

(19)日本国特許庁(JP)

(12)特許公報(B2)

(11)特許番号
特許第7309101号
(P7309101)

(45)発行日 令和5年7月14日(2023.7.14)

(24)登録日 令和5年7月6日(2023.7.6)

(51)国際特許分類
G 0 6 N 20/00 (2019.01)F I
G 0 6 N 20/00

請求項の数 4 (全10頁)

(21)出願番号 特願2023-527359(P2023-527359)
 (86)(22)出願日 令和3年6月16日(2021.6.16)
 (86)国際出願番号 PCT/JP2021/022916
 (87)国際公開番号 WO2022/264331
 (87)国際公開日 令和4年12月22日(2022.12.22)
 審査請求日 令和5年5月8日(2023.5.8)
 早期審査対象出願

(73)特許権者 000006013
 三菱電機株式会社
 東京都千代田区丸の内二丁目7番3号
 (74)代理人 110002491
 弁理士法人クロスボーダー特許事務所
 小関 義博
 東京都千代田区丸の内二丁目7番3号
 三菱電機株式会社内
 審査官 福西 章人

最終頁に続く

(54)【発明の名称】 攻撃検知装置、敵対的サンプルパッチ検知システム、攻撃検知方法、及び、攻撃検知プログラム

(57)【特許請求の範囲】

【請求項1】

撮影時間範囲内の互いに異なる時刻に撮影範囲内の範囲を撮影した複数の画像データそれぞれを用いて計算された複数の認識スコアであって、前記複数の画像データそれぞれにおいて物体を検知した結果を示す複数の認識スコアを用いて生成された時系列データである時系列認識スコアに、敵対的サンプルパッチ攻撃が前記複数の画像データの少なくともいはずれかに対して実施された場合において生じる異常パターンが含まれているか否かを検知する異常パターン検知部
 を備える攻撃検知装置であって、

前記異常パターンは、前記時系列認識スコアが示す認識スコアの値が物体検知閾値未満異常検知閾値以上である時間が異常検知時間以上継続することを示すパターンである攻撃検知装置。

【請求項2】

請求項1に記載の攻撃検知装置と、
 前記複数の画像データそれぞれを用いて前記複数の認識スコアを計算する物体検知部を備える物体検知装置と
 を備える敵対的サンプルパッチ検知システム。

【請求項3】

コンピュータが、撮影時間範囲内の互いに異なる時刻に撮影範囲内の範囲を撮影した複数の画像データそれぞれを用いて計算された複数の認識スコアであって、前記複数の画像デ

ータそれにおいて物体を検知した結果を示す複数の認識スコアを用いて生成された時系列データである時系列認識スコアに、敵対的サンプルパッチ攻撃が前記複数の画像データの少なくともいずれかに対して実施された場合において生じる異常パターンが含まれているか否かを検知する攻撃検知方法であって、

前記異常パターンは、前記時系列認識スコアが示す認識スコアの値が物体検知閾値未満異常検知閾値以上である時間が異常検知時間以上継続することを示すパターンである攻撃検知方法。

【請求項 4】

撮影時間範囲内の互いに異なる時刻に撮影範囲内の範囲を撮影した複数の画像データそれぞれを用いて計算された複数の認識スコアであって、前記複数の画像データそれぞれにおいて物体を検知した結果を示す複数の認識スコアを用いて生成された時系列データである時系列認識スコアに、敵対的サンプルパッチ攻撃が前記複数の画像データの少なくともいずれかに対して実施された場合において生じる異常パターンが含まれているか否かを検知する異常パターン検知処理

10

をコンピュータである攻撃検知装置に実行させる攻撃検知プログラムであって、

前記異常パターンは、前記時系列認識スコアが示す認識スコアの値が物体検知閾値未満異常検知閾値以上である時間が異常検知時間以上継続することを示すパターンである攻撃検知プログラム。

【発明の詳細な説明】

【技術分野】

20

【0001】

本開示は、攻撃検知装置、敵対的サンプルパッチ検知システム、攻撃検知方法、及び、攻撃検知プログラムに関する。

【背景技術】

【0002】

入力画像における各オブジェクトの位置をバウンディングボックスで示し、各オブジェクトの種類をラベルとして示す物体検知のタスクにおいて、ニューラルネットワークを用いた深層学習の手法が近年、非常に高い精度を達成している。非特許文献1は、電子的な振動が加えられた画像を印刷した敵対的サンプルパッチを物理的に配置し、配置された敵対的サンプルパッチを撮影した画像を入力した際に物体検知による検知を逃れる敵対的サンプルパッチ攻撃の手法を開示している。

30

【先行技術文献】

【非特許文献】

【0003】

【文献】S. Thys et al., “Fooling automated surveillance cameras: adversarial patches to attack person detection”, CVPRW (Conference on Computer Vision and Pattern Recognition Workshops), IEEE (Institute of Electrical and Electronics Engineers) / CVF (Computer Vision Foundation), 2019

40

【発明の概要】

【発明が解決しようとする課題】

【0004】

既存技術によれば、敵対的サンプルパッチ攻撃が行われた場合に、当該攻撃を検知することが困難であるという課題がある。

【0005】

本開示は、敵対的サンプルパッチ攻撃が行われた場合に、当該攻撃を検知することを目的とする。

【課題を解決するための手段】

50

【0006】

本開示に係る攻撃検知装置は、

撮影時間範囲内の互いに異なる時刻に撮影範囲内の範囲を撮影した複数の画像データそれぞれを用いて計算された複数の認識スコアであって、前記複数の画像データそれぞれにおいて物体を検知した結果を示す複数の認識スコアを用いて生成された時系列データである時系列認識スコアに、敵対的サンプルパッチ攻撃が前記複数の画像データの少なくともいはずれかに対して実施された場合において生じる異常パターンが含まれているか否かを検知する異常パターン検知部

を備える。

【発明の効果】

【0007】

本開示に係る攻撃検知装置は、物体を検知した結果を示す認識スコアから成る時系列認識スコアに、敵対的サンプルパッチ攻撃が実施された場合において生じる異常パターンが含まれているか否かを検知する異常パターン検知部を備える。従って、本開示によれば、敵対的サンプルパッチ攻撃が行われた場合に、当該攻撃を検知することができる。

【図面の簡単な説明】

【0008】

【図1】実施の形態1に係る敵対的サンプルパッチ検知システム100のシステム構成例を示す図。

【図2】実施の形態1に係る物体検知装置110の機能構成例を示す図。

【図3】実施の形態1に係る攻撃検知装置120の機能構成例を示す図。

【図4】実施の形態1に係る物体検知装置110及び攻撃検知装置120の各々のハードウェア構成例を示す図。

【図5】実施の形態1に係る敵対的サンプルパッチ検知システム100の動作を示すフローチャート。

【図6】実施の形態1の変形例に係る物体検知装置110及び攻撃検知装置120の各々のハードウェア構成例を示す図。

【発明を実施するための形態】

【0009】

実施の形態の説明及び図面において、同じ要素及び対応する要素には同じ符号を付している。同じ符号が付された要素の説明は、適宜に省略又は簡略化する。図中の矢印はデータの流れ又は処理の流れを主に示している。また、「部」を、「回路」、「工程」、「手順」、「処理」又は「サーキットリー」に適宜読み替えてよい。

【0010】

実施の形態1.

以下、本実施の形態について、図面を参照しながら詳細に説明する。

【0011】

構成の説明

図1は、本実施の形態に係る敵対的サンプルパッチ検知システム100のシステム構成例を示している。敵対的サンプルパッチ検知システム100は、物体検知装置110と攻撃検知装置120とを有する。物体検知装置110と攻撃検知装置120とは一体的に構成されてもよい。

物体検知装置110は、入力画像データ x を入力として受け取り、物体検知結果である認識スコア y を出力する。

攻撃検知装置120は、認識スコア y を入力として受け取り、敵対的サンプルパッチ攻撃の検知結果 r として、敵対的サンプルパッチ攻撃による認識スコアの異常パターンが検知された場合には敵対的サンプルパッチ攻撃を検知したことを示す結果、それ以外の場合には敵対的サンプルパッチ攻撃を検知していないことを示す結果を出力する。敵対的サンプルパッチ攻撃は、敵対的サンプル攻撃の一種であり、特にニューラルネットワーク等を用いた物体検知を逃れる攻撃である。異常パターンは、具体例として、認識スコアの値が

10

20

30

40

50

、一定時間継続して物体検知の閾値を多少下回る値であるパターンである。即ち、異常パターンは、時系列認識スコアが示す認識スコアの値が物体検知閾値未満異常検知閾値以上である時間が異常検知時間以上継続することを示すパターンである。現状の敵対的サンプルパッチを用いた攻撃では、認識スコアの値を、物体検知の閾値を下回るよう抑えることができるが、完全に0にすることはできない。そのため、この異常パターンは当該攻撃の検知において有効である。ただし、異常パターンはこのパターンのみに限定されない。

【0012】

図2は、物体検知装置110の機能構成例を示している。物体検知装置110は、データ入力部111と物体検知器112とデータ出力部113とを有する。

データ入力部111は、物体検知の対象である入力画像データ x を受け取り、受け取った入力画像データ x を物体検知器112に入力する。

物体検知器112は、入力された入力画像データ x を用いて認識スコア y を算出し、算出した認識スコア y をデータ出力部113に出力する。物体検知器112は、具体例としてニューラルネットワークによって構築された物体検知器112である。ニューラルネットワークは、具体例として、YOLO (You Only Look Once)、SSD (Single Shot Multibox Detector)、又はFaster R-CNN (Region-based Convolutional Neural Networks) 等である。物体検知器112は、入力された画像に映る各オブジェクトに対応するバウンディングボックスの位置を表す座標と、各バウンディングボックス内のオブジェクトの種類及び確信度を示す確率を認識スコアとして出力する。物体検知器112は、複数の画像データそれぞれを用いて複数の認識スコアを計算する。物体検知器112は物体検知部とも呼ばれる。

データ出力部113は、物体検知器112が計算した認識スコア y を出力する。

【0013】

図3は、攻撃検知装置120の機能構成例を示している。攻撃検知装置120は、データ入力部121と認識スコア集積部122と異常パターン検知部123とデータ出力部124とを有する。

データ入力部121は、認識スコア y を受け取り、受け取った認識スコア y を認識スコア集積部122に入力する。

認識スコア集積部122は、入力された認識スコア y を時系列認識スコア Y' に追加することによって時系列認識スコア Y を生成し、生成した時系列認識スコア Y を異常パターン検知部123に入力する。時系列認識スコア Y は更新された時系列認識スコア Y' に当たる。時系列認識スコア Y' は、新たな認識スコア y が入力されるまでに入力された認識スコア y を集積して生成した時系列データである。認識スコア集積部122は、認識スコア集積部122に認識スコア y が入力される度に時系列認識スコア Y を異常パターン検知部123に入力しなくてもよく、一定時間毎に異常パターン検知部123が認識スコア集積部122から最新の時系列認識スコア Y を取り出し、取り出した時系列認識スコア Y を用いてもよい。時系列認識スコアは、複数の認識スコアを用いて生成された時系列データである。複数の認識スコアは、撮影時間範囲内の互いに異なる時刻に撮影範囲内の範囲を撮影した複数の画像データそれぞれを用いて計算され、また、当該複数の画像データそれぞれにおいて物体を検知した結果を示す。

異常パターン検知部123は、時系列認識スコア Y に対して事前に指定された異常パターンとのマッチングを行う。異常パターン検知部123は、検知結果 r として、時系列認識スコア Y が異常パターンに合致する場合には検知したことを示す結果、時系列認識スコア Y が異常パターンに合致しない場合には敵対的サンプルパッチ攻撃を検知していないことを示す結果をデータ出力部124に入力する。異常パターン検知部123は、通常時の認識スコアの推移と異なる推移である異常パターンを検知した際に敵対的サンプルパッチによる攻撃が行われたと判定する。異常パターン検知部123は、時系列認識スコアに、敵対的サンプルパッチ攻撃が複数の画像データの少なくともいずれかに対して実施された場合において生じる異常パターンが含まれているか否かを検知する。

10

20

30

40

50

データ出力部 124 は、入力された検知結果 r を出力する。

【0014】

図4は、本実施の形態に係る物体検知装置110及び攻撃検知装置120の各々のハードウェア資源の一例を示す図である。物体検知装置110及び攻撃検知装置120の各々は、コンピュータから成り、また、複数のコンピュータから成ってもよい。

【0015】

物体検知装置110及び攻撃検知装置120の各々は、プロセッサ11を備えている。プロセッサ11は、バス12を介してROM13と、RAM14と、通信ボード15と、表示装置であるディスプレイ51と、キーボード52と、マウス53と、ドライブ54と、磁気ディスク装置20等のハードウェアデバイスと接続され、これらのハードウェアデバイスを制御する。プロセッサ11は、演算処理を行うIC (Integrated Circuit) であり、具体例として、CPU (Central Processing Unit)、DSP (Digital Signal Processor)、又はGPU (Graphics Processing Unit) である。物体検知装置110及び攻撃検知装置120の各々は、複数のプロセッサを備えてもよい。複数のプロセッサは、プロセッサ11の役割を分担する。

10

【0016】

ドライブ54は、FD (Flexible Disk Drive)、CD (Compact Disc)、又はDVD (Digital Versatile Disc) 等の記憶媒体を読み書きする装置である。

20

【0017】

ROM13と、RAM14と、磁気ディスク装置20と、ドライブ54との各々は、記憶装置の一例である。記憶装置はコンピュータから独立していてもよい。

【0018】

キーボード52と、マウス53と、通信ボード15との各々は入力装置の一例である。ディスプレイ51及び通信ボード15は出力装置の一例である。

【0019】

通信ボード15は、有線又は無線で、LAN (Local Area Network)、インターネット、又は電話回線等の通信網に接続している。通信ボード15は、具体例として、通信チップ又はNIC (Network Interface Card) から成る。

30

【0020】

磁気ディスク装置20は、OS (オペレーティングシステム) 21と、プログラム群22と、ファイル群23とを記憶している。

【0021】

プログラム群22は、本実施の形態において各部又は各器として説明する機能を実行するプログラムを含む。プログラムは、プロセッサ11により読み出され実行される。即ち、プログラムは、各部又は各器としてコンピュータを機能させるものであり、また、各部又は各器の手順又は方法をコンピュータに実行させるものである。

本明細書に記載されているいずれのプログラムも、コンピュータが読み取り可能な不揮発性の記録媒体に記録されていてもよい。不揮発性の記録媒体は、具体例として、光ディスク又はフラッシュメモリである。本明細書に記載されているいずれのプログラムも、プログラムプロダクトとして提供されてもよい。

40

【0022】

ファイル群23は、本実施の形態において説明する各部又は各器で使用される各種データを含む。

【0023】

*** 動作の説明 ***

物体検知装置110の動作手順は、物体検知方法に相当する。また、物体検知装置110の動作を実現するプログラムは、物体検知プログラムに相当する。攻撃検知装置120

50

の動作手順は、攻撃検知方法に相当する。また、攻撃検知装置 120 の動作を実現するプログラムは、攻撃検知プログラムに相当する。

【0024】

図 5 は、本実施の形態に係る敵対的サンプルパッチ検知システム 100 の処理の一例を示すフローチャートである。本図を参照して敵対的サンプルパッチ検知システム 100 の処理を説明する。

【0025】

(ステップ S11)

データ入力部 111 は、入力画像データ x を受け取り、受け取った入力画像データ x を物体検知器 112 に入力する。

【0026】

(ステップ S12)

物体検知器 112 は、入力された入力画像データ x を用いて認識スコア y を算出する。

【0027】

(ステップ S13)

データ出力部 113 は、算出された認識スコア y を出力する。

【0028】

(ステップ S14)

データ入力部 121 は、認識スコア y を受け取り、受け取った認識スコア y を認識スコア集積部 122 に入力する。

【0029】

(ステップ S15)

認識スコア集積部 122 は、入力された認識スコア y を時系列認識スコア Y' に追加することによって時系列認識スコア Y' を時系列認識スコア Y に更新し、更新した時系列認識スコア Y を異常パターン検知部 123 に入力する。

【0030】

(ステップ S16)

異常パターン検知部 123 は、入力された時系列認識スコア Y が事前に指定された異常パターンに合致するか否かを判定する。

異常パターンが時系列認識スコア Y に合致する場合、攻撃検知装置 120 はステップ S

17 に進む。それ以外の場合、攻撃検知装置 120 はステップ S18 に進む。

【0031】

(ステップ S17)

データ出力部 124 は、検知結果 r として、敵対的サンプルパッチ攻撃を検知したことを示す結果を出力する。

【0032】

(ステップ S18)

データ出力部 124 は、検知結果 r として、敵対的サンプルパッチ攻撃を検知していないことを示す結果を出力する。

【0033】

*** 実施の形態 1 の効果の説明 ***

以上のように、本実施の形態によれば、時系列認識スコアが異常パターンに合致するか否かを判定することによって敵対的サンプルパッチ攻撃を検知することができる。

【0034】

*** 他の構成 ***

<変形例 1 >

図 6 は、本変形例に係る物体検知装置 110 及び攻撃検知装置 120 の各々のハードウェア構成例を示している。

物体検知装置 110 及び攻撃検知装置 120 の各々は、プロセッサ 11、プロセッサ 11 と ROM 13、プロセッサ 11 と RAM 14、あるいはプロセッサ 11 と ROM 13 と

10

20

30

40

50

R A M 1 4 とし代えて、処理回路 1 8 を備える。

処理回路 1 8 は、物体検知装置 1 1 0 及び攻撃検知装置 1 2 0 の各々が備える各部の少なくとも一部を実現するハードウェアである。

処理回路 1 8 は、専用のハードウェアであってもよく、また、磁気ディスク装置 2 0 に格納されるプログラムを実行するプロセッサであってもよい。

【 0 0 3 5 】

処理回路 1 8 が専用のハードウェアである場合、処理回路 1 8 は、具体例として、単一回路、複合回路、プログラム化したプロセッサ、並列プログラム化したプロセッサ、A S I C (A p p l i c a t i o n S p e c i f i c I n t e g r a t e d C i r c u i t) 、 F P G A (F i e l d P r o g r a m m a b l e G a t e A r r a y) 又はこれらの組み合わせである。

物体検知装置 1 1 0 及び攻撃検知装置 1 2 0 の各々は、処理回路 1 8 を代替する複数の処理回路を備えてよい。複数の処理回路は、処理回路 1 8 の役割を分担する。

【 0 0 3 6 】

物体検知装置 1 1 0 及び攻撃検知装置 1 2 0 の各々において、一部の機能が専用のハードウェアによって実現されて、残りの機能がソフトウェア又はファームウェアによって実現されてもよい。

【 0 0 3 7 】

処理回路 1 8 は、具体例として、ハードウェア、ソフトウェア、ファームウェア、又はこれらの組み合わせにより実現される。

プロセッサ 1 1 と R O M 1 3 と R A M 1 4 と処理回路 1 8 とを、総称して「プロセッシングサーキットリー」という。つまり、物体検知装置 1 1 0 及び攻撃検知装置 1 2 0 の各々の各機能構成要素の機能は、プロセッシングサーキットリーにより実現される。

【 0 0 3 8 】

* * * 他の実施の形態 * * *

実施の形態 1 について説明したが、本実施の形態のうち、複数の部分を組み合わせて実施しても構わない。あるいは、本実施の形態を部分的に実施しても構わない。その他、本実施の形態は、必要に応じて種々の変更がなされても構わず、全体としてあるいは部分的に、どのように組み合わせて実施されても構わない。各部又は各器として説明するものは、ファームウェア、ソフトウェア、ハードウェア又はこれらの組み合わせのいずれで実装されても構わない。

なお、前述した実施の形態は、本質的に好ましい例示であって、本開示と、その適用物と、用途の範囲とを制限することを意図するものではない。フローチャート等を用いて説明した手順は、適宜変更されてもよい。

【 符号の説明 】

【 0 0 3 9 】

1 1 プロセッサ、1 2 バス、1 3 R O M 、1 4 R A M 、1 5 通信ボード、1 8 処理回路、2 0 磁気ディスク装置、2 1 O S 、2 2 プログラム群、2 3 ファイル群、5 1 ディスプレイ、5 2 キーボード、5 3 マウス、5 4 ドライブ、1 0 0 敵対的サンプルパッチ検知システム、1 1 0 物体検知装置、1 1 1 データ入力部、1 1 2 物体検知器、1 1 3 データ出力部、1 2 0 攻撃検知装置、1 2 1 データ入力部、1 2 2 認識スコア集積部、1 2 3 異常パターン検知部、1 2 4 データ出力部。

10

20

30

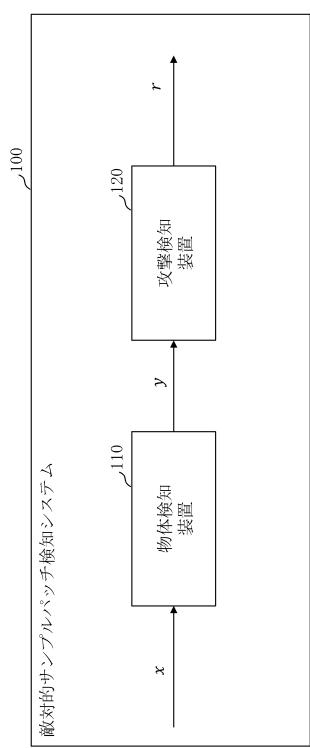
40

50

【図面】

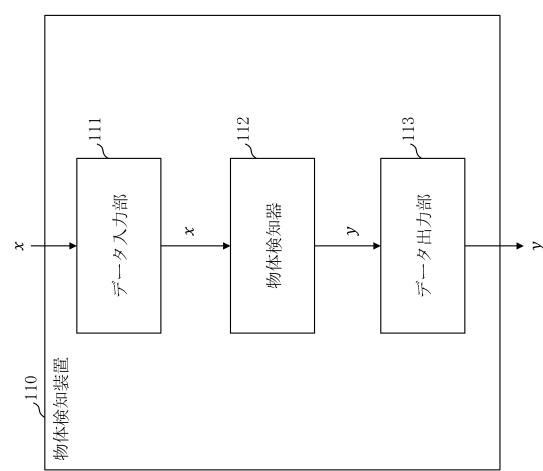
【図1】

図1



【図2】

図2

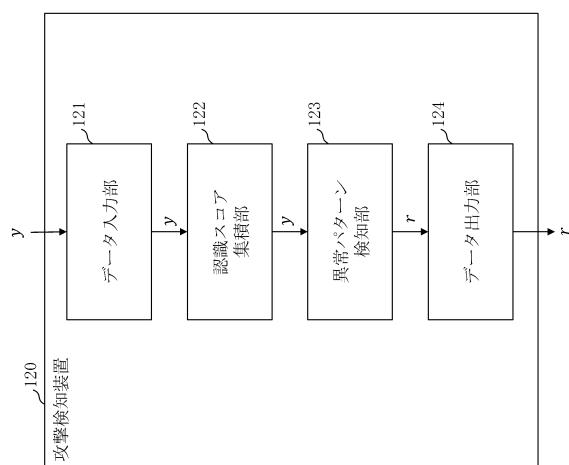


10

20

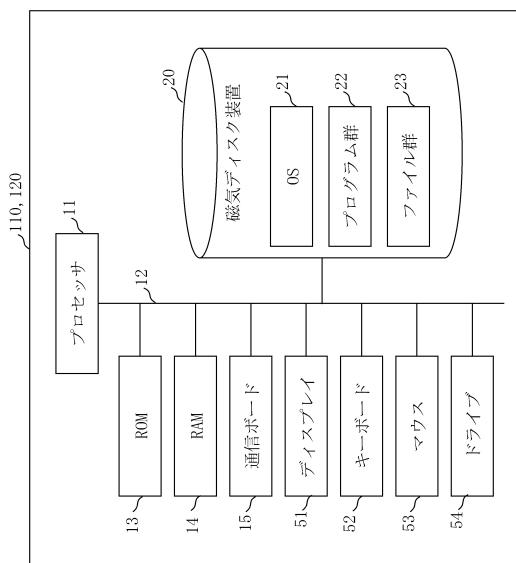
【図3】

図3



【図4】

図4



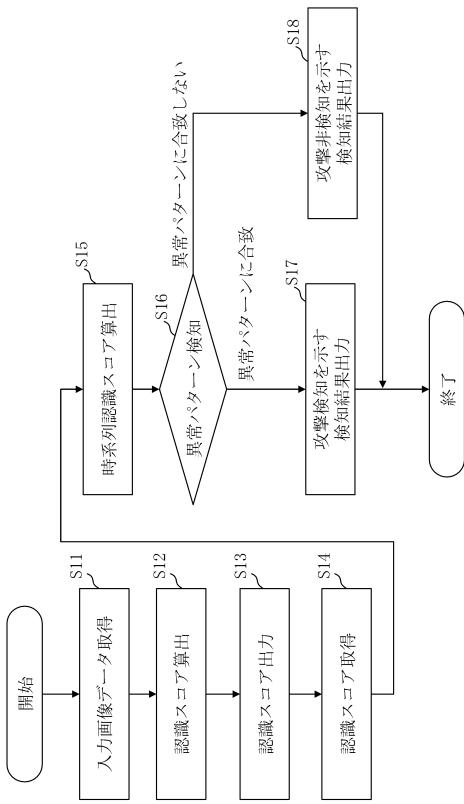
30

40

50

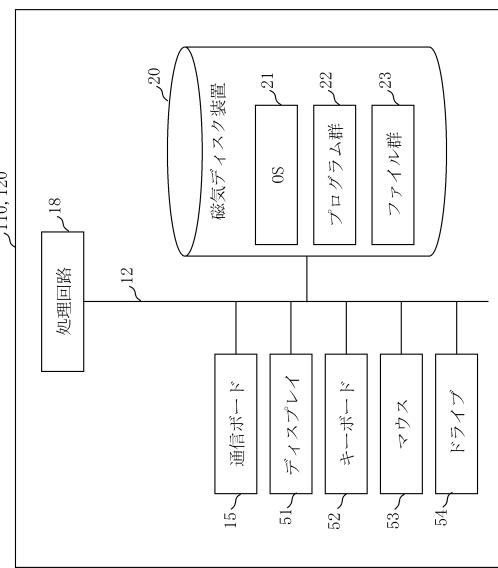
【図 5】

図5



【図 6】

図6



10

20

30

40

50

フロントページの続き

(56)参考文献

特開2019-125867(JP,A)

JI, N., et al. , Adversarial YOLO: Defense Human Detection Patch Attacks via Detecting Adversarial Patches , arXiv.org [online] , arXiv:2103.08860 , 2021年03月16日 , Retrieved from the Internet: URL: <https://arxiv.org/abs/2103.08860> [Retrieved on 2021-08-31]

XIAO, C., et al. , AdvIT: Adversarial Frames Identifier Based on Temporal Consistency in Videos , Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV) , 2019年10月27日 , pp. 3967-3976 , Retrieved from the Internet: URL: <https://ieeexplore.ieee.org/document/9010733> [Retrieved on 2021-08-31] , DOI: 10.1109/ICCV.2019.00407

DOSHI, K. et al. , Continual Learning for Anomaly Detection in Surveillance Videos , arXiv.org [online] , arXiv:2004.07941 , 2020年04月15日 , Retrieved from the Internet: URL: <https://arxiv.org/abs/2004.07941> [Retrieved on 2021-08-31]

(58)調査した分野 (Int.Cl. , DB名)

G 06 F 21/00 - 21/88

G 06 N 3/00 - 99/00