Human-behavior-understanding Engine: Video-recognition AI Library for Understanding Human Behavior

Motohiro Takagi

Abstract

To enable all many types of work to be conducted remotely as needed and overcome geographical constraints, NTT Human Informatics Laboratories is developing telepresence technology for understanding the conditions at a remote location and remotely operate an artificial body, such as a robot, in real time. Remote operation of an artificial body requires that conditions at a remote location be analyzed and information on operating the artificial body and on people and the environment in the vicinity of the artificial body be fed back to the operator in real time. This article introduces the development of a human-behavior-understanding engine for automatically recognizing human behavior from camera video to recognize how people at a remote location might behave. Real-time recognition processing is vital to providing a remote operator with feedback in real time. For this engine, we developed technology for achieving lightweight recognition processing, enabling the behavior of multiple people to be recognized in real time on a central-processing-unit-based machine.

Keywords: remote world, telepresence, video recognition

1. Telepresence technology

Due to the COVID-19 pandemic in 2020, the need arose for technology to enable various types of remote work and overcome geographical and temporal constraints. Essential work, such as construction and care-giving/nursing, often requires on-site human physical operations, which makes remote work difficult to achieve. One method of enabling such work to be conducted remotely is to remotely operate an artificial body, such as a robot, located at a remote site. In remote work involving the operation of an artificial body, the operator must be able to recognize the state of the artificial body at the remote site and the peripheral environment in real time and carry out operations in a seamless, comfortable manner. At NTT Human Informatics Laboratories, our goal is to develop telepresence technology that enables conditions at a remote site to be recognized in real time and an artificial body to be remotely operated in a seamless and efficient manner [1]. For such technology to be effective, there is a need for recognition technology that can recognize the behavior of people and the peripheral environment at the remote site at low cost on the basis of video data obtained from sensors such as robot-mounted cameras or surveillance cameras. This article introduces a human-behavior-understanding engine NTT Human Informatics Laboratories developed for recognizing the behavior of people and the peripheral environment at a remote site by using camera video. This engine enables lightweight recognition processing and can recognize the behavior of multiple people in real time on a central processing unit (CPU)-based machine. It can also reduce the costs of generating training data and of constructing a machine environment, which are often barriers to introduction.

2. Human-behavior-understanding engine featuring low-cost introduction

There has been growing interest in digital

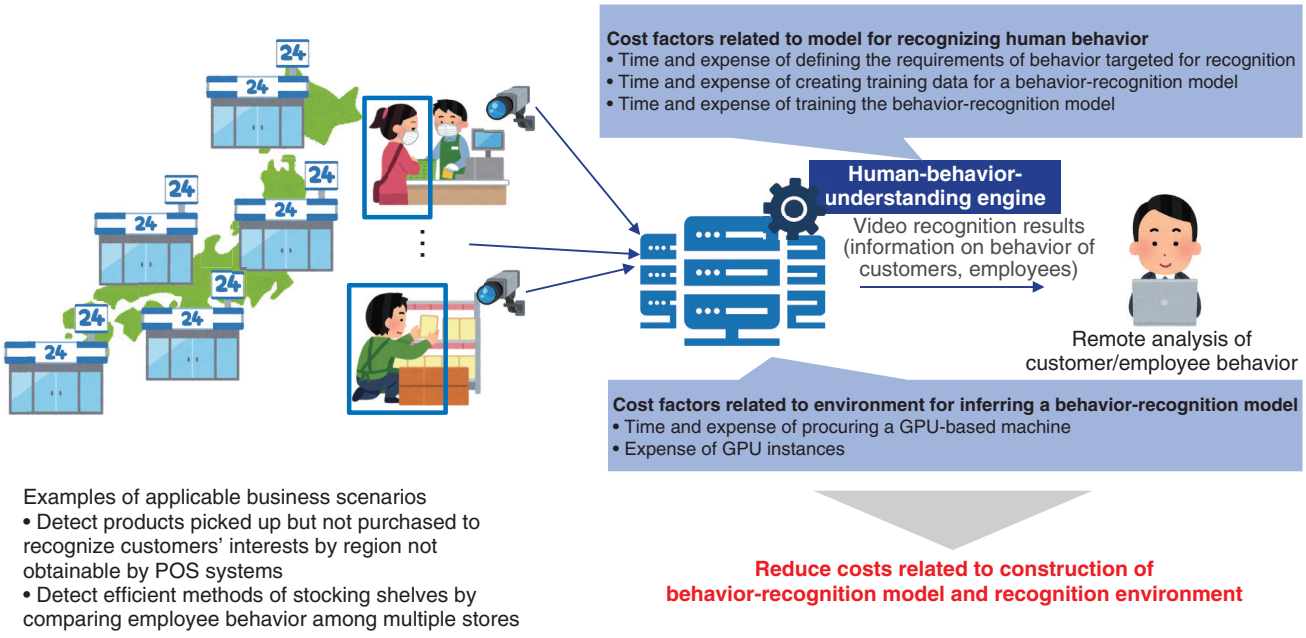


Fig. 1. Costs of introducing video AI for remote monitoring, etc.

transformation, particularly in technology that can automatically analyze the behavior of people (such as customers and employees) from video obtained from retail stores, factories, or other sites (**Fig. 1**). Recognizing the behavior of people appearing in video makes it possible to, for example, detect which products people pick up but do not purchase at a retail store. Thus, the interests of customers, which cannot be obtained through point-of-sale (POS) systems, can be monitored remotely by video and needs can be analyzed by tabulating results on a regional basis. Such remote monitoring could also enable the analysis of efficient behavior in stocking shelves, for example, by detecting inefficient actions of employees and comparing employee behavior among multiple stores. This type of analysis is expected to improve customer satisfaction and employee productivity.

However, the cost of preparing the training data needed for creating a model for recognizing human behavior (behavior-recognition model) is high, which becomes a barrier to introducing such a model. Preparing training data requires the acquisition of a large amount of video, extraction of behavior targeted for recognition by visual checking, as well as the work of labeling and annotating that video. Recognition processing often also requires the use of a graphics processing unit (GPU)-based machine, but the cost of constructing such a machine environment becomes an impediment to introduction.

At NTT Human Informatics Laboratories, we have addressed the above issues by developing a human-behavior-understanding engine that can shorten the time needed for generating training data and reduce associated expenses by presetting a pre-trained behavior-recognition model. This engine can also reduce the cost of constructing a machine environment by enabling operation not only using GPUs but also CPUs.

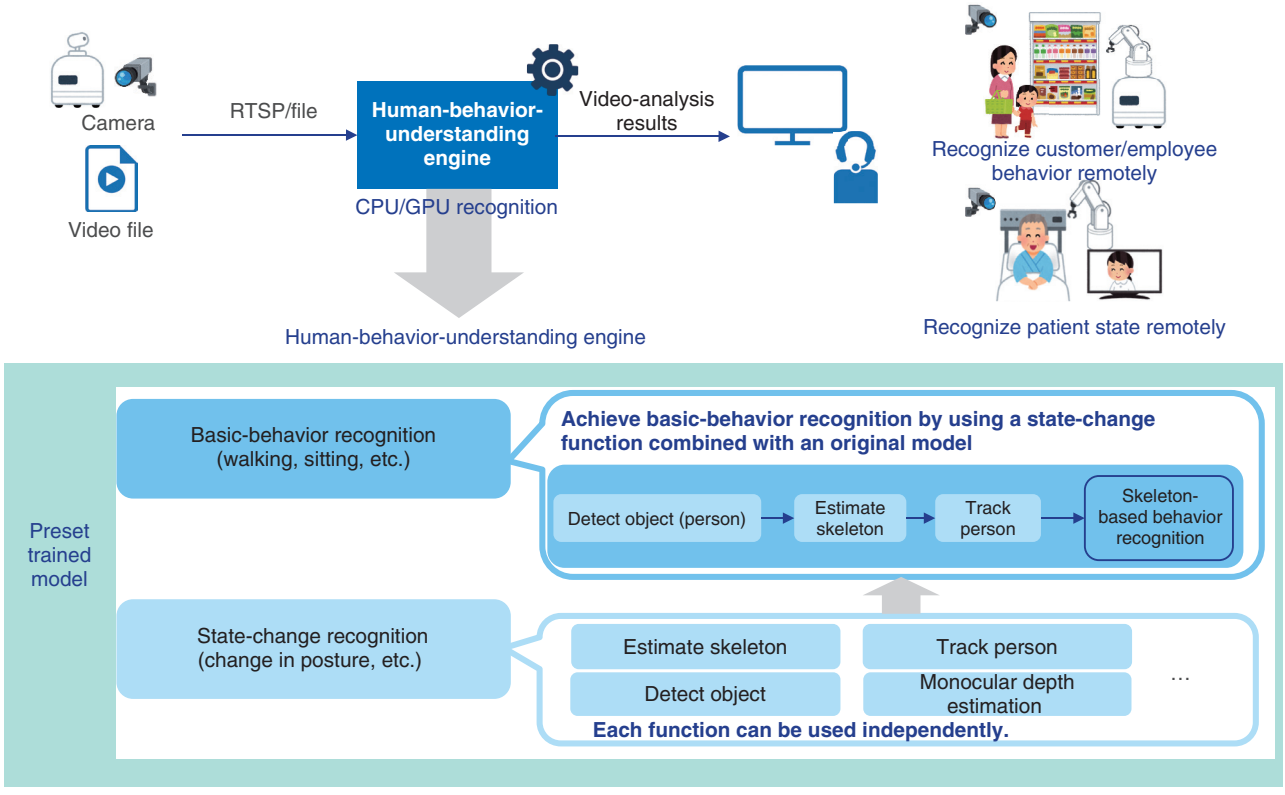
The following are the two key points of this human-behavior-understanding engine.

- Point 1: Presetting a pre-trained basic-behavior-recognition model makes additional training unnecessary and enables recognition/searching of behavior.
- Point 2: Applying lightweight recognition processing within the convolutional neural network (CNN) used in the basic-behavior-recognition model enables operation in an inexpensive machine environment on par with CPU-based machines.

These two points are explained below.

2.1 Basic-behavior recognition and behavior-searching function

Our human-behavior-understanding engine



RTSP: Real Time Streaming Protocol

Fig. 2. Human-behavior-understanding engine.

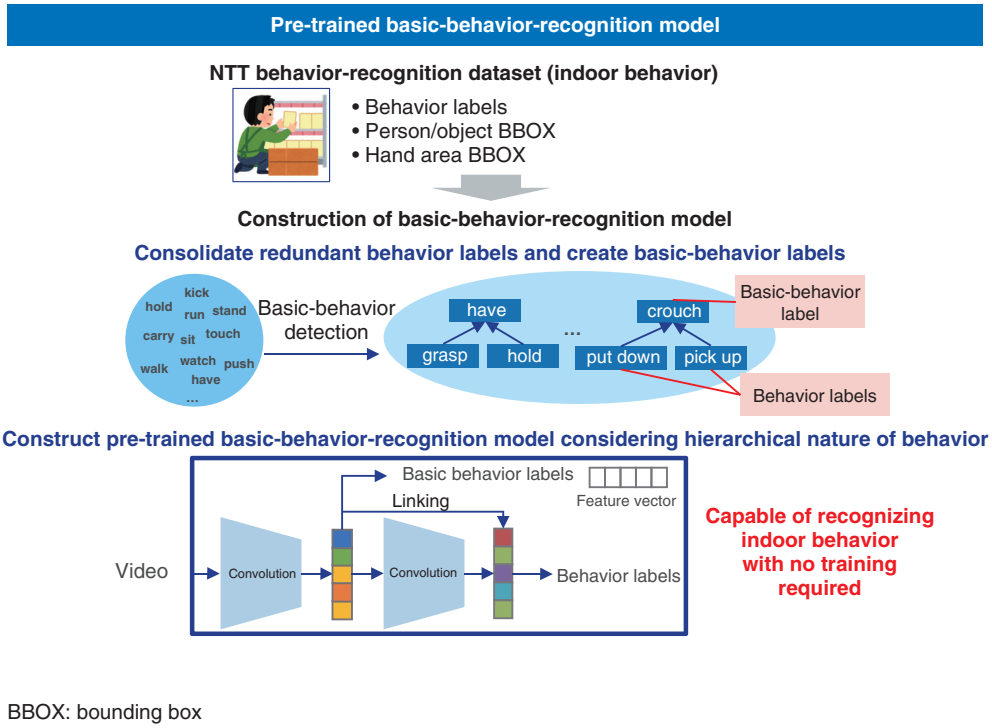
incorporates an NTT-original behavior-recognition model that can robustly recognize human behavior by defining and recognizing a highly general, basic-behavior-recognition model common to a variety of industries by taking the hierarchical nature of behavior into account. Specifically, the engine first executes recognition processing of state changes, such as changes in posture, then recognizes basic behavior using the recognition results of those state changes (**Fig. 2**). The approach to basic-behavior recognition is to define basic-behavior labels with high generality considering the hierarchical nature of behavior and then preset a basic-behavior-recognition model trained to recognize such defined basic behavior in the human-behavior-understanding engine (**Fig. 3**).

A preset basic-behavior training model is constructed using training data targeting real indoor environments (obtained, for example, from convenience store video) having relatively good lighting conditions. Given an indoor environment similar to the training dataset, this preset, pre-trained basic-behavior-recognition model can be used directly in its current form, so the period for constructing the model can be reduced and expenses decreased.

It is also possible to specify important behaviors in video through visual observation against a large amount of accumulated video and to search for and collect behaviors similar to those important behaviors from a large amount of video (**Fig. 4**). This process configures multiple features from video information, distance information, and movement information obtained from camera video and uses a combination of those features to identify the places and times for which important behaviors and similar behaviors can easily occur.

2.2 CNN model lightweight-conversion function

Some of the processing of human-behavior recognition in video is executed using a CNN, which repeats and applies convolutional operations. Such processing, however, is one factor increasing computational complexity. It achieves lightweight processing while preserving accuracy by optimizing the parameters of convolutional filters used in these



BBOX: bounding box

Fig. 3. Basic-behavior-recognition model.

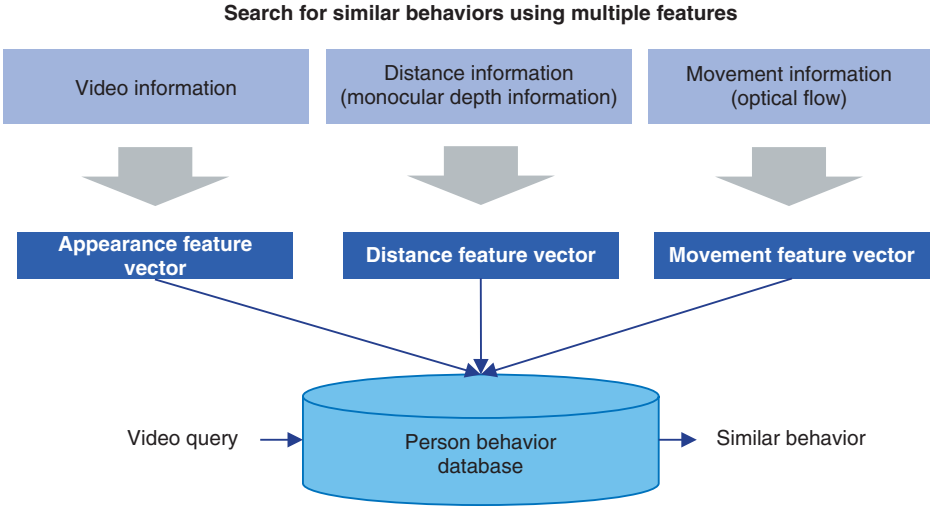


Fig. 4. Behavior-search function.

convolutional operations (**Fig. 5**). This is done by making a convolutional filter sparse using the input/output relationship of the filter as a condition on the basis of the human visual characteristics and decreasing computational complexity by changing the filter-application range while suppressing a drop in accuracy. Optimizing computational complexity in this manner can also enable lightweight processing in the temporal dimension, which in turn makes it possible to recognize the behavior of multiple people in real

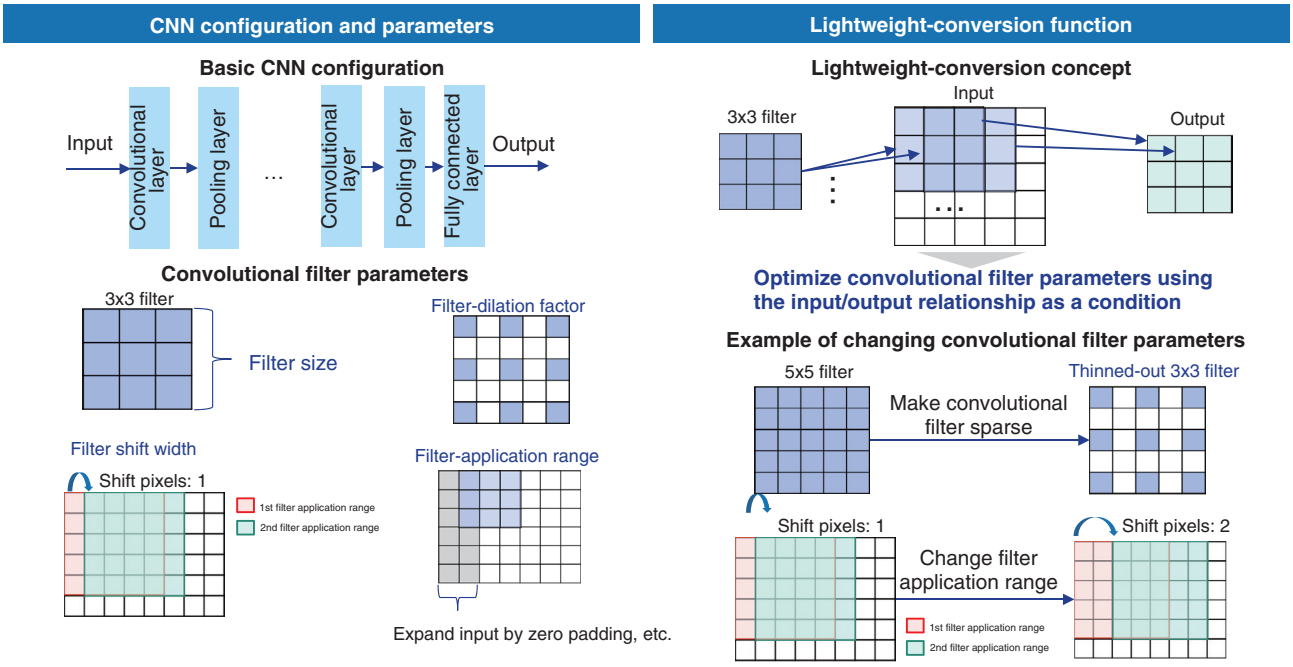


Fig. 5. Achieving lightweight recognition in a CNN.

time even on a machine with no GPUs.

3. Summary and future developments

We have reduced the costs of creating training data and constructing a machine environment by loading a pre-trained basic-behavior-recognition model in an engine and making CNN-recognition processing lightweight. For basic-behavior recognition in an indoor environment, this engine makes it unnecessary to collect video data for training purposes and assign labels at the time of introduction or to train a behavior-recognition model. It also enables the basic behavior of multiple people to be recognized in real time on a CPU-based machine. For future research, we plan to achieve the recognition of composite behavior composed of series of operations such as product replacement, which can be useful in work analysis, by combining the basic behavior or objects recognized by our human-behavior-understanding engine with interaction information. By enabling the recognition of composite behavior described in work manuals with such titles as "stocking the shelves with products in a retail store," we will promote research and development that will make it easy to use human-behavior-understanding technology in work analysis (**Fig. 6**).

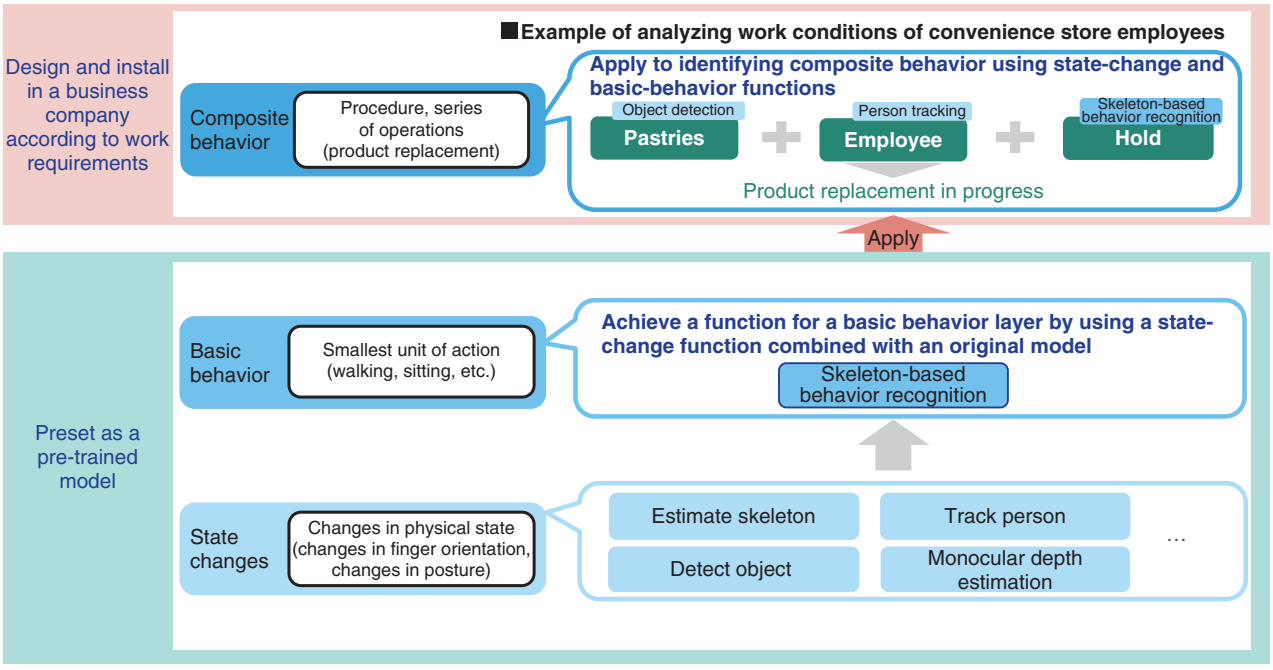
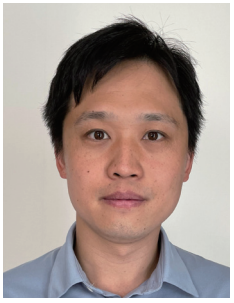


Fig. 6. Composite behavior recognition tailored to usage scenario.

Reference

[1] S. Kondo, D. Sato, M. Goto, M. Takagi, and N. Matsumura, "Telepresence Technology to Facilitate Remote Working with Human Physicality," NTT Technical Review, Vol. 20, No. 11, pp. 27–32, Nov. 2022. https://doi.org/10.53829/ntr202211fa3



Motohiro Takagi
Senior Research Engineer, NTT Human Informatics Laboratories.
He received a B.E., M.E., and Ph.D. from Keio University, Kanagawa, in 2009, 2011, and 2020 and joined NTT in 2011. His research interests include human-behavior understanding through machine learning, computer vision, and natural language processing.