an object in a video.

Description

BRIEF DESCRIPTION OF DRAWINGS

[0020] FIG. **1** is a block diagram illustrating a configuration of a video processing system according to an overview of an example embodiment.

[0021] FIG. **2** is a block diagram illustrating a configuration of a video processing apparatus according to an overview of an example embodiment.

[0022] FIG. **3** is a flowchart illustrating a video processing method according to an overview of the example embodiment.

[0023] FIG. **4** is a block diagram illustrating a configuration of a video processing system according to a first example embodiment.

[0024] FIG. **5** is a block diagram illustrating a configuration of a terminal according to the first example embodiment.

[0025] FIG. **6** is a block diagram illustrating a configuration of a center server according to the first example embodiment.

[0026] FIG. **7** is a flowchart illustrating an operation of the video processing system according to the first example embodiment.

[0027] FIG. **8** is a diagram illustrating an example of input information of a trained recognition model according to the first example embodiment.

[0028] FIG. **9** is a diagram illustrating an example of a configuration of a trained recognition model and a recognition operation according to the first example embodiment.

[0029] FIG. **10** is a block diagram illustrating a configuration of a center server according to a second example embodiment.

[0030] FIG. **11** is a flowchart illustrating an example of an operation of a video processing system according to the second example embodiment.

[0031] FIG. **12** is a diagram illustrating an example of a configuration of a trained recognition model and a recognition operation according to the second example embodiment.

[0032] FIG. **13** is a diagram illustrating an example of a first training operation of a recognition model according to the second example embodiment.

[0033] FIG. **14** is a diagram illustrating an example of a second training operation of the recognition model according to the second example embodiment.

[0034] FIG. **15** is a diagram illustrating another example of the configuration of the trained recognition model and the recognition operation according to the second example embodiment.

[0035] FIG. **16** is a block diagram illustrating a configuration of a computer according to the present example embodiment.

EXAMPLE EMBODIMENT

[0036] Hereinafter, example embodiments of the present disclosure will be described in detail with reference to the drawings. In the drawings, the same or corresponding elements are denoted by the same reference numerals and signs, and repeated description of the elements will be omitted to

clarify the description as necessary.

OVERVIEW OF EXAMPLE EMBODIMENT

[0037] First, a video processing system **10** according to an overview of an example embodiment will be described with reference to FIG. **1**. FIG. **1** is a block diagram illustrating a configuration of the video processing system **10** according to an overview of an example embodiment. The video processing system **10** is applicable to, for example, a remote monitoring system that collects videos via a network and recognizes the videos.

[0038] As illustrated in FIG. **1**, the video processing system **10** includes a video acquisition unit **11**, a time difference information acquisition unit **12**, and a recognition unit **13**. The video acquisition unit **11** acquires an input video. The time difference information acquisition unit **12** acquires first time difference information between frames of the input video. The recognition unit **13** inputs the input video and the first time difference information between the frames of the input video to the trained recognition model trained using the training video and second time difference information between the frames of the training video, and recognizes an object in the input video.

[0039] Next, a configuration of the video processing apparatus **20** according to the overview of the example embodiment will be described with reference to FIG. **2**. FIG. **2** is a block diagram illustrating a configuration of the video processing apparatus **20** according to the overview of the example embodiment. As illustrated in FIG. **2**, the video processing apparatus **20** includes the video acquisition unit **11**, the time difference information acquisition unit **12**, and the recognition unit **13** illustrated in FIG. **1**. When the video processing apparatus **20** is realized by edge computing, a part or all of the video processing apparatus **20** may be disposed on an edge or a cloud. For example, the video acquisition unit **11** and the time difference information acquisition unit **12** may be disposed in a terminal of the edge, and the recognition unit **13** may be disposed in a server of the cloud. Further, functions may be distributed and disposed in the cloud. The video processing apparatus **20** may be realized by a virtualization technique such as a virtualization server. A part or all of the video processing apparatus **20** may be disposed on a site side or a server side. An apparatus disposed at a site where the terminal is installed, an apparatus arranged at a place close to the site, or an apparatus disposed at a network hierarchy close to the terminal is defined as an apparatus disposed at the site side. An apparatus located away from the site is defined as an apparatus disposed on the center side. Since the apparatus disposed on the center side may be disposed on a cloud, the center side may be referred to as a cloud side.

[0040] Next, a video processing method according to an overview of the example embodiment will be described with reference to FIG. **3**. FIG. **3** is a flowchart illustrating a video processing method according to an overview of the example embodiment. For example, the video processing method according to the example embodiment is performed by the video processing system **10** in FIG. **1** or the video processing apparatus **20** in FIG. **2**.

[0041] As illustrated in FIG. **3**, the input video is acquired (step S**11**). Next, the first time difference information between the frames of the input video is acquired (step S**12**). Next, the input video and the first time difference information between the frames of the input video are input to the trained recognition model trained using the training video and the second time difference information between the frames of the training video, and an object in the input video is recognized (step S**13**).

[0042] As described above, the video processing system **10** can recognize the object corresponding to a change in a frame rate of the video by considering the time difference information between the frames of the input video. In the video processing system **10**, it can be expected that recognition accuracy of the object in the video will be improved.

(Basic Configuration of Video Processing System)

[0043] Next, a video processing system **1** which is an example of a system to which the example embodiment is applied will be described with reference to FIG. **4**. FIG. **4** is a block diagram illustrating a configuration of the video processing system **1** according to the first example embodiment. As illustrated in FIG. **4**, the video processing system **1** is a system that monitors a

captured area by a video captured by a camera. Hereinafter, in the present example embodiment, the video processing system **1** will be described below as a system that remotely monitors work of a worker at a site. For example, the site may be an area such as a work site such as a construction site, a square where people gather, or a school where people and machines operate. In the present example embodiment, hereinafter, the work will be described as construction work, civil engineering work, or the like, but the work is not limited thereto. Since the video includes time-series frames that are time-series images, the videos and the images are terms that can be used interchangeably. That is, the video processing system can be said to be a video processing system that processes a video and an image processing system that processes an image.

[0044] The video processing system **1** includes a plurality of terminals **100**, a center server **200**, a base station **300**, and a MEC **400**. The terminal **100**, the base station **300**, and the MEC **400** are disposed on the site side, and the center server **200** is disposed on the center side. For example, the center server **200** is disposed in a data center or the like disposed at a position away from the site. The site side is an edge side of the system, and the center side is also a cloud side.

[0045] The terminal **100** and the base station **300** are communicatively connected by a network NW**1**. The network NW**1** is, for example, a wireless network such as a 4G, local 5G/5G, long term evolution (LTE), or a wireless LAN network. The base station **300** and the center server **200** are communicatively connected by a network NW**2**. The network NW**2** includes, for example, a core network such as a 5th generation core network (5GC) or an evolved packet core (EPC), the Internet, or the like. It can also be said that the terminal **100** and the center server **200** are communicatively connected via the base station **300**. The base station **300** and the MEC **400** are communicatively connected by any communication method, but the base station **300** and the MEC **400** may be one apparatus.

[0046] The terminal **100** is a terminal apparatus connected to the network NW**1**, and is also a video generation apparatus that generates a video of a site. The terminal **100** acquires a video captured by the camera **101** installed at the site, and transmits the acquired video to the center server **200** via the base station **300**. The camera **101** may be disposed outside of the terminal **100** or inside the terminal **100**.

[0047] The terminal **100** compresses a video of the camera **101** at a predetermined bit rate and transmits the compressed video. The terminal **100** has a compression efficiency optimization function **102** of optimizing compression efficiency and a video delivery function **103**. The compression efficiency optimization function **102** performs region of interest (ROI) (also referred to as a gaze region) control to control quality of the ROI. The compression efficiency optimization function **102** reduces the bit rate by reducing image quality of a region around the ROI including a person or an object while maintaining the image quality of the ROI. The video delivery function **103** delivers a video of which the image quality is controlled to the center server **200**.

[0048] The base station **300** is a base station apparatus of the network NW**1**, and is also a relay apparatus that relays communication between the terminal **100** and the center server **200**. For example, the base station **300** is a local 5G base station, a 5G next generation node B (gNB), an LTE evolved node B (eNB), an access point of a wireless LAN, or the like, but may be another relay apparatus.

[0049] A multi-access edge computing (MEC) **400** is an edge processing apparatus disposed on an edge side of the system. The MEC **400** is an edge server that controls the terminal **100**, and has a compression bit rate control function **401** that control a bit rate of the termina and a terminal control function **402**. The compression bit rate control function **401** controls a bit rate of terminal **100** by adaptive video delivery control or quality of experience (QoE) control. For example, the compression bit rate control function **401** predicts the recognition accuracy to be obtained while curbing the bit rate according to a communication environment of the networks NW**1** and NW**2**, and allocates the bit rate to the camera **101** of each terminal **100** so that recognition accuracy is improved. The terminal control function **402** controls the terminal **100** so that a video of the

allocated bit rate is delivered. The terminal **100** encodes the video so that the video has the allocated bit rate, and delivers the encoded video.

[0050] The center server **200** is a server installed on the center side of the system. The center server **200** may be one or a plurality of physical servers, a cloud server constructed on a cloud, or another virtualization server. The center server **200** is a monitoring apparatus that monitors work of a site by recognizing work of a person from a camera image of the site. The center server **200** is also a video recognition apparatus that recognizes an action or the like of a person in the video transmitted from the terminal **100**.

[0051] The center server **200** has a video recognition function **201**, an alert generation function **202**, a GUI drawing function **203**, and a screen display function **204**. The video recognition function **201** recognizes work performed by the worker, that is, a type of action of the person, by inputting the video transmitted from the terminal **100** to an AI engine (for example, a trained recognition model). The alert generation function **202** generates an alert according to the recognized work. The GUI drawing function **203** displays a graphical user interface (GUI) on a screen of the display apparatus. The screen display function **204** displays a video, a recognition result, an alert, and the like of the terminal **100** on the GUI.

First Example Embodiment

[0052] First, a configuration of a video processing system **1** according to a first example embodiment will be described with reference to FIG. **4**. As illustrated in FIG. **4**, the video processing system **1** includes a plurality of terminals **100**, a center server **200**, a base station **300**, and a MEC **400**. The configuration of each apparatus is exemplary, and another configuration may be used as long as an operation according to the present example embodiment described below can be performed. For example, some functions of the terminal **100** may be disposed in the center server **200** or another apparatus, or some functions of the center server **200** may be disposed in the terminal **100** or another apparatus.

[0053] The video processing system **1** is a concrete implementation of a video processing system **10** according to the overview of the example embodiment. The center server **200** is a concrete implementation of the video processing apparatus **20** according to the overview of the example embodiment.

[0054] Next, a configuration of the terminal **100** of the video processing system **1** according to the first example embodiment will be described with reference to FIG. **5**. FIG. **5** is a block diagram illustrating a configuration of the terminal **100** of the video processing system **1** according to the first example embodiment. As illustrated in FIG. **5**, the terminal **100** includes a video acquisition unit **110**, a frame filtering unit **120**, an encoding unit **130**, and a terminal communication unit **140**.

[0055] The video acquisition unit **110** acquires a video (also referred to as an input video) captured by the camera **101**. The input video is, for example, data obtained by imaging a person who is a worker who performs work on a site, a work object used by the person, or the like. The input video includes time-series frames.

[0056] The frame filtering unit **120** filters (sorts) the time-series frames included in the input video. The frame filtering unit **120** performs filtering to adjust a bit rate of a video to be transmitted to the center server **200**, for example. Here, frames that are not filtered among the frames included in the input video are skipped.

[0057] The encoding unit **130** encodes the filtered input video. The encoding unit **130** changes a frame rate of the input video by filtering the frames.

[0058] The encoding unit **130** may encode the input video such that a gaze region of the frame has higher image quality than other regions. Specifically, the encoding unit **130** detects an object in the input video using a trained neural network model (for example, a model such as a convolutional neural network), and surrounds the detected object with a box. The encoding unit **130** may enclose the detected object not only in a box but also in a circle, an ellipse, an irregular shape suitable for a silhouette, or the like. Then, the encoding unit **130** recognizes the object inside the box. The

encoding unit **130** extracts an object of which a class is a person or a work object from the recognition objects, and determines the inside of the box of the extracted object as a gaze region. The encoding unit **130** encodes the input video such that the gaze region has higher image quality than other regions.

[0059] The terminal communication unit **140** transmits the encoded data to the center server **200**.

[0060] Next, a configuration of the center server **200** of the video processing system **1** according to the first example embodiment will be described with reference to FIG. **6**. FIG. **6** is a block diagram illustrating an example of a configuration of the center server **200** of the video processing system **1** according to the first example embodiment. As illustrated in FIG. **6**, the center server **200** includes a center communication unit **210**, a decoding unit **220**, a time difference information acquisition unit **230**, a recognition unit **240**, a storage unit **250**, and a training unit **260**.

[0061] The center communication unit **210** receives the encoded data transmitted from the terminal **100** via the base station **300**. The center communication unit **210** is an interface capable of communicating with the Internet or a core network, and is, for example, a wired interface for IP communication, but may be a wired or wireless interface of any other communication system.

[0062] The decoding unit **220** decodes the encoded data received from the terminal **100**. The decoding unit **220** corresponds to an encoding system of the terminal **100** and performs decoding in conformity with a moving image encoding system such as H.264 or H.265, for example. The decoding unit **220** decodes each region in the frames according to the compression rate to generate the decoded input video.

[0063] The time difference information acquisition unit **230** acquires time difference information ΔT (where ΔT is a natural number) based on time stamp information acquired from a video compression codec or the like. The time difference information ΔT corresponds to a predetermined frame included in the input video and is information indicating a time difference from a previous frame in the predetermined frame. That is, the time difference information ΔT is 1 when the frame is not skipped between the predetermined frame and the previous frame. On the other hand, the time difference information ΔT is 1+n when n frames are skipped between the predetermined frame and the previous frame. The time stamp information is information indicating a timing at which each frame included in the input video is captured by the camera **101**. The time stamp information may be information indicating a timing at which each frame is encoded by the encoding unit **130** of the terminal **100**.

[0064] The recognition unit **240** inputs the time series frames included in the input video and the time difference information ΔT between the frames of the input video as input information to the trained recognition model M**1**, and recognizes the object in the input video. The recognition unit **240** recognizes, for example, work performed by the worker in the input video, that is, a type of action of the person. Specifically, the trained recognition model M**1** is a model of a recurrent neural network (RNN) that inputs time-series frames included in the input video, and includes a plurality of cells of the RNN. The plurality of cells of the RNN input parameters corresponding to the time difference information between the frames of the input video. More specifically, the plurality of cells of the RNN input the parameters in which the time difference information between the frames of the input video is decoded by a decoder.

[0065] The storage unit **250** stores the trained recognition model M**1**.

[0066] The training unit **260** generates the trained recognition model M**1** by training using the training video, the time difference information ΔT between the frames of the training video, and the correct data.

[0067] Next, a recognition operation of the video processing system **1** according to the first example embodiment will be described with reference to FIGS. **7** to **9**.

[0068] FIG. **7** is a flowchart illustrating an operation of the video processing system **1** according to the first example embodiment. As illustrated in FIG. **7**, the video acquisition unit **110** of the terminal **100** of the video processing system **1** first acquires an input video obtained by imaging a

site from the camera **101** (step S**101**). The input video includes time-series frames.
[0069] The frame filtering unit **120** filters the time-series frames included in the input video (step S**102**). Here, frames that are not filtered among the frames included in the input video are skipped.
[0070] The encoding unit **130** encodes the filtered input video (step S**103**). Subsequently, the terminal communication unit **140** transmits the encoded data to the center server **200** via the base station **300** (step S**104**).
[0071] Subsequently, the center communication unit **210** of the center server **200** receives the encoded data from the terminal **100** (step S**105**). Subsequently, the decoding unit **220** decodes the encoded data to acquire the input video (step S**106**).
[0072] Subsequently, the time difference information acquisition unit **230** acquires time difference information ΔT between the frames corresponding to the frames of the input video (step S**107**). Specifically, the time difference information acquisition unit **230** acquires the time difference information ΔT based on the time stamp information acquired from a video compression codec or the like. The time stamp information is, for example, information regarding a timing at which each frame included in the input video is imaged by the camera **101**.
[0073] Subsequently, the recognition unit **240** inputs the time-series frames included in the input video and the time difference information ΔT corresponding to the frames of the input video as input information to the trained recognition model M**1** (step S**108**).
[0074] FIG. **8** is a diagram illustrating an example of input information input to the trained recognition model M**1**. As illustrated in FIG. **8**, the input information includes time-series frames included in the input video and the time difference information ΔT corresponding to the frames. The time difference information ΔT indicates a time difference from the previous frame in the corresponding predetermined frame. For example, the time difference information ΔT is 1 when no frame is skipped between the corresponding predetermined frame and the previous frame. The time difference information ΔT is 1+n when n frames are skipped between the corresponding predetermined frame and the previous frame.
[0075] The description returns to FIG. **7**. Subsequently, the recognition unit **240** recognizes an object in the input video by the trained recognition model M**1** (step S**109**). The recognition unit **240** recognizes, for example, work performed by the worker in the input video, that is, a type of action of the person.
[0076] FIG. **9** is a diagram illustrating an example of a configuration and a recognition operation of the trained recognition model M**1**. As illustrated in FIG. **9**, the trained recognition model M**1** is a model of a recurrent neural network (RNN) and includes a plurality of cells M**11** in a time series of the RNN. When a structure of the RNN is classified into an input layer, an intermediate layer, and an output layer, the cell M**11** corresponds to the intermediate layer of the RNN. Further, the trained recognition model M**1** includes a decoder M**12** corresponding to each cell M**11**.
[0077] At the predetermined time, the decoder M**12** receives an input of the time difference information ΔT and outputs a parameter with which the input time difference information ΔT is decoded to the cell M**11**. Subsequently, the cell M**11** receives an input of the frames, a state vector output by the cell M**11** at a previous time, and parameter information output by the decoder M**12**, and outputs the state vector to the cell M**11** at a subsequent time. In an initial state of the state vector input to the cell M**11**, for example, all elements may be 0.
[0078] For example, at time t, the decoder M**12** receives an input of **1** of the time difference information ΔT and outputs a parameter with which 1 of the time difference information ΔT is decoded to the cell M**11**. Here, since no frame skipping incurs between a frame input to the cell M**11** at time t-**1** and a frame input to the cell M**11** at time t, the time difference information ΔT input to the decoder M**12** at time t is 1. Then, at time t, the cell M**11** receives an input of the frames and the state vector and the parameter output by the cell M**11** at time t-**1**, and outputs the state vector to the cell M**11** at time t+1.
[0079] On the other hand, at time t+1, the decoder M**12** receives an input of **2** of the time

difference information ΔT, and outputs a parameter with which 2 of the time difference information ΔT is decoded to the cell M**11**. Here, since skipping of one frame incurs between the frame input to the cell M**11** at time t and the frame input to the cell M**11** at time t+1, the time difference information ΔT input to the decoder M**12** at time t+1 is 2. Then, at time t+1, the cell M**11** receives an input of the frames, the state vector, and the parameter output by the cell M**11** at time t, and outputs the state vector to the cell M**11** at time t+2.

[0080] Next, a training operation of the recognition model M**1** in the video processing system **1** according to the first example embodiment will be described.

[0081] The training unit **260** inputs the time-series frames included in the training video and the time difference information ΔT corresponding to the frames to the recognition model M**1**. The training video includes, for example, time-series frames in which frame skipping incurs by a predetermined pattern. The configuration of the recognition model M**1** has been described above (see FIG. **9**). The training unit **260** trains the recognition model M**1** by comparing the output result by the recognition model M**1** with the correct data, and generates the trained recognition model M**1**.

[0082] As described above, the trained recognition model M**1** of the video processing system **1** decodes the time difference information ΔT and dynamically determines the parameter to be input to the cell M**11**. That is, the trained recognition model M**1** can improve the recognition accuracy of the object by reflecting the time difference information of the frames in the recognition of the object in consideration of a case where the frame is skipped due to a change in the frame rate of the video or the like.

Second Example Embodiment

[0083] Next, a configuration of a video processing system **2** according to a second example embodiment will be described. The video processing system **2** includes a plurality of terminals **100**, a center server **200**, a base station **300**, and a MEC **400**, similarly to the video processing system **1** according to the first example embodiment. Here, the center server **200** of the video processing system **2** is different from the center server **200** of the video processing system **1** in the following configuration.

[0084] FIG. **10** is a diagram illustrating a configuration of the center server **200** of the video processing system **2**. As illustrated in FIG. **10**, the center server **200** of the video processing system **2** includes a center communication unit **210**, a decoding unit **220**, a time difference information acquisition unit **230**, a recognition unit **270**, a storage unit **280**, and a training unit **290**.

[0085] The recognition unit **270** inputs the time series frames included in the input video and the time difference information ΔT between the frames of the input video as input information to a trained recognition model M**2**, and recognizes the object in the input video. The trained recognition model M**2** includes a plurality of cells of a recurrent neural network (RNN) that inputs time-series frames included in the input video, and inputs and outputs a state vector chronologically. The trained recognition model M**2** inserts a state predictor that predicts a state vector based on an inter-frame time difference information ΔT between predetermined cells such as between cells in which frame skipping has incurred.

[0086] The storage unit **280** stores the recognition model M**2**.

[0087] The training unit **290** trains the trained recognition model M**2** into which the state predictor is inserted using the correct data and the time-series frames in which frame skipping has incurred in a predetermined pattern included in the training video, and the time difference information between the frames of the training video.

[0088] The training unit **290** trains a plurality of cells of the trained recognition model M**2** using the correct data and the time-series frames which are included in the training video and in which no frame skipping incurs. Then, the training unit **290** inputs the time-series frames which are included in the training video and in which no frame skipping incurs to the plurality of cells of the trained recognition model M**2**. In this case, the training unit **290** trains the state predictor using the state

vector output at time t (where t is a natural number) and the state vector output at time t+N (where N is a natural number) by a plurality of cells.
[0089] On the other hand, the recognition unit **270** inputs the input video, the time difference information between the frames of the input video, and a motion between the frames of the input video to the trained recognition model trained using the training video, the time difference information between the frames of the training video, and the motion between the frames of the input video, and recognizes an object in the input video.
[0090] The training unit **290** trains the trained recognition model M**2** into which the state predictor is inserted using the time difference information between the time-series frames in which the frame skipping incurs in the predetermined pattern included in the training video, the frames of the training video, and the motion between the frames of the input video, and the correct data.
[0091] Next, a recognition operation of the video processing system **2** according to the second example embodiment will be described with reference to FIG. **11**.
[0092] FIG. **11** is a flowchart illustrating an example of an operation of the video processing system **2** according to the second example embodiment. As illustrated in FIG. **11**, the video processing system **2** first performs the process of step S**101** to the process of step S**107** (see FIG. **7**) described above. Description of the process of step S**101** to the process of step S**107** will be omitted.
[0093] Subsequently, the recognition unit **270** of the center server **200** inputs the time-series frames included in the input video and the time difference information ΔT corresponding to the frames as input information to the trained recognition model M**2** (step S**201**). An example of the input information has been described above (see FIG. **8**). In the present example embodiment, however, the recognition unit **240** sets the time difference information ΔT of ΔT #**1** as input information to the trained recognition model M**2**.
[0094] Subsequently, the recognition unit **270** recognizes the object in the input video by the trained recognition model M**2** (step S**202**). The recognition unit **240** recognizes, for example, work performed by the worker in the input video, that is, a type of action of the person.
[0095] FIG. **12** is a diagram illustrating an example of a configuration and a recognition operation of the trained recognition model M**2** according to the second example embodiment.
[0096] As illustrated in FIG. **12**, the trained recognition model M**2** is a recurrent neural network (RNN) and includes a plurality of cells M**21** of a time-series RNN. At a predetermined time, the cell M**21** receives an input of the frame and the state vector output by the cell M**21** at the previous time, and outputs the state vector to the cell M**21** at the subsequent time. In the initial state of the state vector, for example, elements may all be **0**.
[0097] Further, when frame skipping incurs between a frame input to the cell M**21** at a predetermined time and a frame input to the cell M**21** at a previous time of the cell M**21** at the predetermined time, the trained recognition model M**2** inserts the state predictor M**22** between the cell M**21** at the predetermined time and the cell M**21** at the previous time. The incurrence of the frame skipping can be determined from the time difference information ΔT corresponding to the frame input to the cell M**21** at the predetermined time. The inserted state predictor M**22** receives an input of the state vector output by the cell M**21** at the previous time and the time difference information ΔT corresponding to the frame input to the cell M**21** at the predetermined time. Then, the state predictor M**22** predicts a state vector and outputs the predicted state vector to the cell M**21** at the predetermined time.
[0098] For example, frame skipping incurs between a frame input to the cell M**21** at time t+1 and a frame input to the cell M**21** at time t. In this case, the trained recognition model M**2** inserts the state predictor M**22** between the cell M**21** at time t+1 and the cell M**21** at time t. The state predictor M**22** receives an input of the state vector output by the cell M**21** at time t and **2** of the time difference information ΔT corresponding to the frame input to the cell M**21** at time t+1. Here, the input time difference information ΔT is 2 since one-frame skipping incurs between the frame input to the cell

M**21** at time t+1 and the frame input to the cell M**21** at time t. The state predictor M**22** predicts a state vector and outputs the predicted state vector to the cell M**21** at time t+1.

[0099] Next, a training operation of the recognition model M**2** of the video processing system **2** according to the second example embodiment will be described with reference to FIGS. **13** and **14**.

[0100] FIG. **13** is a diagram illustrating an example of the first training operation of the recognition model M**2**.

[0101] As illustrated in FIG. **13**, the training unit **290** inputs the time-series frames included in the training video and the time difference information ΔT (ΔT #**1**) corresponding to the frames to the recognition model M**2**. Specifically, the training unit **290** inputs the time-series frames included in the training video to the plurality of cells M**21** of the recognition model M**2**. In the input time-series frames, the frame skipping incurs by a predetermined pattern. Further, when the frame skipping incurs between a frame input to the cell M**21** at the predetermined time of the recognition model M**2** and a frame input to the cell M**21** at a time previous to the predetermined time, the training unit **290** inserts the state predictor M**22** between the cell M**21** at the predetermined time and the cell M**21** at the previous time. The training unit **290** inputs the time difference information ΔT corresponding to the frame input to the cell M**21** at the predetermined time to the state predictor M**22**. For example, one-frame skipping incurs between the frame input to the predetermined cell M**21** at time t+1 and the frame input to the cell M**21** at time t. In this case, the training unit **290** inserts the state predictor M**22** between the cell M**21** at time t+1 and the cell M**21** at time t, and inputs **2** of the time difference information ΔT corresponding to the frame input to the predetermined cell M**21** at time t+1.

[0102] Then, the training unit **290** trains the recognition model M**2** into which the state predictor M**22** is inserted by comparing an output result by the recognition model M**2** into which the state predictor M**22** is inserted with the correct data. The training unit **290** may separate the training of the state predictor M**22** from the training of the recognition model M**2**.

[0103] FIG. **14** is a diagram illustrating an example of the second training operation of the recognition model M**2**.

[0104] As illustrated in FIG. **14**, the training unit **290** inputs the time-series frames included in the training video to the plurality of cells M**21** of the recognition model M**2**. In the input time-series frames, the frame skipping does not incur. Then, the training unit **290** trains the recognition model M**2** by comparing the output result by the recognition model M**2** with the correct data.

[0105] Subsequently, the training unit **290** inputs the time-series frames included in the training video to the plurality of cells M**21** of the trained recognition model M**2**. In the input time-series frames, the frame skipping does not incur.

[0106] Subsequently, the training unit **290** acquires a data set including the state vector output from the cell M**21** at time t and the state vector output from the cell M**21** at time t+N (where N is a natural number), and trains the state predictor M**22** using the acquired data set as training data. Specifically, the training unit **290** trains the state predictor M**22** by performing regression analysis so that the output result at the time of inputting of the state vector at time t and N to the state predictor M**22** approaches a state vector at time t+N.

[0107] FIG. **15** is a diagram illustrating another example of the recognition operation of the trained recognition model M**2** according to the second example embodiment.

[0108] As illustrated in FIG. **15**, the trained recognition model M**2** is a recurrent neural network (RNN) and includes the plurality of cells M**21** of the time-series RNN. At a predetermined time, the cell M**21** receives an input of the frame and the state vector output by the cell M**21** at the previous time, and outputs the state vector to the cell M**21** at the subsequent time. In the initial state of the state vector, for example, elements may all be **0**.

[0109] Further, when the frame skipping incurs between the frame input to the cell M**21** at the predetermined time and the frame input to the cell M**21** at the previous time of the cell M**21** at the predetermined time, the trained recognition model M**2** inserts a state predictor M**23** between the

predetermined cell M**21** and the cell M**21** at the previous time. The state predictor M**23** receives an input of the state vector output by the cell M**21** at the previous time, the time difference information ΔT corresponding to the frame input to the cell M**21** at the predetermined time, and a motion vector. The motion vector is information obtained by vectorizing a difference between the frame at the predetermined time and the frame at the previous time, that is, a motion. The state predictor M**23** predicts a state vector and outputs the predicted state vector to the cell M**21** at the predetermined time.

[0110] For example, frame skipping incurs between a frame input to the cell M**21** at time t+1 and a frame input to the cell M**21** at time t. In this case, the trained recognition model M**2** inserts the state predictor M**23** between the cell M**21** at time t+1 and the cell M**21** at time t. The state predictor M**23** receives an input of the motion vector together with the state vector output by the cell M**21** at time t and **2** of the time difference information ΔT corresponding to the frame input to the cell M**21** at time t+1. The input motion vector indicates a difference between the frame input to the cell M**21** at time t and the frame input to the cell M**21** at time t+1, that is, a motion. The state predictor M**23** predicts a state vector and outputs the predicted state vector to the cell M**21** at time t+1.

[0111] Next, a training operation of the recognition model M**2** in the video processing system **2** will be described.

[0112] The training unit **290** inputs the time-series frames included in the training video, the time difference information ΔT (ΔT #**1**) corresponding to the frames, and the motion vector to the recognition model M**2**. The training unit **290** trains the recognition model M**2** into which the state predictor M**22** is inserted by comparing an output result of the recognition model M**2** into which the state predictor M**22** is inserted with the correct data. The training unit **290** may separate the training of the state predictor M**22** from the training of the recognition model M**2**.

[0113] As described above, when the frame skipping incurs, the trained recognition model M**2** of the video processing system **2** according to the second example embodiment inserts the state predictor M**22** or the state predictor M**23** between the cells M**21** and predicts the state vector. The trained recognition model M**1** can improve the recognition accuracy of the object by reflecting the time difference information of the frame in the recognition of the object in consideration of a case where the frame is skipped due to a change in the frame rate of the video or the like.

[0114] Each configuration in the above-described example embodiments may be implemented by hardware, software, or both, and may be implemented by one piece of hardware or software or by a plurality of pieces of hardware or software. Each apparatus and each function (processing) may be realized by a computer **1000** including a processor **1001** such as a central processing unit (CPU) and a memory **1002** which is a storage device as illustrated in FIG. **19**. For example, a program that performs the method (video processing method) in the example embodiment may be stored in the memory **1002**, and each function may be realized by the processor **1001** executing the program stored in the memory **1002**.

[0115] The program includes a group of instructions (or software codes) causing a computer to perform one or more of the functions described in the example embodiments when the program is read by the computer. The program may be stored in a non-transitory computer-readable medium or a tangible storage medium. As an example and not by way of limitation, the computer-readable medium or the tangible storage medium includes a random-access memory (RAM), a read-only memory (ROM), a flash memory, a solid-state drive (SSD) or any other memory technique, a CD-ROM, a digital versatile disc (DVD), a Blu-ray (registered trademark) disc or any other optical disc storage, a magnetic cassette, a magnetic tape, and a magnetic disk storage or any other magnetic storage device. The program may be transmitted on a transitory computer-readable medium or a communication medium. As an example and not by way of limitation, transitory computer-readable or communication media include electrical, optical, and acoustic propagated signals or other forms of propagated signals.

[0116] Some or all of the above-described example embodiments may be described as in the

following Supplementary Notes, but are not limited to the following Supplementary Notes.
(Supplementary Note 1)
[0117] A video processing system including: [0118] video acquisition means for acquiring an input video; [0119] time difference information acquisition means for acquiring first time difference information between frames of the input video; and [0120] recognition means for inputting the input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference information between frames of the training video and recognizing an object in the input video.
(Supplementary Note 2)
[0121] The video processing system according to Supplementary Note 1, wherein [0122] the trained recognition model is a model including a plurality of cells of a recurrent neural network (RNN) that inputs time-series frames included in the input video, and [0123] the plurality of cells input a parameter corresponding to first time difference information between the frames of the input video.
(Supplementary Note 3)
[0124] The video processing system according to Supplementary Note 2, wherein the plurality of cells input a parameter obtained by decoding the first time difference information between the frames of the input video.
(Supplementary Note 4)
[0125] The video processing system according to Supplementary Note 1, wherein [0126] the trained recognition model includes a plurality of cells of a recurrent neural network that inputs time-series frames included in the input video, and inputs and outputs state vectors chronologically, and [0127] a state predictor that predicts the state vectors based on the first time difference information between the frames of the input video is inserted between predetermined cells.
(Supplementary Note 5)
[0128] The video processing system according to Supplementary Note 4, wherein the trained recognition model into which the state predictor is inserted is trained using time-series frames in which frame skipping incurs in a predetermined pattern included in the training video, the second time difference information between the frames of the training video, and correct data.
(Supplementary Note 6)
[0129] The video processing system according to Supplementary Note 4, wherein [0130] the plurality of cells of the trained recognition model are trained using time-series frames which are included in the training video and in which no frame skipping incurs and correct data, and [0131] the state predictor inserted into the trained recognition model is trained using a state vector output at time t (where t is natural number) and a state vector output at time t+N (where N is a natural number) by the plurality of cells when time-series frames which are included in the training video and in which no frame skipping incurs are input to the plurality of trained cells.
(Supplementary Note 7)
[0132] The video processing system according to Supplementary Note 1, wherein the recognition means inputs the input video, the first time difference information between the frames of the input video, and a motion between frames of the input video to a trained recognition model trained by using the training video, the second time difference information between the frames of the training video, and the motion between the frames of the training video, and recognizes an object in the input video.
(Supplementary Note 8)
[0133] A video processing apparatus including: [0134] video acquisition means for acquiring an input video; [0135] time difference information acquisition means for acquiring first time difference information between frames of the input video; and [0136] recognition means for inputting the input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference

information between frames of the training video and recognizing an object in the input video.
(Supplementary Note 9)
[0137] The video processing apparatus according to Supplementary Note 8, wherein [0138] the trained recognition model is a model including a plurality of cells of a recurrent neural network (RNN) that inputs time-series frames included in the input video, and [0139] the plurality of cells input a parameter corresponding to first time difference information between the frames of the input video.
(Supplementary Note 10)
[0140] The video processing apparatus according to Supplementary Note 9, wherein the plurality of cells input a parameter obtained by decoding the first time difference information between the frames of the input video.
(Supplementary Note 11)
[0141] The video processing apparatus according to Supplementary Note 8, wherein [0142] the trained recognition model includes a plurality of cells of a recurrent neural network that inputs time-series frames included in the input video, and inputs and outputs state vectors chronologically, and [0143] a state predictor that predicts the state vectors based on the first time difference information between the frames of the input video is inserted between predetermined cells.
(Supplementary Note 12)
[0144] The video processing apparatus according to Supplementary Note 11, wherein the trained recognition model into which the state predictor is inserted is trained using time-series frames in which frame skipping incurs in a predetermined pattern included in the training video, the second time difference information between the frames of the training video, and correct data.
(Supplementary Note 13)
[0145] The video processing apparatus according to Supplementary Note 11, wherein [0146] the plurality of cells of the trained recognition model are trained using time-series frames which are included in the training video and in which no frame skipping incurs and correct data, and [0147] the state predictor inserted into the trained recognition model is trained using a state vector output at time t (where t is natural number) and a state vector output at time t+N (where N is a natural number) by the plurality of cells when time-series frames which are included in the training video and in which no frame skipping incurs are input to the plurality of trained cells.
(Supplementary Note 14)
[0148] The video processing apparatus according to Supplementary Note 8, wherein the recognition means inputs the input video, the first time difference information between the frames of the input video, and a motion between frames of the input video to a trained recognition model trained by using the training video, the second time difference information between the frames of the training video, and the motion between the frames of the training video, and recognizes an object in the input video.
(Supplementary Note 15)
[0149] A video processing method comprising: by a computer, [0150] acquiring an input video; [0151] acquiring first time difference information between frames of the input video; and [0152] inputting the input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference information between frames of the training video and recognizing an object in the input video.
(Supplementary Note 16)
[0153] The video processing method according to Supplementary Note 15, wherein [0154] the trained recognition model is a model including a plurality of cells of a recurrent neural network (RNN) that inputs time-series frames included in the input video, and [0155] the plurality of cells input a parameter corresponding to first time difference information between the frames of the input video.
(Supplementary Note 17)

[0156] The video processing method according to Supplementary Note 16, wherein the plurality of cells input a parameter obtained by decoding the first time difference information between the frames of the input video.
(Supplementary Note 18)
[0157] The video processing method according to Supplementary Note 15, wherein [0158] the trained recognition model includes a plurality of cells of a recurrent neural network that inputs time-series frames included in the input video, and inputs and outputs state vectors chronologically, and [0159] a state predictor that predicts the state vectors based on the first time difference information between the frames of the input video is inserted between predetermined cells.
(Supplementary Note 19)
[0160] The video processing method according to Supplementary Note 18, wherein the trained recognition model into which the state predictor is inserted is trained using time-series frames in which frame skipping incurs in a predetermined pattern included in the training video, the second time difference information between the frames of the training video, and correct data.
(Supplementary Note 20)
[0161] The video processing method according to Supplementary Note 18, wherein [0162] the plurality of cells of the trained recognition model are trained using time-series frames which are included in the training video and in which no frame skipping incurs and correct data, and [0163] the state predictor inserted into the trained recognition model is trained using a state vector output at time t (where t is natural number) and a state vector output at time t+N (where N is a natural number) by the plurality of cells when time-series frames which are included in the training video and in which no frame skipping incurs are input to the plurality of trained cells.
(Supplementary Note 21)
[0164] The video processing method according to Supplementary Note 15, wherein the computer inputs the input video, the first time difference information between the frames of the input video, and a motion between frames of the input video to a trained recognition model trained by using the training video, the second time difference information between the frames of the training video, and the motion between the frames of the training video, and recognizes an object in the input video.
REFERENCE SIGNS LIST
[0165] **1**, **2**, **10** VIDEO PROCESSING SYSTEM [0166] **11** VIDEO ACQUISITION UNIT [0167] **12** TIME DIFFERENCE INFORMATION ACQUISITION UNIT [0168] **13** RECOGNITION UNIT [0169] **20** VIDEO PROCESSING APPARATUS [0170] **100** TERMINAL [0171] **101** CAMERA [0172] **102** COMPRESSION EFFICIENCY OPTIMIZATION FUNCTION [0173] **110** VIDEO ACQUISITION UNIT [0174] **120** FRAME FILTER UNIT [0175] **130** ENCODING UNIT [0176] **140** TERMINAL COMMUNICATION UNIT [0177] **200** CENTER SERVER [0178] **201** VIDEO RECOGNITION FUNCTION [0179] **202** ALERT GENERATION FUNCTION [0180] **203** GUI DRAWING FUNCTION [0181] **204** SCREEN DISPLAY FUNCTION [0182] **210** CENTER COMMUNICATION UNIT [0183] **220** DECODING UNIT [0184] **230** TIME DIFFERENCE INFORMATION ACQUISITION UNIT [0185] **240**, **270** RECOGNITION UNIT [0186] **250**, **280** STORAGE UNIT [0187] **260**, **290** TRAINING UNIT [0188] **300** BASE STATION [0189] **401** COMPRESSION BIT RATE CONTROL FUNCTION [0190] **1000** COMPUTER [0191] **1001** PROCESSOR [0192] **1002** MEMORY [0193] M**1**, M**2** RECOGNITION MODEL [0194] M**11**, M**21** CELL [0195] M**12** DECODER [0196] M**22**, M**23** STATE PREDICTOR

Claims

1. A video processing system comprising: at least one memory storing instructions, and at least one processor configured to execute the instructions to: acquire an input video; acquire first time

difference information between frames of the input video; and input the input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference information between frames of the training video and recognizing an object in the input video.

2. The video processing system according to claim 1, wherein the trained recognition model is a model including a plurality of cells of a recurrent neural network (RNN) that inputs time-series frames included in the input video, and the plurality of cells input a parameter corresponding to first time difference information between the frames of the input video.

3. The video processing system according to claim 1, wherein the trained recognition model includes a plurality of cells of a recurrent neural network that input time-series frames included in the input video, and input and output state vectors chronologically, and a state predictor that predicts the state vectors based on the first time difference information between the frames of the input video is inserted between predetermined cells.

4. The video processing system according to claim 3, wherein the trained recognition model into which the state predictor is inserted is trained using time-series frames in which frame skipping incurs in a predetermined pattern included in the training video, the second time difference information between the frames of the training video, and correct data.

5. The video processing system according to claim 3, wherein the plurality of cells of the trained recognition model are trained using time-series frames which are included in the training video and in which no frame skipping incurs and correct data, and the state predictor inserted into the trained recognition model is trained using a state vector output at time t (where t is natural number) and a state vector output at time t+N (where N is a natural number) by the plurality of cells when time-series frames which are included in the training video and in which no frame skipping incurs are input to the plurality of trained cells.

6. The video processing system according to claim 1, wherein the recognition means inputs the input video, the first time difference information between the frames of the input video, and a motion between frames of the input video to a trained recognition model trained by using the training video, the second time difference information between the frames of the training video, and the motion between the frames of the training video, and recognizes an object in the input video.

7. A video processing apparatus comprising: at least one memory storing instructions, and at least one processor configured to execute the instructions to: acquire an input video; acquire first time difference information between frames of the input video; and input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference information between frames of the training video and recognizing an object in the input video.

8. The video processing apparatus according to claim 7, wherein the trained recognition model is a model including a plurality of cells of a recurrent neural network (RNN) that inputs time-series frames included in the input video, and the plurality of cells input a parameter corresponding to first time difference information between the frames of the input video.

9. The video processing apparatus according to claim 7, wherein the trained recognition model includes a plurality of cells of a recurrent neural network that input time-series frames included in the input video, and input and output state vectors chronologically, and a state predictor that predicts the state vectors based on the first time difference information between the frames of the input video is inserted between predetermined cells.

10. The video processing apparatus according to claim 9, wherein the trained recognition model into which the state predictor is inserted is trained using time-series frames in which frame skipping incurs in a predetermined pattern included in the training video, the second time difference information between the frames of the training video, and correct data.

11. The video processing apparatus according to claim 9, wherein the plurality of cells of the

trained recognition model are trained using time-series frames which are included in the training video and in which no frame skipping incurs and correct data, and the state predictor inserted into the trained recognition model is trained using a state vector output at time t (where t is natural number) and a state vector output at time t+N (where N is a natural number) by the plurality of cells when time-series frames which are included in the training video and in which no frame skipping incurs are input to the plurality of trained cells.

12. The video processing apparatus according to claim 7, wherein the recognition means inputs the input video, the first time difference information between the frames of the input video, and a motion between frames of the input video to a trained recognition model trained by using the training video, the second time difference information between the frames of the training video, and the motion between the frames of the training video, and recognizes an object in the input video.

13. A video processing method comprising: by a computer, acquiring an input video; acquiring first time difference information between frames of the input video; and inputting the input video and the first time difference information between the frames of the input video to a trained recognition model trained using a training video and second time difference information between frames of the training video and recognizing an object in the input video.

14. The video processing method according to claim 13, wherein the trained recognition model is a model including a plurality of cells of a recurrent neural network (RNN) that inputs time-series frames included in the input video, and the plurality of cells input a parameter corresponding to first time difference information between the frames of the input video.

15. The video processing method according to claim 13, wherein the trained recognition model includes a plurality of cells of a recurrent neural network that input time-series frames included in the input video, and input and output state vectors chronologically, and a state predictor that predicts the state vectors based on the first time difference information between the frames of the input video is inserted between predetermined cells.

16. The video processing method according to claim 15, wherein the trained recognition model into which the state predictor is inserted is trained using time-series frames in which frame skipping incurs in a predetermined pattern included in the training video, the second time difference information between the frames of the training video, and correct data.

17. The video processing method according to claim 15, wherein the plurality of cells of the trained recognition model are trained using time-series frames which are included in the training video and in which no frame skipping incurs and correct data, and the state predictor inserted into the trained recognition model is trained using a state vector output at time t (where t is natural number) and a state vector output at time t+N (where N is a natural number) by the plurality of cells when time-series frames which are included in the training video and in which no frame skipping incurs are input to the plurality of trained cells.

18. The video processing method according to claim 13, wherein the computer inputs the input video, the first time difference information between the frames of the input video, and a motion between frames of the input video to a trained recognition model trained by using the training video, the second time difference information between the frames of the training video, and the motion between the frames of the training video, and recognizes an object in the input video.