

(12) 特許協力条約に基づいて公開された国際出願

(19) 世界知的所有権機関
国際事務局

(43) 国際公開日
2017 年 9 月 8 日(08.09.2017)



(10) 国際公開番号
WO 2017/149971 A1

- (51) 国際特許分類:
G06N 3/04 (2006.01) **G06T 7/00** (2017.01)
- (21) 国際出願番号: PCT/JP2017/001554
- (22) 国際出願日: 2017 年 1 月 18 日(18.01.2017)
- (25) 国際出願の言語: 日本語
- (26) 国際公開の言語: 日本語
- (30) 優先権データ:
特願 2016-038956 2016 年 3 月 1 日(01.03.2016) JP
- (71) 出願人: 株式会社デンソー(DENSO CORPORATION) [JP/JP]; 〒4488661 愛知県刈谷市昭和町 1 丁目 1 番地 Aichi (JP).
- (72) 発明者: 蓑谷 顕一(MINOYA, Kenichi); 〒4488661 愛知県刈谷市昭和町 1 丁目 1 番地 株式会社デンソー内 Aichi (JP). 尾崎 智章(OZAKI, Tomoaki); 〒4488661 愛知県刈谷市昭和町 1 丁目 1 番地 株式会社デンソー内 Aichi (JP).
- (74) 代理人: 特許業務法人 サトー国際特許事務所 (SATO INTERNATIONAL PATENT FIRM); 〒4600008 愛知県名古屋市中区栄四丁目 6 番 15 号 フォーティーンヒルズセンタービル Aichi (JP).

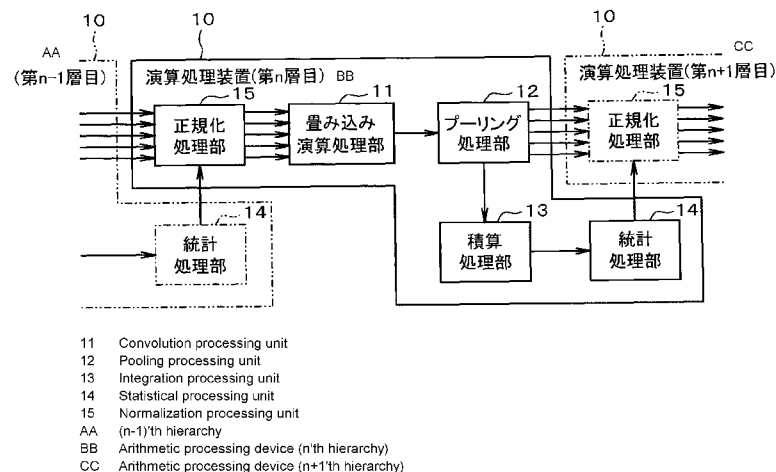
- (81) 指定国 (表示のない限り、全ての種類の国内保護が可能): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) 指定国 (表示のない限り、全ての種類の広域保護が可能): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), ユーロパ (AM, AZ, BY, KG, KZ, RU, TJ, TM), ユーロッパ (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

添付公開書類:

— 国際調査報告 (条約第 21 条(3))

(54) Title: ARITHMETIC PROCESSING DEVICE

(54) 発明の名称: 演算処理装置



(57) Abstract: This arithmetic processing device 10 is provided with: a convolution arithmetic processing unit 11 for executing a convolution arithmetic process on input data that is inputted from a preceding hierarchy; a pooling processing unit 12 for executing a pooling process on the result data of processing by the convolution arithmetic processing unit; an integration processing unit 13 for integrating the result data of processing by the pooling processing unit before an arithmetic operation of the next hierarchy begins; a statistical processing unit 14 for calculating the average value and standard deviation of the data integrated by the integration processing unit; and a normalization processing unit 15 which, when the arithmetic operation of the next hierarchy begins, executes a normalization process on the result data of processing that is outputted from the pooling processing unit in the preceding hierarchy using the average value and standard deviation calculated by the statistical processing unit in the preceding hierarchy.

(57) 要約:

[続葉有]

WO 2017/149971 A1



演算処理装置 10 は、前階層から入力される入力データに対して畳み込み演算処理を実行する畳み込み演算処理部 11 と、畳み込み演算処理部による処理結果データに対してプーリング処理を実行するプーリング処理部 12 と、次回層の演算が開始される前に、プーリング処理部による処理結果データを積算する積算処理部 13 と、積算処理部が積算するデータの平均値および標準偏差を算出する統計処理部 14 と、次回層の演算が開始されると、前階層において統計処理部により算出された平均値および標準偏差を用いて、前階層においてプーリング処理部により出力された処理結果データに対して正規化処理を実行する正規化処理部 15 と、を備える。

明 細 書

発明の名称：演算処理装置

関連出願の相互参照

[0001] 本出願は、2016年3月1日に出願された日本出願番号2016-038956号に基づくもので、ここにその記載内容を援用する。

技術分野

[0002] 本開示は、演算処理装置に関する。

背景技術

[0003] 従来より、複数の処理層が階層的に接続されたニューラルネットワークによる演算を実行する演算処理装置が考えられている。特に画像認識を行う演算処理装置においては、いわゆる畳み込みニューラルネットワーク（CNN：Convolutional Neural Network）が中核的な存在となっている。

先行技術文献

特許文献

[0004] 特許文献1：特許第5184824号公報

発明の概要

[0005] 従来の畳み込みニューラルネットワークでは、前段の階層により得られる複数の異なる演算結果データ、つまり特徴量の抽出結果データに対して畳み込み演算処理を実行し、活性化処理を実行し、プーリング処理を実行することで、より高次元の特徴量の抽出を行っている。そして、さらに、プーリング処理による処理結果データに対して正規化処理を施すことにより、特徴量の認識率を向上することができ、特徴量の抽出処理を一層優位に行うことができる。

[0006] ところで、処理結果データに対して正規化処理を行う場合、従来では、まず、畳み込み演算処理時にメモリからデータを読み出し、読み出したデータに対して畳み込み演算処理およびプーリング処理を行い、その演算結果データをメモリに格納する。そして、その後、同一の処理階層において、そのメ

メモリに格納したデータを読み出し、読み出したデータに対して正規化処理を行う。即ち、従来技術では、同一の処理階層においてメモリからのデータの読み出しを2回行う必要があり、大きなレイテンシが発生してしまう。

[0007] そこで、本開示は、ニューラルネットワークによる演算処理を実現する演算処理装置において、正規化処理を行うための構成に改良を施すことにより、レイテンシの発生を抑えることを目的とする。

[0008] 本開示の一態様において、演算処理装置は、複数の処理層が階層的に接続されたニューラルネットワークによる演算を実行する演算処理装置であって、畳み込み演算処理部、プーリング処理部、積算処理部、統計処理部、正規化処理部を備える。畳み込み演算処理部は、前階層から入力される入力データに対して畳み込み演算処理を実行する。プーリング処理部は、前記畳み込み演算部による処理結果データに対してプーリング処理を実行する。積算処理部は、次回層の演算が開始される前に、前記プーリング処理部による処理結果データを積算する。統計処理部は、前記積算処理部が積算するデータの平均値および標準偏差を算出する。正規化処理部は、次回層の演算が開始されると、前階層において前記統計処理部により算出された平均値および標準偏差を用いて、前階層において前記プーリング処理部により出力された処理結果データに対して正規化処理を実行する。

[0009] この構成によれば、次回層における演算処理の入力部分において、前階層の演算処理時に得られた統計値を用いて正規化処理を行うことができる。そのため、同一の処理階層においてメモリからのデータの読み出しを2回行う必要が無く、レイテンシの発生を抑えることができる。

図面の簡単な説明

[0010] 本開示についての上記目的およびその他の目的、特徴や利点は、添付の図面を参照しながら下記の詳細な記述により、より明確になる。その図面は、
[図1]図1は、畳み込みニューラルネットワークの構成例を概念的に示す図であり、
[図2]図2は、中間層における演算処理の流れを視覚的に示す図「その1」で

あり、

[図3]図3は、中間層における演算処理の流れを視覚的に示す図「その2」であり、

[図4]図4は、特徴量抽出処理に用いられる一般的な演算式および関数を示す図であり、

[図5]図5は、第1実施形態に係る演算処理装置の構成例を概略的に示すブロック図であり、

[図6]図6は、正規化関数の一例を示す図であり、

[図7]図7は、演算処理装置による演算処理の流れを視覚的に示す図であり、

[図8]図8は、第2実施形態に係る演算処理装置の構成例を概略的に示すブロック図であり、

[図9]図9は、正規化処理を構成する減算式および除算式の一部を示す図であり、

[図10]図10は、第3実施形態に係る演算処理装置の構成例を概略的に示すブロック図であり、

[図11]図11は、プーリング処理部の処理結果データのサンプリング例を示す図であり、

[図12]図12は、第4実施形態に係る演算処理装置の構成例を概略的に示すブロック図であり、

[図13]図13は、プーリング処理部の処理結果データの一部を2進数により示す図であり、

[図14]図14は、第5実施形態に係る演算処理装置の構成例を概略的に示すブロック図「その1」であり、

[図15]図15は、第5実施形態に係る演算処理装置の構成例を概略的に示すブロック図「その2」であり、

[図16]図16は、非線形の活性化関数の一例を示す図であり、

[図17]図17は、第6実施形態に係る演算処理装置の構成例を概略的に示すブロック図である。

発明を実施するための形態

[0011] 以下、演算処理装置の複数の実施形態について図面を参照しながら説明する。なお、各実施形態において実質的に同一の要素には同一の符号を付し、説明を省略する。

(ニューラルネットワーク)

図1には、詳しくは後述する演算処理装置10、20、30、40、50、60に適用されるニューラルネットワーク、この場合、畳み込みニューラルネットワークの構成を概念的に示している。即ち、畳み込みニューラルネットワークNは、入力データである画像データD1から所定の形状やパターンを認識する画像認識技術に応用されるものであり、中間層Naと全結合層Nbとを有する。中間層Naは、複数の特徴量抽出処理層Na1、Na2・・・が階層的に接続された構成である。各特徴量抽出処理層Na1、Na2・・・は、それぞれ畳み込み層Cおよびプーリング層Pを備える。

[0012] 次に、中間層Naにおける処理の流れについて説明する。図2に例示するように、第1層目の特徴量抽出処理層Na1では、演算処理装置は、入力される画像データD1を例えばラスタスキャンにより所定サイズごとに走査する。そして、走査したデータに対して周知の特徴量抽出処理を施すことにより入力画像に含まれる複数の特徴量を抽出する。なお、第1層目の特徴量抽出処理層Na1では、例えば水平方向に延びる線状の特徴量や斜め方向に延びる線状の特徴量などといった比較的シンプルな単独の特徴量を抽出する。このとき、演算処理装置は、入力画像に含まれる複数の特徴にそれぞれ対応する複数の特徴マップを生成する。

[0013] 第2層目の特徴量抽出処理層Na2では、演算処理装置は、前階層の特徴量抽出処理層Na1から入力される入力データを例えばラスタスキャンにより所定サイズごとに走査する。そして、走査したデータに対して周知の特徴量抽出処理を施すことにより入力画像に含まれる複数の特徴量を抽出する。なお、第2層目の特徴量抽出処理層Na2では、第1層目の特徴量抽出処理層Na1で抽出された複数の特徴量の空間的な位置関係などを考慮しながら

統合させることで、より高次元の複合的な特徴量を抽出する。このとき、演算処理装置は、入力画像に含まれる複数の特徴にそれぞれ対応する複数の特徴マップを生成する。

[0014] 第3層目の特徴量抽出処理層N a 3では、演算処理装置は、前階層の特徴量抽出処理層N a 2から入力される入力データを例えばラスタスキャンにより所定サイズごとに走査する。そして、走査したデータに対して周知の特徴量抽出処理を施すことにより入力画像に含まれる複数の特徴量を抽出する。なお、第3層目の特徴量抽出処理層N a 3では、第2層目の特徴量抽出処理層N a 2で抽出された複数の特徴量の空間的な位置関係などを考慮しながら統合させることで、より高次元の複合的な特徴量を抽出する。このとき、演算処理装置は、入力画像に含まれる複数の特徴にそれぞれ対応する複数の特徴マップを生成する。このように、複数の特徴量抽出処理層による特徴量の抽出処理を繰り返すことで、演算処理装置は、画像データD 1に含まれる検出対象物体の画像認識を行う。

[0015] 演算処理装置は、中間層N aにおいて複数の特徴量抽出処理層N a 1, N a 2, N a 3・・・による処理を繰り返すことで入力画像データD 1に含まれる種々の特徴量を高次元で抽出していく。そして、演算処理装置は、中間層N aの処理により得られた結果を中間演算結果データとして全結合層N bに出力する。

[0016] 全結合層N bは、中間層N aから得られる複数の中間演算結果データを結合して最終的な演算結果データを出力する。即ち、全結合層N bは、中間層N aから得られる複数の中間演算結果データを結合し、さらに、その結合結果に対して重み係数を異ならせながら積和演算を行うことにより、最終的な演算結果データ、即ち、入力データである画像データD 1に含まれる検出対象物を認識した画像データを出力する。このとき、積和演算による演算結果の値が大きい部分が検出対象物の一部または全部として認識される。

[0017] 次に、演算処理装置による特徴量抽出処理の流れについて説明する。図3に例示するように、演算処理装置は、前階層の特徴量抽出処理層から入力さ

れる入力データ D_n を所定サイズ、この場合、図にてハッチングで示す 3×3 画素ごとのフィルタサイズにより走査する。なお、画素サイズは、 3×3 画素に限られず、例えば 5×5 画素など適宜変更することができる。

[0018] そして、演算処理装置は、走査したデータに対して、それぞれ周知の畳み込み演算を行う。そして、演算処理装置は、畳み込み演算後のデータに対して周知の活性化処理を行い、畳み込み層 C の出力とする。そして、演算処理装置は、畳み込み層 C の出力データ C_n に対して、所定サイズ、この場合、 2×2 画素ごとに周知のプーリング処理を行い、プーリング層 P の出力とする。そして、演算処理装置は、プーリング層 P の出力データ P_n を次の階層の特徴量抽出処理層に出力する。なお、画素サイズは、 2×2 画素に限られず適宜変更することができる。

[0019] ここで、演算処理装置は、プーリング処理後のデータ P_{n1} , P_{n2} , $\dots$ に対して周知の正規化処理を施すことにより、プーリングデータ P_n を所定の基準形式である正規化データ N_{n1} , P_{n2} , $\dots$ に変換してから次の階層に出力する。これにより、より統一された形式でプーリングデータ P_n を次の階層に出力することができる。従って、特徴量の認識率を向上することができ、特徴量の抽出処理を一層優位に行うことができる。後述する各実施形態においては、演算処理装置には、この正規化処理を行うための構成に改良が施されている。

[0020] 図4には、畳み込み演算処理に用いられる畳み込み関数、活性化処理に用いられる関数、プーリング処理に用いられる関数の一般的な例を示している。即ち、畳み込み関数 Y_{ij} は、直前の層の出力 X_{ij} に学習により得られる重み係数 $W_{p,q}$ を乗算した値を累積する関数となっている。なお、「 N 」は1サイクルの畳み込み演算処理により処理される画素サイズを示す。即ち、例えば1演算サイクルの画素サイズが「 3×3 」画素である場合、 N の値は「2」である。また、畳み込み関数 Y_{ij} は、累積値に所定のバイアス値を加算する関数としてもよい。また、畳み込み関数は、全結合処理にも対応し得る積和演算が可能な関数であれば、種々の関数を採用することができる。

る。また、活性化処理には、周知のロジスティックジグモイド関数やReLU関数 (Rectified Linear Units) などが用いられる。また、プーリング処理には、入力されるデータの最大値を出力する周知の最大プーリング関数や、入力されるデータの平均値を出力する周知の平均プーリング関数などが用いられる。

[0021] 上述した畳み込みニューラルネットワークNによれば、コンボルーション層Cによる処理およびプーリング層Pによる処理が繰り返されることにより、より高次元の特徴量の抽出が可能となる。次に、この畳み込みニューラルネットワークNを適用した演算処理装置に係る複数の実施形態について説明する。なお、各実施形態に係る図においては、第n層目の処理を行っている演算処理装置を実線により示し、次回層である第n+1層目の処理を行っている演算処理装置を二点鎖線により示している。また、第n層目の処理を行っている演算ブロックを実線により示し、次回層である第n+1層目の処理を行っている演算ブロックを二点鎖線により示している。

[0022] (第1実施形態)

図5に例示する演算処理装置10は、畳み込み演算処理部11、プーリング処理部12、積算処理部13、統計処理部14、正規化処理部15を備える。畳み込み演算処理部11は、前階層から入力される入力データに対して周知の畳み込み演算処理を実行して、その処理結果データをプーリング処理部12に出力する。なお、演算処理装置10は、畳み込み演算処理部11の処理結果データに対して図示しない活性化処理部により周知の活性化処理を実行してからプーリング処理部12に出力する。プーリング処理部12は、畳み込み演算処理部11から入力される処理結果データに対して周知のプーリング処理を実行して、その処理結果データを、次回層の演算処理時における正規化処理部15に出力するようになっている。

[0023] 積算処理部13は、プーリング処理部12が出力する処理結果データを、次回層における演算処理が開始される前に積算する。統計処理部14は、積算処理部13が積算するデータ、つまり、プーリング処理部12が出力する

処理結果データの平均値および標準偏差を算出する。正規化処理部 15 は、図示しない減算器や除算器などを備えており、次回層における演算処理が開始されると、前階層において統計処理部 14 により算出された平均値および標準偏差を用いて、前階層においてプーリング処理部 12 により出力された処理結果データに対して周知の正規化処理を実行する。正規化処理部 15 は、例えば図 6 に示す正規化関数に基づいて正規化処理を実行する。

[0024] 演算処理装置 10 によれば、図 7 に例示するように、次回層における演算処理の入力部分において、前階層の演算処理時に得られた統計値を用いて正規化処理を行うようにした。そのため、同一の処理階層においてメモリからのデータの読み出しを 2 回行う必要が無く、連続する 2 層にわたって正規化処理をパイプライン化して行うことができる。よって、処理の高速化を図ることができ、レイテンシの発生を抑えることができる。

[0025] (第 2 実施形態)

図 8 に例示する演算処理装置 20 は、畳み込み演算処理部 21、プーリング処理部 22、積算処理部 23、統計処理部 24、正規化処理部 25 を備える。正規化処理部 25 は、減算処理部 25a および除算処理部 25b を備える。

[0026] 減算処理部 25a は、図 9 に例示する減算式 (1) に基づいて、前階層において統計処理部 24 により算出された平均値を、前階層においてプーリング処理部 22 により出力された処理結果データから減算する。畳み込み演算処理部 21 は、減算処理部 25a が出力する処理結果データに対し周知の畳み込み演算処理を実行する。除算処理部 25b は、図 9 に例示する除算式 (2) に基づいて、畳み込み演算処理部 21 が出力する処理結果データを、前階層において統計処理部 24 により算出された標準偏差により除算する。

[0027] 減算および除算からなる正規化処理を行った上で畳み込み演算処理を行う構成では、正規化処理部に入力されるデータの数に応じた除算回路が必要であり、回路規模が大きくなるという課題がある。演算処理装置 20 によれば、減算からなる正規化処理を行い、その処理結果データに対して畳み込み演

算処理を行い、その処理結果データに対して除算からなる正規化処理を行うようにした。畳み込み演算処理において、入力されるデータは複数であるのに対し、出力されるデータは1つである。そのため、演算処理装置20の構成によれば、畳み込み演算処理の後に除算を行うようにしたので、除算回路の数を削減することができる。

[0028] (第3実施形態)

図10に例示する演算処理装置30は、畳み込み演算処理部31、プーリング処理部32、積算処理部33、統計処理部34、正規化処理部35、サンプリング処理部36を備える。図11に例示するように、サンプリング処理部36は、プーリング処理部32が出力する処理結果データの一部をサンプリングする。そして、サンプリング処理部36は、サンプリングしたデータを積算処理部33に出力する。そして、積算処理部33は、サンプリング処理部36によってサンプリングされたデータ、つまり、プーリング処理部32による処理結果データの一部のみを積算する。そして、統計処理部34は、積算処理部33により積算されたプーリング処理部32による処理結果データの一部のみに基づいて、その平均値および標準偏差を算出する。

[0029] 演算処理装置30によれば、プーリング処理部32による処理結果データの全てではなく一部のみに基づいて、その平均値と標準偏差を算出して正規化処理に用いる。このように、プーリング処理部32による処理結果データの一部のみに基づく統計値によっても十分精度が高い正規化処理を行うことができる。また、演算処理装置30によれば、積算処理部33の処理負荷を低減することができる。

[0030] (第4実施形態)

図12に例示する演算処理装置40は、畳み込み演算処理部41、プーリング処理部42、積算処理部43、統計処理部44、正規化処理部45、右シフト処理部46、左シフト処理部47を備える。右シフト処理部46は、プーリング処理部42と積算処理部43との間に設けられており、プーリング処理部42による処理結果データを所定ビットだけ右にシフトさせて積算

処理部 4 3 に出力する。左シフト処理部 4 7 は、積算処理部 4 3 と統計処理部 4 4 との間に設けられており、積算処理部 4 3 による処理結果データを所定ビットだけ左にシフトさせて統計処理部 4 4 に出力する。

[0031] なお、右シフト処理部 4 6 がデータを右シフトするときのビット数と、左シフト処理部 4 7 がデータを左シフトするときのビット数は同じである。そのため、図 1 3 に例示するように、積算処理部 4 3 は、プーリング処理部 4 2 による処理結果データの上位の所定ビットのみを積算する構成となる。演算処理装置 4 0 によれば、積算処理部 4 3 により、プーリング処理部 4 2 による処理結果データの全ビットではなく上位の所定ビットのみを積算する。この構成によれば、積算処理部 4 3 の処理負荷を低減することができる。

[0032] (第 5 実施形態)

図 1 4 に例示する演算処理装置 5 0 は、畳み込み演算処理部 5 1、プーリング処理部 5 2、積算処理部 5 3、統計処理部 5 4、正規化処理部 5 5 を有する演算ブロック 5 0 0 を複数備える構成である。なお、図 1 4 には、1 つの演算ブロック 5 0 0 のみを示している。演算ブロック 5 0 0 には、それぞれ重み調整処理部 5 6 が備えられている。重み調整処理部 5 6 は、各演算ブロック 5 0 0 の畳み込み演算処理部 5 1 がそれぞれ畳み込み演算処理を実行する際に用いる重み係数のうち最大の値を示す重み係数を特定する。そして、重み調整処理部 5 6 は、その特定した重み係数の最大値の絶対値により、畳み込み演算処理部 5 1 が用いる重み係数を除算する。これにより、重み調整処理部 5 6 は、重み係数を $-1 \sim 1$ の範囲に収まるように調整する。

[0033] 演算処理装置 5 0 は、複数の畳み込み演算処理部 5 1 がそれぞれ畳み込み演算処理を実行する際に用いる重み係数の最大値を特定し、その特定した最大値の絶対値により、複数の畳み込み演算処理部 5 1 が用いる重み係数を除算して、重み係数を $-1 \sim 1$ の範囲に調整する。このように重み係数を所定の範囲内に収めることで、当該重み係数を用いて算出される畳み込み演算処理の処理結果データを所定の範囲内に収めることができ、処理結果データのぶれを抑えることができる。よって、処理結果データを、浮動小数点数より

も表現範囲が狭い固定小数点数によっても表現することができ、また、固定小数点数による処理により、より高速で精度の良い演算が可能となる。

[0034] なお、図 1 5 に例示するように、演算処理装置 5 0 は、さらに、記憶部 5 7、乗算処理部 5 8 を備える構成とするとよい。この構成において、記憶部 5 7 は、重み調整処理部 5 6 が特定した重み係数の最大値の絶対値を記憶する。乗算処理部 5 8 は、例えば乗算器で構成されており、畳み込み演算処理部 5 1 が出力する処理結果データに、記憶部 5 7 に記憶されている重み係数の最大値の絶対値を乗算して、活性化処理部 5 9 に出力する。活性化処理部 5 9 は、入力される処理結果データに対して周知の活性化処理を実行して、その処理結果データをプーリング処理部 5 2 に出力する。

[0035] この構成は、活性化処理に用いる活性化関数が非線形である場合に特に効果を発揮する。即ち、図 1 6 に例示する非線形の活性化関数により処理結果データにおいては、畳み込み演算の重み係数を調整しない場合の処理結果 R 1 付近における処理結果の変化量と、畳み込み演算の重み係数を調整した場合の処理結果 R 2 付近における処理結果の変化量とが異なってくる。そのため、演算処理装置 5 0 は、重み係数を調整した場合の処理結果データ R 2 に対し、重み係数の調整の際に用いた最大値の絶対値を乗算することにより、矢印 r で示すように、畳み込み演算処理の処理結果データを、重み係数を調整しない場合の処理結果データ R 1 に戻すようにした。これにより、調整された重み係数により高速で精度の良い畳み込み演算を行いつつも、その処理結果データを、重み係数を調整しなかった場合の処理結果データに近似させることができ、特に活性化関数が非線形である場合であっても、その影響を抑えることができる。

[0036] (第 6 実施形態)

図 1 7 に例示する演算処理装置 6 0 は、畳み込み演算処理部 6 1、プーリング処理部 6 2、積算処理部 6 3、正規化処理部 6 5 を有する演算ブロック 6 0 0 を複数備える。演算ブロック 6 0 0 には、それぞれ、加算処理部 6 6、記憶部 6 7 が備えられている。また、複数の演算ブロック 6 0 0 のうちの

何れか1つ、この場合、最上流の演算ブロック600には、統計処理部64が備えられている。なお、説明の便宜上、図の下側を下流側、図の上側を上流側と定義する。

[0037] 加算処理部66は、自身と同じ演算ブロック600を構成する積算処理部63から入力されるデータに、自身よりも下流側の演算ブロック600から入力されるデータを加算して、記憶部67に出力する。記憶部67は、加算処理部66から入力されるデータを記憶するとともに、そのデータを自身よりも上流側の演算ブロック600の加算処理部66に出力する。最も上流側の演算ブロック600の記憶部67は、記憶したデータを統計処理部64に出力する。統計処理部64は、入力されたデータについて平均値と標準偏差を算出し、各演算ブロック600の正規化処理部65にそれぞれ出力する。即ち、統計処理部64は、複数の積算処理部63が算出するデータの平均値および標準偏差を算出して、複数の正規化処理部65にそれぞれ提供する。

[0038] 演算処理装置60によれば、複数の演算ブロック600を備える構成において、各演算ブロック600にそれぞれ統計処理部64を設けるのではなく、1つの演算ブロック600のみに統計処理部64を設け、当該統計処理部64が出力する統計値を複数の正規化処理部65により共用するようにした。この構成によれば、複数の演算ブロック600を備える場合であっても、回路規模が比較的大きくなる統計処理部64を1つだけに抑えることができる、装置全体としてコンパクト化や低コスト化を図ることができる。

[0039] (その他の実施形態)

なお、本開示は、上述した実施形態に限定されるものではなく、その要旨を逸脱しない範囲で種々の実施形態に適用可能である。例えば、上述した複数の実施形態を適宜組み合わせる実施してもよい。

[0040] 本開示は、実施例に準拠して記述されたが、本開示は当該実施例や構造に限定されるものではないと理解される。本開示は、様々な変形例や均等範囲内の変形をも包含する。加えて、様々な組み合わせや形態、さらには、それらに一要素のみ、それ以上、あるいはそれ以下、を含む他の組み合わせや形

態をも、本開示の範疇や思想範囲に入るものである。

請求の範囲

[請求項1] 複数の処理層が階層的に接続されたニューラルネットワークによる演算を実行する演算処理装置（１０，２０，３０，４０，５０，６０）であって、

前階層から入力される入力データに対して畳み込み演算処理を実行する畳み込み演算処理部（１１，２１，３１，４１，５１，６１）と、

前記畳み込み演算処理部による処理結果データに対してプーリング処理を実行するプーリング処理部（１２，２２，３２，４２，５２，６２）と、

次回層の演算が開始される前に、前記プーリング処理部による処理結果データを積算する積算処理部（１３，２３，３３，４３，５３，６３）と、

前記積算処理部が積算するデータの平均値および標準偏差を算出する統計処理部（１４，２４，３４，４４，５４，６４）と、

次回層の演算が開始されると、前階層において前記統計処理部により算出された平均値および標準偏差を用いて、前階層において前記プーリング処理部により出力された処理結果データに対して正規化処理を実行する正規化処理部（１５，２５，３５，４５，５５，６５）と、

を備える演算処理装置。

[請求項2] 前記正規化処理部（２５）は、減算処理部（２５ａ）および除算処理部（２５ｂ）を備え、

前記減算処理部は、前階層において前記統計処理部（２４）により算出された平均値を、前階層において前記プーリング処理部（２２）により出力された処理結果データから減算し、

前記畳み込み演算処理部（２１）は、前記減算処理部が出力する処理結果データに対して畳み込み演算処理を実行し、

前記除算処理部は、前記畳み込み演算処理部が出力する処理結果データを、前階層において前記統計処理部（２４）により算出された標準偏差により除算する請求項１に記載の演算処理装置。

[請求項３] 前記積算処理部（３３）は、前記プーリング処理部（３２）による処理結果データの一部を積算する請求項１または２に記載の演算処理装置。

[請求項４] 前記積算処理部（４３）は、前記プーリング処理部（４２）による処理結果データの上位の所定ビットを積算する請求項１から３の何れか１項に記載の演算処理装置。

[請求項５] 前記畳み込み演算処理部（６１）、前記プーリング処理部（６２）、前記積算処理部（６３）、前記正規化処理部（６５）を有する演算ブロック（６００）が複数備えられ、

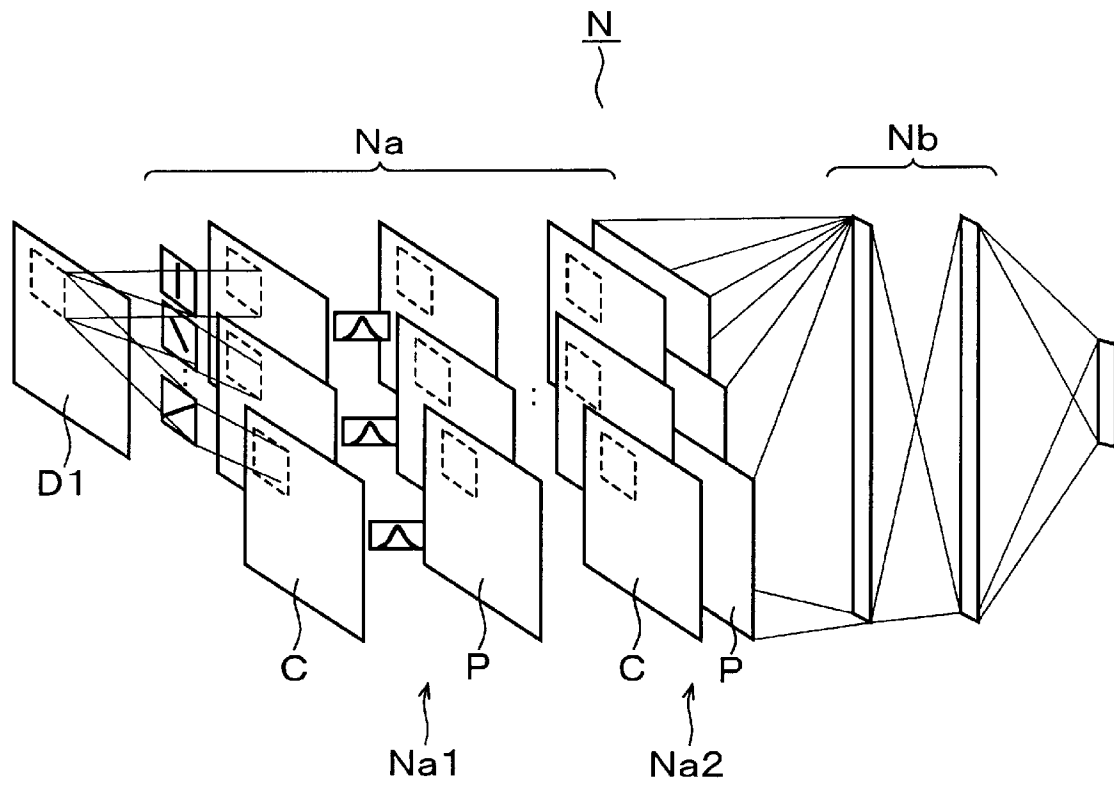
前記統計処理部（６４）は、複数の前記積算処理部（６３）が算出するデータの平均値および標準偏差を算出する請求項１から４の何れか１項に記載の演算処理装置。

[請求項６] 前記畳み込み演算処理部（５１）、前記プーリング処理部（５２）、前記積算処理部（５３）、前記正規化処理部（５５）を有する演算ブロック（５００）が複数備えられ、

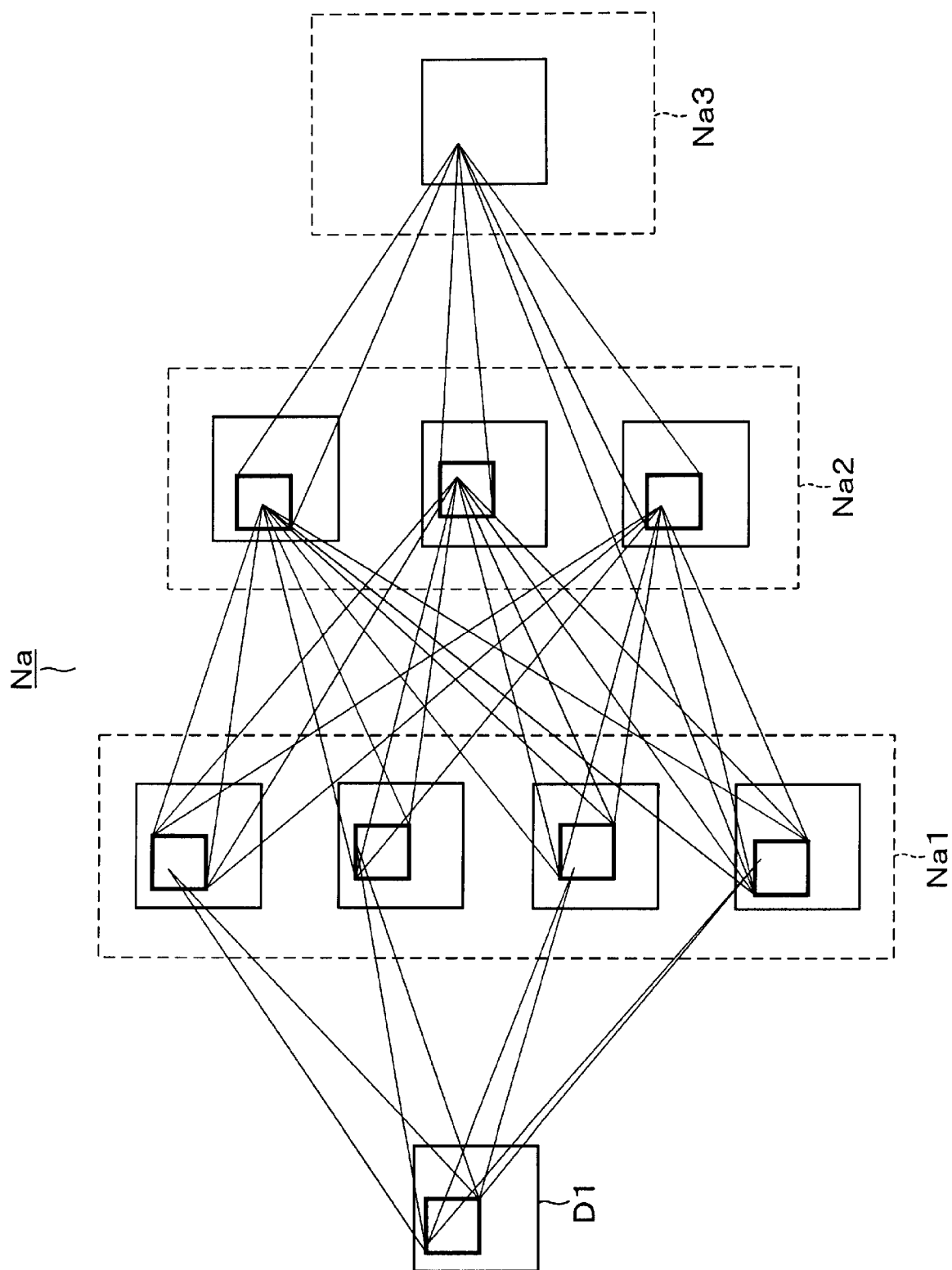
複数の前記畳み込み演算処理部がそれぞれ畳み込み演算処理を実行する際に用いる重み係数の最大値を特定し、その特定した最大値の絶対値により、複数の前記畳み込み演算処理部が用いる重み係数を除算する重み調整処理部（５６）を備える請求項１から５の何れか１項に記載の演算処理装置。

[請求項７] 前記畳み込み演算処理部（５１）が前記重み調整処理部により調整された重み係数を用いて演算した演算結果データに、前記重み係数の最大値の絶対値を乗算する乗算処理部（５８）を備える請求項６に記載の演算処理装置。

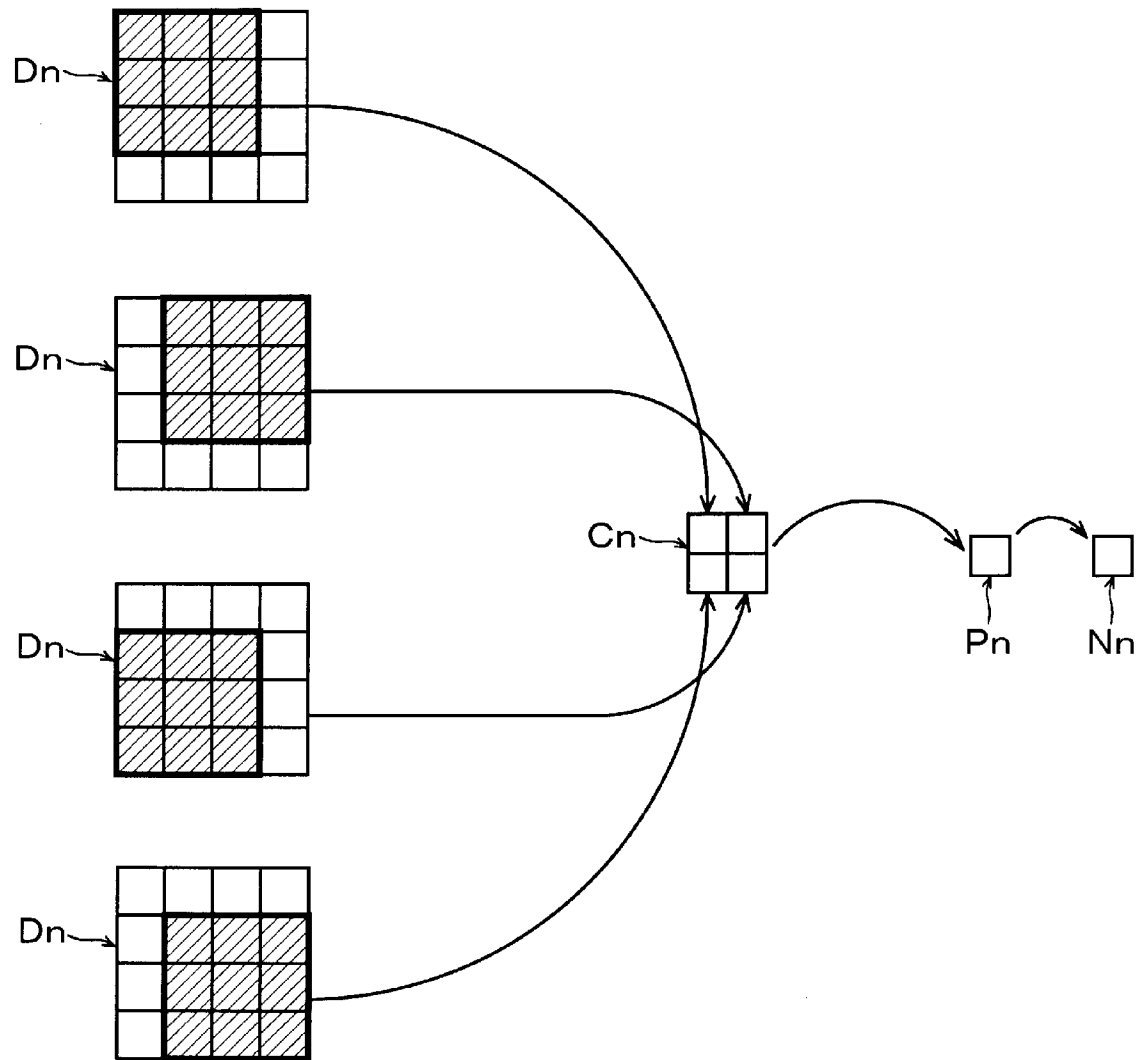
[図1]



[図2]



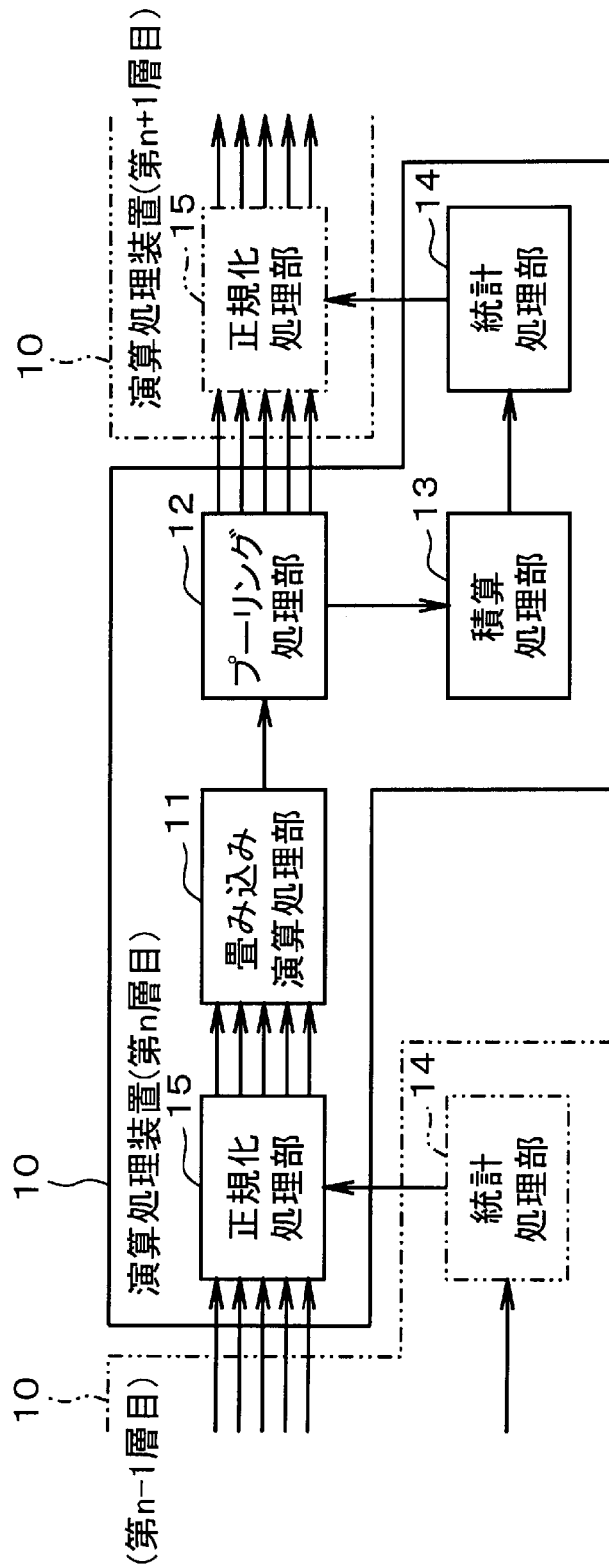
[図3]



[図4]

処理分類	関数名	関数式
畳み込み演算	畳み込み関数	$Y_{ij} = \sum_{p=0}^N \sum_{q=0}^N X_{i+p, j+q} \times W_{p,q}$ $(i, j = 1, 2, \dots)$
活性化処理	ロジスティックシグモイド	$f(y_j) = \frac{1}{1 + e^{-y_j}}$
	Rectified-Liner-Units (ReLU)	$f(y_j) = \max(y_j, 0)$
プーリング処理	最大プーリング	$h' = \max_j h_j$
	平均プーリング	$h' = \frac{1}{N} \sum_j h_j$

[図5]

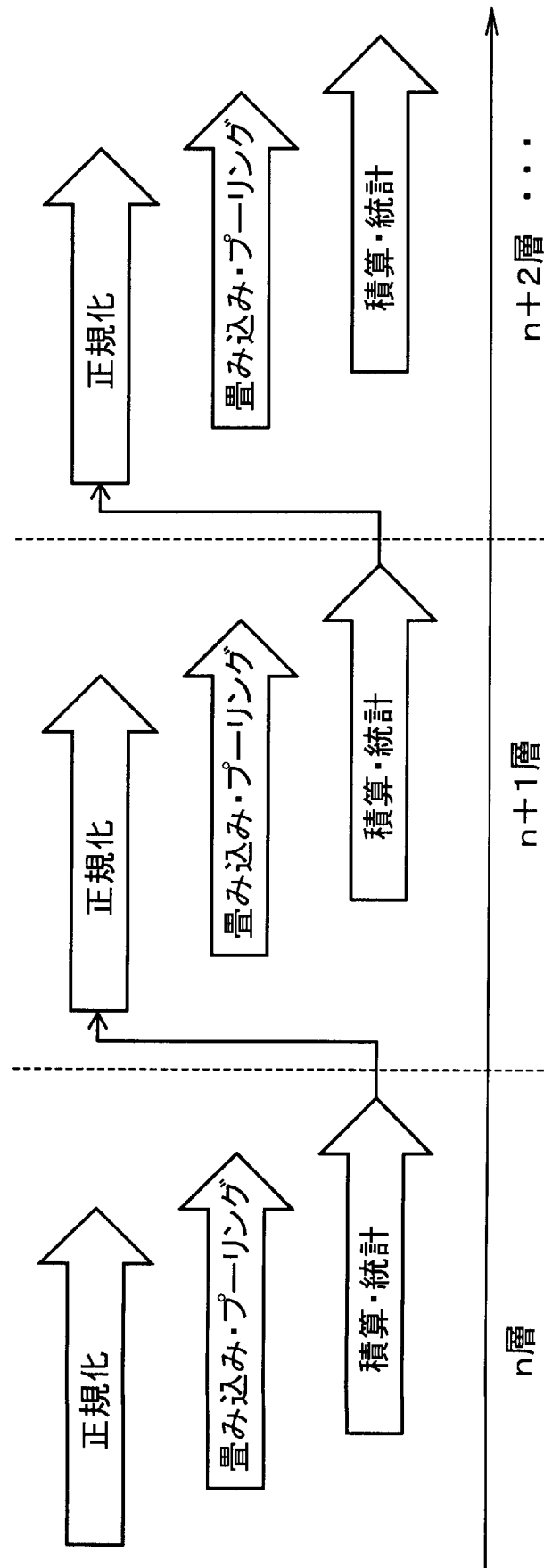


[図6]

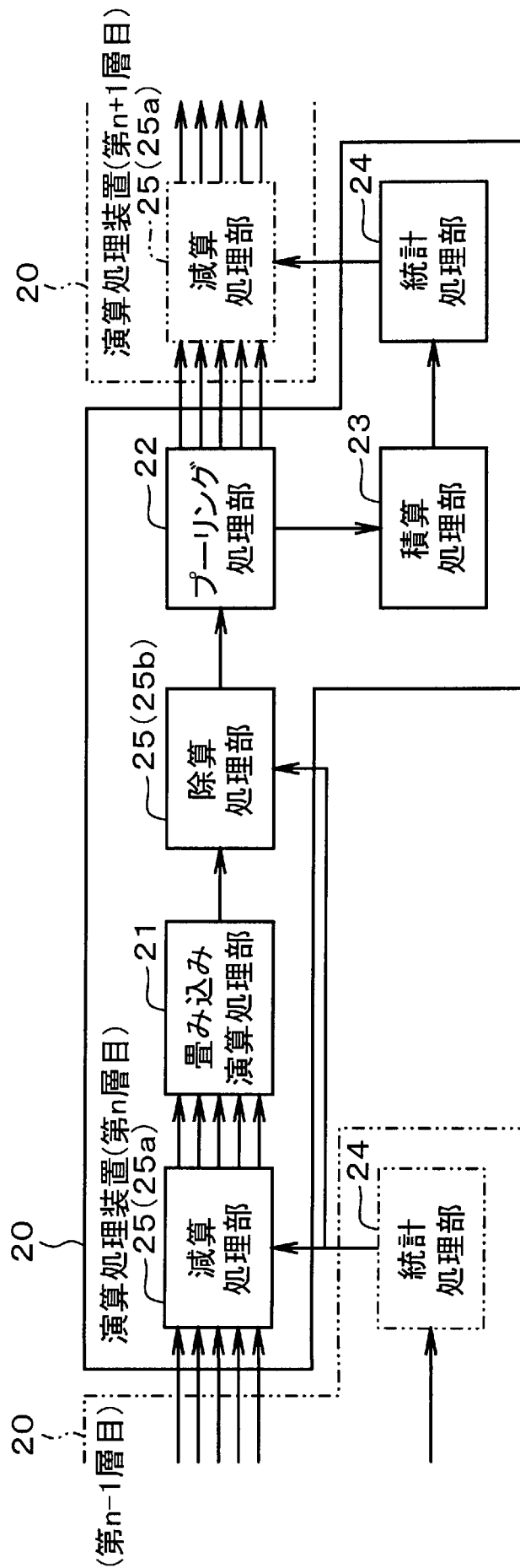
$$Y_i = \frac{(X_i - \bar{X})}{SD_x}$$

 $\bar{X}$: 平均値SD_x: 標準偏差

[図7]



[図8]



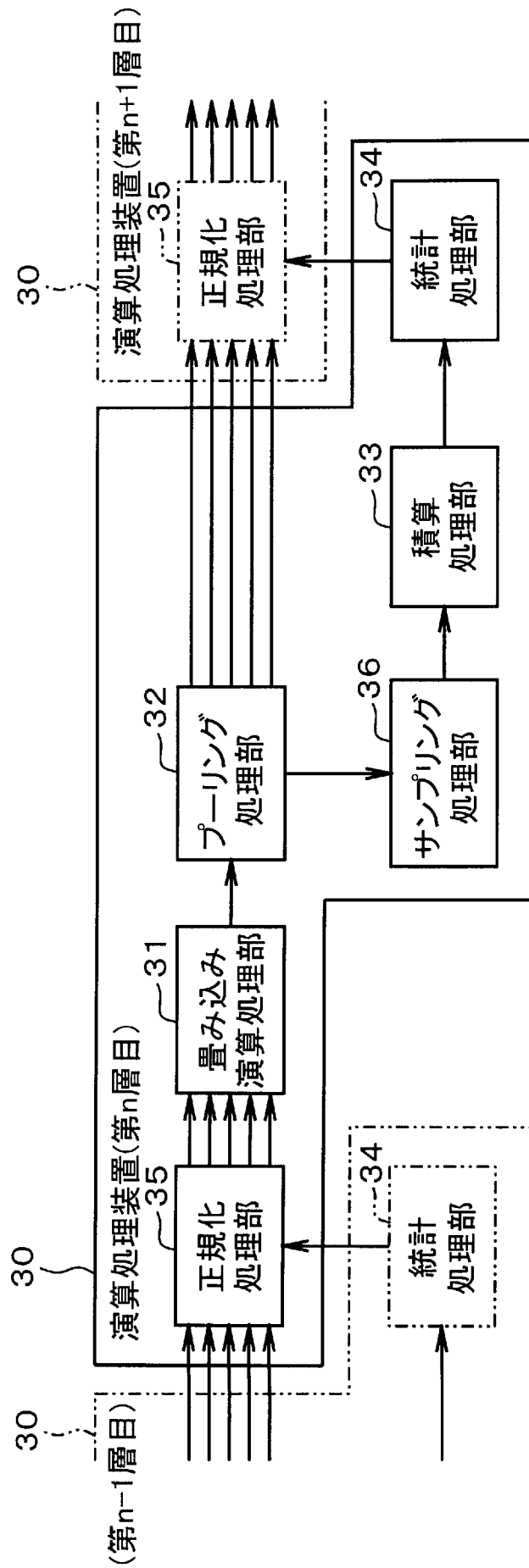
[図9]

$$Yi' = (Xi - \bar{X}) \quad \dots (1)$$

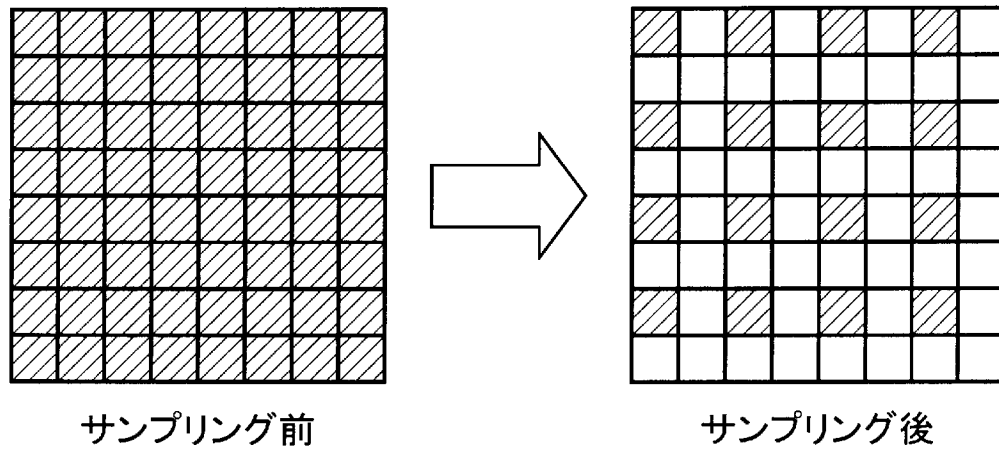
$$Yi'' = \frac{C}{SDx} \quad \dots (2)$$

 $\bar{X}$: 平均値SD_x: 標準偏差

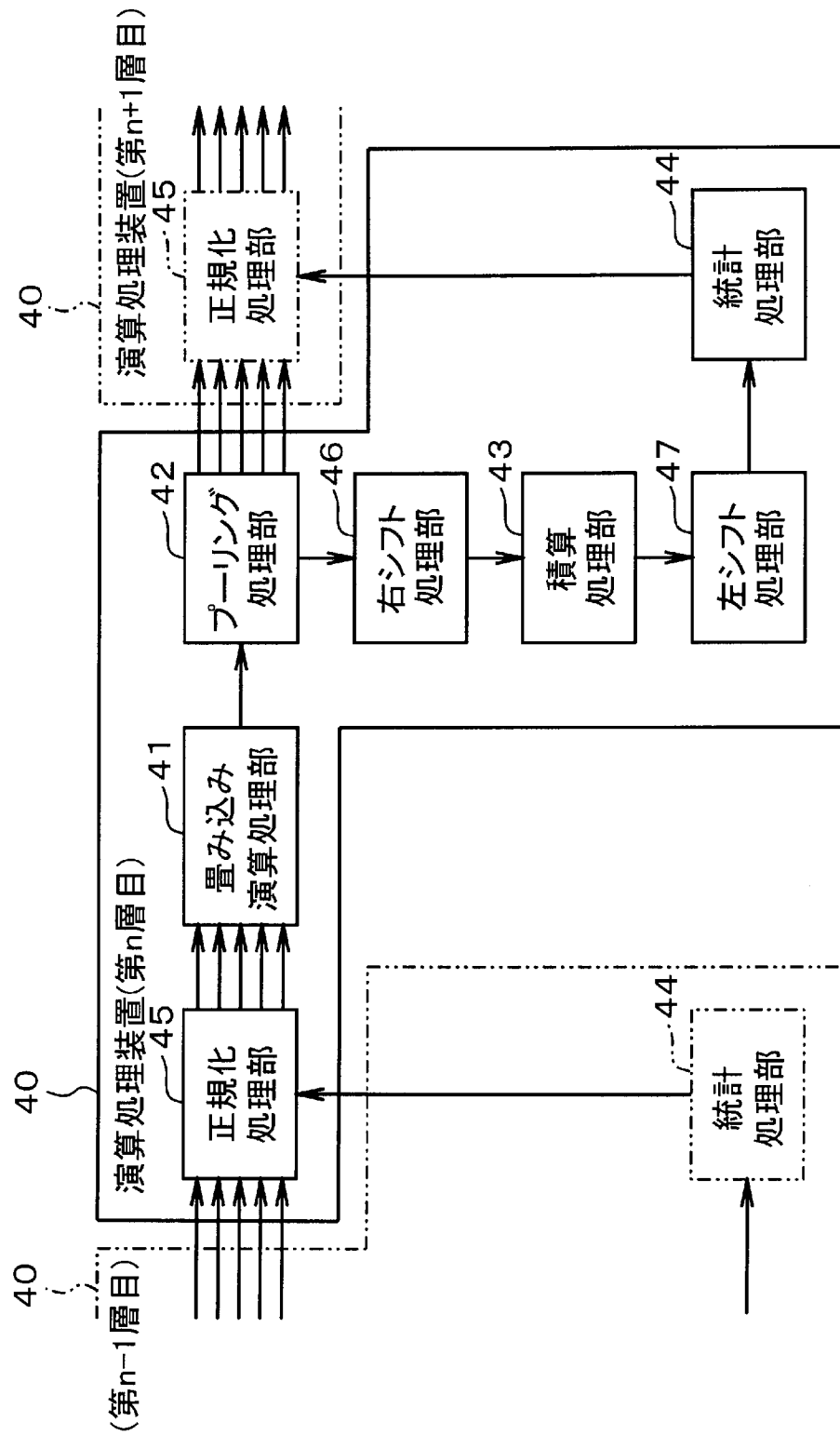
[図10]



[図11]



[図12]

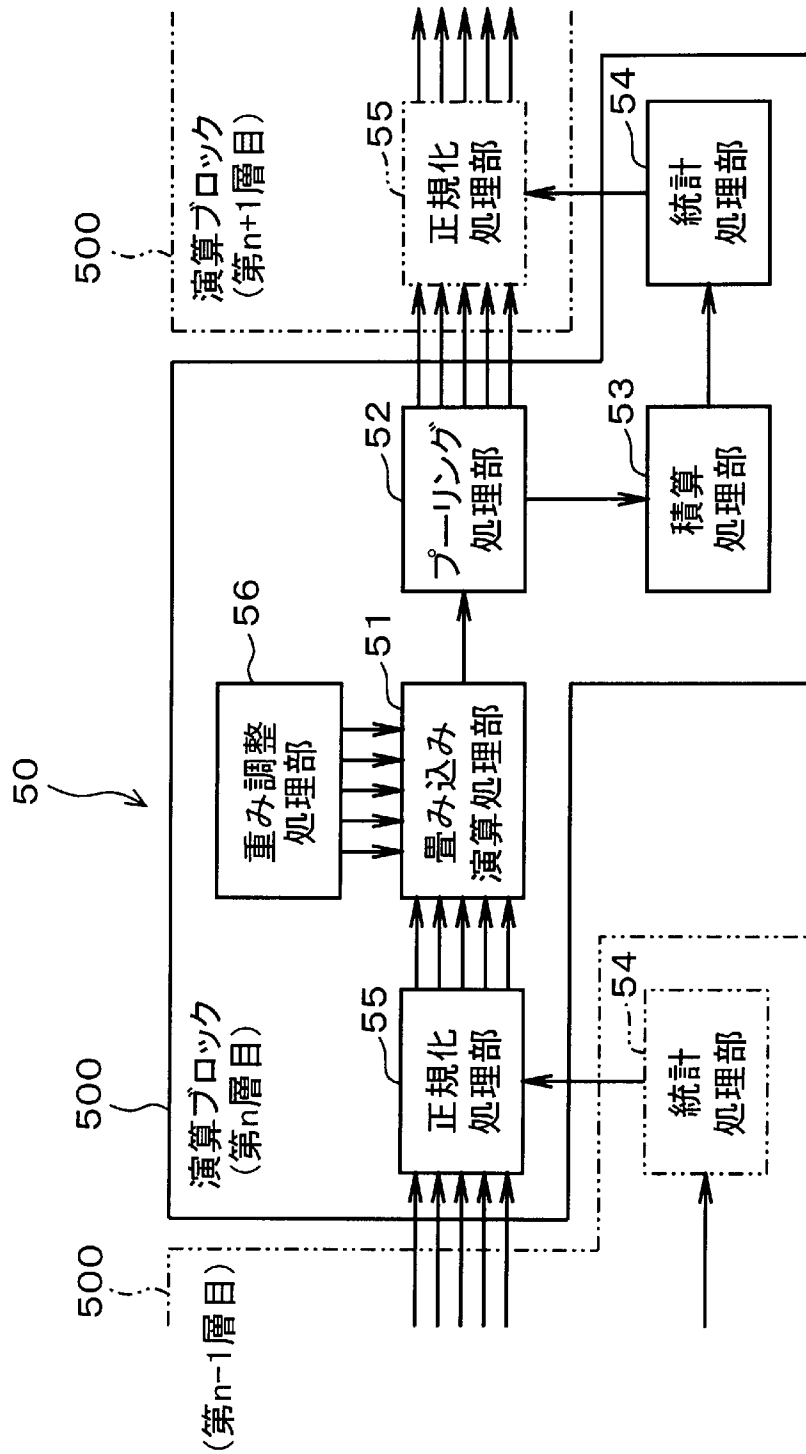


[図13]

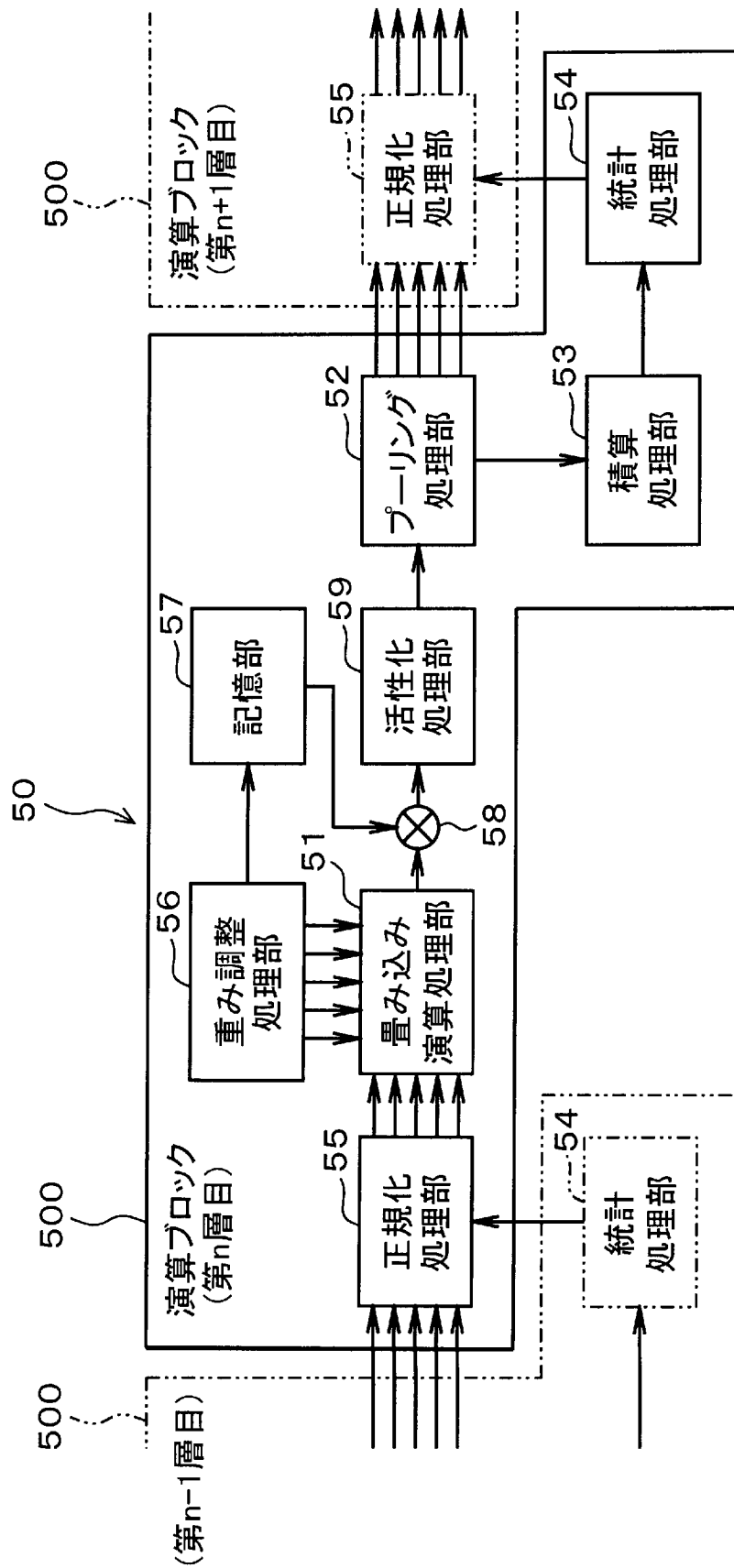
0000001100001100

上位ビット

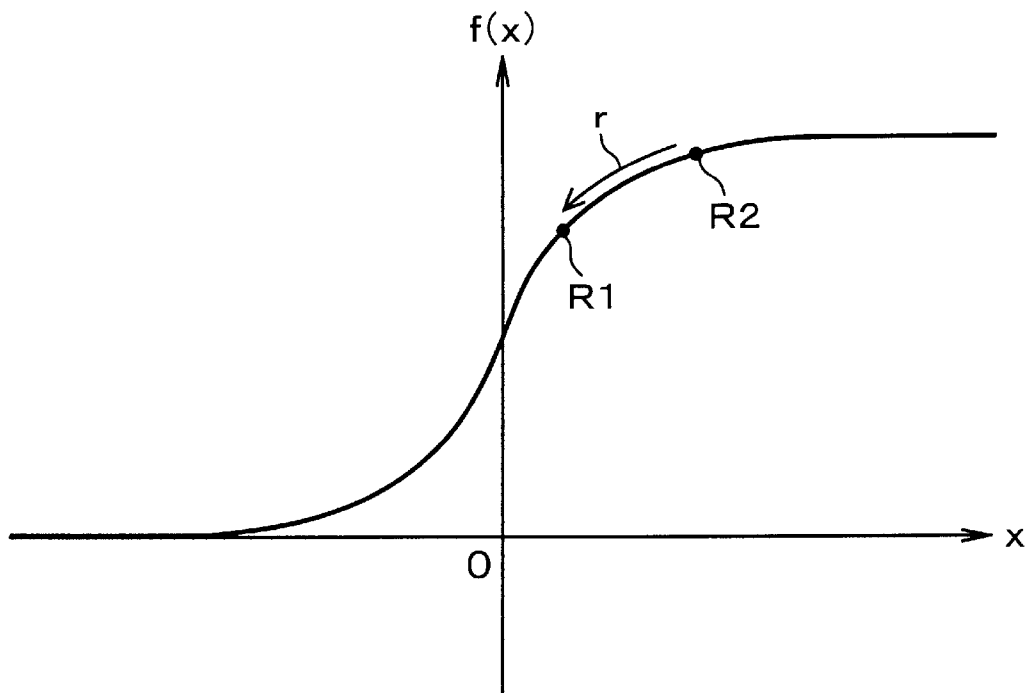
[図14]



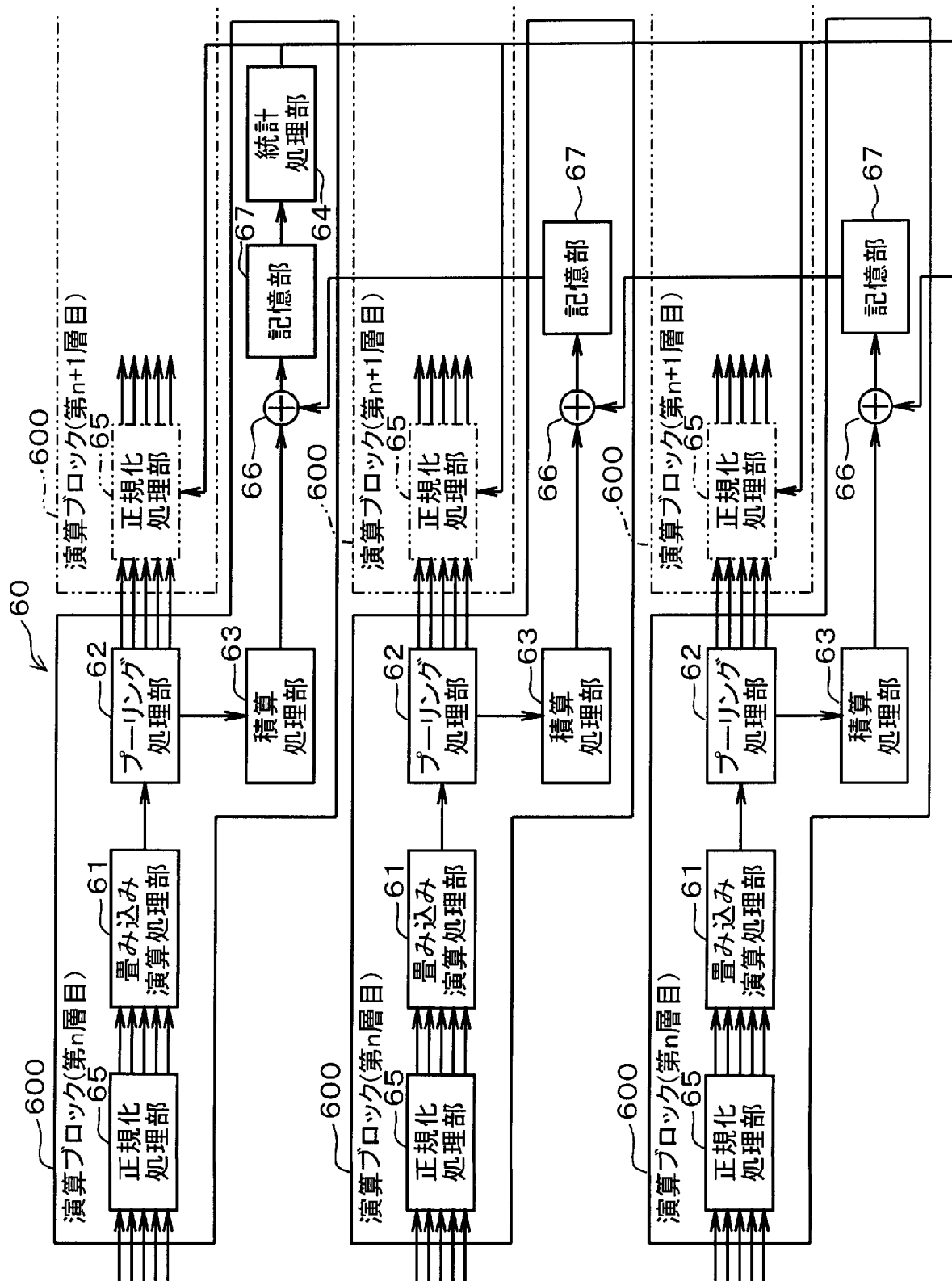
[図15]



[図16]



[図17]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2017/001554

A. CLASSIFICATION OF SUBJECT MATTER

G06N3/04(2006.01)i, G06T7/00(2017.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N3/04, G06T7/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Jitsuyo Shinan Koho	1922-1996	Jitsuyo Shinan Toroku Koho	1996-2017
Kokai Jitsuyo Shinan Koho	1971-2017	Toroku Jitsuyo Shinan Koho	1994-2017

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2013-069132 A (Fuji Xerox Co., Ltd.), 18 April 2013 (18.04.2013), paragraphs [0033] to [0035], [0052]; fig. 4, 10 to 11 (Family: none)	1-7
A	JP 2015-215837 A (Denso Corp.), 03 December 2015 (03.12.2015), claim 1; paragraphs [0020] to [0036]; fig. 4 to 5 & US 2015/0331832 A1 claim 1; paragraphs [0032] to [0051]; fig. 4 to 5	1-7

☐ Further documents are listed in the continuation of Box C.

☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search
11 April 2017 (11.04.17)

Date of mailing of the international search report
18 April 2017 (18.04.17)

Name and mailing address of the ISA/
Japan Patent Office
3-4-3, Kasumigaseki, Chiyoda-ku,
Tokyo 100-8915, Japan

Authorized officer

Telephone No.

A. 発明の属する分野の分類（国際特許分類（IPC））

Int.Cl. G06N3/04(2006.01)i, G06T7/00(2017.01)i

B. 調査を行った分野

調査を行った最小限資料（国際特許分類（IPC））

Int.Cl. G06N3/04, G06T7/00

最小限資料以外の資料で調査を行った分野に含まれるもの

日本国実用新案公報	1922-1996年
日本国公開実用新案公報	1971-2017年
日本国実用新案登録公報	1996-2017年
日本国登録実用新案公報	1994-2017年

国際調査で使用した電子データベース（データベースの名称、調査に使用した用語）

C. 関連すると認められる文献

引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	JP 2013-069132 A（富士ゼロックス株式会社） 2013.04.18, 段落 [0033] - [0035], [0052], 図4, 10-11 (ファミリーなし)	1-7

☒ C欄の続きにも文献が列挙されている。

☐ パテントファミリーに関する別紙を参照。

* 引用文献のカテゴリー

「A」特に関連のある文献ではなく、一般的技術水準を示すもの

「E」国際出願日前の出願または特許であるが、国際出願日以後に公表されたもの

「L」優先権主張に疑義を提起する文献又は他の文献の発行日若しくは他の特別な理由を確立するために引用する文献（理由を付す）

「O」口頭による開示、使用、展示等に言及する文献

「P」国際出願日前で、かつ優先権の主張の基礎となる出願

の日の後に公表された文献

「T」国際出願日又は優先日後に公表された文献であって出願と矛盾するものではなく、発明の原理又は理論の理解のために引用するもの

「X」特に関連のある文献であって、当該文献のみで発明の新規性又は進歩性がないと考えられるもの

「Y」特に関連のある文献であって、当該文献と他の1以上の文献との、当業者にとって自明である組合せによって進歩性がないと考えられるもの

「&」同一パテントファミリー文献

国際調査を完了した日

11.04.2017

国際調査報告の発送日

18.04.2017

国際調査機関の名称及びあて先

日本国特許庁（ISA/J P）

郵便番号100-8915

東京都千代田区霞が関三丁目4番3号

特許庁審査官（権限のある職員）

佐藤 卓馬

5H

5581

電話番号 03-3581-1101 内線 3531

C (続き) . 関連すると認められる文献		
引用文献の カテゴリー*	引用文献名 及び一部の箇所が関連するときは、その関連する箇所の表示	関連する 請求項の番号
A	JP 2015-215837 A (株式会社デンソー) 2015.12.03, [請求項1], 段落 [0020] – [0036], 図4 – 5 & US 2015/0331832 A1, [請求項1], 段落 [0032] – [0051], 図4 – 5	1-7