

(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
2020年10月1日(01.10.2020)



(10) 国際公開番号

WO 2020/194792 A1

- (51) 国際特許分類:
G06F 16/43 (2019.01)
- (21) 国際出願番号: PCT/JP2019/035526
- (22) 国際出願日: 2019年9月10日(10.09.2019)
- (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 2019-059437 2019年3月26日(26.03.2019) JP
- (71) 出願人: 日本電信電話株式会社 (NIPPON TELEGRAPH AND TELEPHONE CORPORATION) [JP/JP]; 〒1008116 東京都千代田区大手町一丁目5番1号 Tokyo (JP).
- (72) 発明者: オンカー クリシュナ (ONKAR, Krishna); 〒1808585 東京都武蔵野市緑町3丁

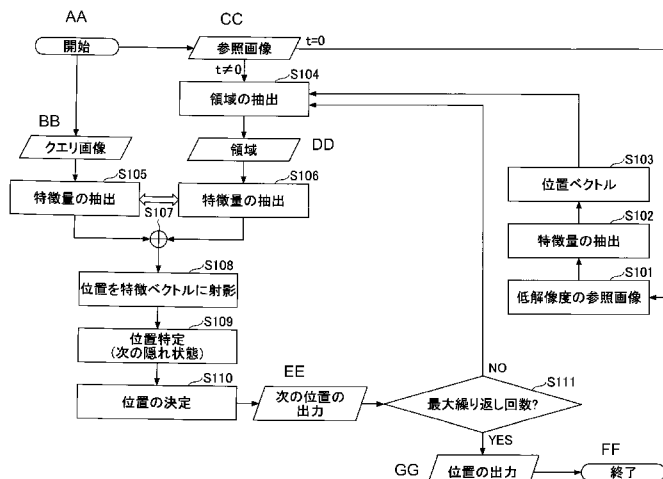
目9-11 NTT知的財産センタ内 Tokyo (JP). 入江 豪(IRIE, Go); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP). 武小萌(WU, Xiaomeng); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP). 川西 隆仁(KAWANISHI, Takahito); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP). 柏野 邦夫(KASHINO, Kunio); 〒1808585 東京都武蔵野市緑町3丁目9-11 NTT知的財産センタ内 Tokyo (JP).

(74) 代理人: 伊東 忠重, 外(ITO, Tadashige et al.); 〒1000005 東京都千代田区丸の内二丁目1番1号 丸の内 M Y P L A Z A (明治安田生命ビル) 16階 Tokyo (JP).

(54) Title: SEARCH DEVICE, LEARNING DEVICE, SEARCH METHOD, LEARNING METHOD, AND PROGRAM

(54) 発明の名称: 検索装置、学習装置、検索方法、学習方法及びプログラム

[図4]



S101 Low-resolution reference image
S102, S105, S106 Extract feature quantity
S103 Position vector
S104 Extract region
S108 Project position onto feature vector
S109 Specify position (next hidden state)
S110 Determine position
S111 Is maximum number of times of repetition reached?
AA Start
BB Query image
CC Reference image
DD Region
EE Output next position
FF End
GG Output position

(57) Abstract: Provided is a search device for searching for media data by determining a target region matching query data, the search device comprising: a first feature extracting unit which extracts a first feature vector from the query data, using a first trained neural network; a second feature extracting unit which acquires a first region from the media data and, using a second trained neural network, extracts a second feature vector from the first region; a position specifying unit which, using a third trained neural network and on the basis of the first feature vector, the second feature vector, and the first region or the position of the first region, determines a candidate for the target region; and a control unit which, by using the determined candidate for the target region as the first region used by the second feature extracting unit, repeats the operations of the second feature extracting unit and the position specifying unit until a predetermined condition is satisfied.

(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, JO, JP, 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), ユーラシア (AM, AZ, BY, KG, KZ, RU, TJ, TM), ヨーロッパ (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:

一 国際調査報告(条約第21条(3))

(57) 要約: クエリデータに一致する目標領域を求めてメディアデータを検索する検索装置は、第1の学習済みニューラルネットワークを用いて、前記クエリデータから第1の特徴ベクトルを抽出する第1の特徴抽出部と、前記メディアデータから第1の領域を取得し、第2の学習済みニューラルネットワークを用いて、前記第1の領域から第2の特徴ベクトルを抽出する第2の特徴抽出部と、第3の学習済みニューラルネットワークを用いて、前記第1の特徴ベクトルと、前記第2の特徴ベクトルと、前記第1の領域又は前記第1の領域の位置とに基づいて、前記目標領域の候補を決定する位置特定部と、前記決定された目標領域の候補を、前記第2の特徴抽出部により用いられる前記第1の領域として使用することにより、所定の条件が満たされるまで、前記第2の特徴抽出部及び前記位置特定部の動作を繰り返す制御部とを有する。

明 細 書

発明の名称：

検索装置、学習装置、検索方法、学習方法及びプログラム

技術分野

[0001] 本発明は、クエリデータに一致する目標領域を求めてメディアデータを検索するために用いることができる検索装置、学習装置、検索方法、学習方法及びプログラムに関する。

背景技術

[0002] 画像キャプチャ技術の近年の進展によって、様々なフォーマットの大量の画像が世界中で日々利用可能になっている。さらに、これらの画像の量、複雑性、多様性及び次元も増加している。画像の増加に伴い、参照画像の中の多数の候補領域からクエリ画像に一致する目標領域を効率的且つ正確に見つけるための要求が増加している。これは、画像の位置合わせ、検証、トラッキング、ステレオ画像処理、圧縮、画像の繋ぎ合わせ及びレンダリングのような広範囲のアプリケーションにおける基本的な課題である。

[0003] この課題を解決するための望ましい方法は、複雑背景 (background clutter)、オクルージョン及び幾何学変換のような実世界のシナリオで発生する状況において正確な一致を見つけることができるように、十分にロバストな方法であることが望まれる。また、妥当な時間内に参照画像の中の多数の候補領域からクエリ画像を特定することができるように、十分に高速な方法であることが望まれる。

[0004] スライディングウィンドウを用いた網羅的な検索方法は、十分な一致精度を提供する。しかし、網羅的な検索方法は、多数のウィンドウを評価する必要があるため、しばしば高コストとなる。既存の方法のほとんどは、最終的な結果に影響及ぼさないようなウィンドウ又は画素をスキップすることによって実行時間を減少させることを目的として、枝刈り (pruning) の概念を取り入れることによりこの問題を克服している。

[0005] 非特許文献1は、ウィンドウを1画素より多く適応的にスライドさせることにより、実行時間を減少させる。スライドの量についての判断は、パターンにおける特徴量毎に定義されたランク (rank) に基づく。非特許文献2は、画素のペアの部分集合 (通常では小さい部分集合) を用いることで、複雑背景に対する頑健性と高速性を担保する方法を提案している。非特許文献3は、幾何変換と距離近似法及び分岐限定法 (branch-and-bound search) のランダムなサンプリングを組み合わせることにより、検索処理を高速化している。非特許文献4は、主成分方向の差分特徴量 (principal orientation difference features) に基づいて一致しない位置をスキップすることにより、計算コストを減少させている。

先行技術文献

非特許文献

[0006] 非特許文献1: Pele O, Werman M, "Accelerating pattern matching or how much can you slide?", In Asian Conference on Computer Vision, 2007 Nov. 18 (pp. 435-446), Springer, Berlin, Heidelberg.

非特許文献2: Dekel T, Oron S, Rubinstein M, Avidan S, Freeman WT, "Best-buddies similarity for robust template matching", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015 (pp. 2021-2029).

非特許文献3: Korman S, Reichman D, Tsur G, Avidan S, "Fast-match: Fast affine template matching", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013 (pp. 2331-2338).

非特許文献4: Jiao J, Wang X, Deng Z, Cao J, Tang W, "A fast template matching algorithm based on principal orientation difference", International Journal of Advanced Robotic Systems, 2018 May 25; 15(3):17298-17308.

非特許文献5: Ba J, Mnih V, Kavukcuoglu K, "Multiple object recognition with visual attention", arXiv preprint arXiv:1412.7755, 2014 Dec. 2

4.

非特許文献6: Ablavatski A, Lu S, Cai J, "Enriched deep recurrent visual attention model for multiple object recognition", In Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, 2017 Mar. 24 (pp. 971-978).

非特許文献7: Mnih V, Heess N, Graves A, "Recurrent models of visual attention", In Advances in neural information processing systems, 2014 (pp. 2204-2212).

非特許文献8: Hel-Or Y, Hel-Or H, David E, "Fast template matching in non-linear tone-mapped images", In Computer Vision (ICCV), 2011 IEEE International Conference on, 2011 Nov. 6 (pp. 1355-1362).

発明の概要

発明が解決しようとする課題

[0007] 上記の非特許文献1～4に記載の枝刈りに基づく画像検索方法は、クエリ画像に一致する目標領域を特定するために、依然として多数のウィンドウ又は画像を評価する必要がある、これは全体の検索処理を非効率にしている。

[0008] さらに、上記の方法は、目標領域を特定しつつ、目標領域を特定するための検索パス（すなわち、評価対象のウィンドウ又は画素の順序）を最適化することを考慮していない。したがって、上記の方法は、例えば、複雑背景、オクルージョン、幾何学変換等に対してロバストではない。

[0009] 評価対象のウィンドウ又は画素のような候補領域の数を減少させることは、画像検索用途だけでなく、動画像や音響信号のようなメディアデータを扱う他の用途においても望まれる。

[0010] 本発明は、上記の課題に鑑みてなされたものであり、本発明は、メディアデータにおける評価対象の候補領域の数を減少させ、メディアデータからクエリデータに一致する目標領域をより効率的に見つけることができる検索装置、学習装置、検索方法、学習方法及びプログラムを提供することを目的とする。

課題を解決するための手段

[0011] 本発明の一形態は、クエリデータに一致する目標領域を求めてメディアデータを検索する検索装置を提供し、当該検索装置は、

第1の学習済みニューラルネットワークを用いて、前記クエリデータから第1の特徴ベクトルを抽出する第1の特徴抽出部と、

前記メディアデータから第1の領域を取得し、第2の学習済みニューラルネットワークを用いて、前記第1の領域から第2の特徴ベクトルを抽出する第2の特徴抽出部と、

第3の学習済みニューラルネットワークを用いて、前記第1の特徴ベクトルと、前記第2の特徴ベクトルと、前記第1の領域又は前記第1の領域の位置とに基づいて、前記目標領域の候補を決定する位置特定部と、

前記決定された目標領域の候補を、前記第2の特徴抽出部により用いられる前記第1の領域として使用することにより、所定の条件が満たされるまで、前記第2の特徴抽出部及び前記位置特定部の動作を繰り返す制御部と、を有する。

[0012] 本発明の他の形態は、クエリデータに一致する目標領域を求めてメディアデータを検索するために用いられるニューラルネットワークを学習する学習装置を提供し、当該学習装置は、

第1のニューラルネットワークを用いて、学習用クエリデータから第1の特徴ベクトルを抽出する第1の特徴抽出部と、

学習用メディアデータから第1の領域を取得し、第2のニューラルネットワークを用いて、前記第1の領域から第2の特徴ベクトルを抽出する第2の特徴抽出部と、

第3のニューラルネットワークを用いて、前記第1の特徴ベクトルと、前記第2の特徴ベクトルと、前記第1の領域又は前記第1の領域の位置とに基づいて、前記目標領域の候補を決定する位置特定部と、

前記決定された目標領域の候補を、前記第2の特徴抽出部により用いられる前記第1の領域として使用することにより、所定の条件が満たされるまで

、前記第2の特徴抽出部及び前記位置特定部の動作を繰り返す制御部と、

前記位置特定部により決定された前記目標領域の候補が、前記目標領域を捉えたか否かを判定し、当該判定結果に基づいて、前記第1のニューラルネットワーク、前記第2のニューラルネットワーク及び前記第3のニューラルネットワークのパラメータを更新する学習部と、

を有する。

[0013] 本発明の他の形態は、クエリデータに一致する目標領域を求めてメディアデータを検索する検索装置により使用される検索方法を提供し、当該検索方法は、

第1の学習済みニューラルネットワークを用いて、前記クエリデータから第1の特徴ベクトルを抽出する第1のステップと、

前記メディアデータから第1の領域を取得し、第2の学習済みニューラルネットワークを用いて、前記第1の領域から第2の特徴ベクトルを抽出する第2のステップと、

第3の学習済みニューラルネットワークを用いて、前記第1の特徴ベクトルと、前記第2の特徴ベクトルと、前記第1の領域又は前記第1の領域の位置とに基づいて、前記目標領域の候補を決定する第3のステップと、

前記決定された目標領域の候補を、前記第2のステップにより用いられる前記第1の領域として使用することにより、所定の条件が満たされるまで、前記第2のステップ及び前記第3のステップを繰り返す第4のステップと、
を有する。

[0014] 本発明の他の形態は、クエリデータに一致する目標領域を求めてメディアデータを検索するために用いられるニューラルネットワークを学習する学習装置により使用される学習方法を提供し、当該学習方法は、

第1のニューラルネットワークを用いて、学習用クエリデータから第1の特徴ベクトルを抽出する第1のステップと、

学習用メディアデータから第1の領域を取得し、第2のニューラルネットワークを用いて、前記第1の領域から第2の特徴ベクトルを抽出する第2の

ステップと、

第3のニューラルネットワークを用いて、前記第1の特徴ベクトルと、前記第2の特徴ベクトルと、前記第1の領域又は前記第1の領域の位置とに基づいて、前記目標領域の候補を決定する第3のステップと、

前記決定された目標領域の候補を、前記第2のステップにより用いられる前記第1の領域として使用することにより、所定の条件が満たされるまで、前記第2のステップ及び前記第3のステップを繰り返す第4のステップと、

前記目標領域の候補が、前記目標領域を捉えたか否かを判定し、当該判定結果に基づいて、前記第1のニューラルネットワーク、前記第2のニューラルネットワーク及び前記第3のニューラルネットワークのパラメータを更新する第5のステップと、

を有する。

[0015] 本発明の他の形態は、上記の検索装置又は学習装置としてコンピュータを機能させるプログラムを提供する。

発明の効果

[0016] 本発明によれば、メディアデータにおける評価対象の候補領域の数を減少させ、メディアデータからクエリデータに一致する目標領域をより効率的に見つけることが可能になる。

図面の簡単な説明

[0017] [図1]本発明の第1の実施例に係る検索装置の機能構成を示す図である。

[図2]本発明の第1の実施例に係る学習装置の機能構成を示す図である。

[図3]本発明の第2の実施例に係る検索装置の概念図である。

[図4]本発明の第2の実施例に係る画像検索方法のフローチャートである。

[図5]初期領域予測部におけるCNNの詳細図である。

[図6]第1の特徴抽出部及び第2の特徴抽出部におけるCNNの詳細図である。

[図7]本発明の第2の実施例に係る学習方法のフローチャートである。

[図8]画像検索成功率を示す表である。

[図9]クエリ画像を特定するために評価されたウィンドウの数を示す表である。

[図10]クエリ画像を特定するための平均実行時間を示す表である。

[図11]クエリ画像を特定するための検索パスを示す図である。

[図12]本発明の実施例に係る各装置のハードウェア構成例を示す図である。

発明を実施するための形態

[0018] 本発明の実施例について図面を参照して以下に説明する。

[0019] <第1の実施例>

第1の実施例では、クエリデータに一致する目標領域（以下、「目標ウィンドウ」とも呼ばれる）を求めてメディアデータを検索する検索装置の全体構成について説明する。例えば、メディアデータ及びクエリデータは、それぞれ、静止画像、動画像、音響信号又は他のデータである。検索装置は、メディアデータの中の複数の目標領域の候補（以下、「候補領域」又は「候補ウィンドウ」とも呼ばれる）からクエリデータに一致する目標領域を見つけるために、学習装置により学習されたモデル（具体的には、ニューラルネットワーク）を用いる。具体的には、検索装置は、学習済みモデルを用いることにより、クエリデータと候補領域とを繰り返し比較し、候補領域がクエリデータに一致するか否かを判断することによって、目標領域を見つける。モデルは、評価対象の候補領域を減少させるように学習されているため、検索装置は、より少ない数の候補領域で目標領域を見つけることができる。

[0020] 第1の実施例では、検索装置により使用されるモデルを学習する学習装置について更に説明する。学習装置は、検索装置と異なってもよく、検索装置と同じでもよい。以下の説明では、学習装置が検索装置と異なる装置であると仮定する。

[0021] [検索装置]

図1は、本発明の第1の実施例に係る検索装置100の機能構成を示す図である。検索装置100の目的は、メディアデータRの中の特定の位置 l_0 に存在する、クエリデータQによって表される目標領域を特定することである。

位置 l_0 は領域の中心でもよく、領域の角でもよく、領域を定義するために用いることができる他の位置でもよい。メディアデータRが静止画像である場合、位置 l_0 はxy画像座標により表すことができる。メディアデータRが動画である場合、位置 l_0 はタイムスタンプ（又はフレームインデックス）により表すことができる。この場合、位置 l_0 は時間フレームのxy画像座標を含んでもよい。メディアデータRが音響信号である場合、位置 l_0 はタイムスタンプにより表すことができる。検索装置100は、ウィンドウサイズ、方向又は上記のいずれかの組み合わせ等のように、目標領域を定義するために用いられる他の種類の情報を決定してもよい。

[0022] 検索装置100は、クエリデータQとメディアデータRの中の候補領域とを繰り返し比較し、候補領域がクエリデータQに一致するか否かを判断する。位置 l におけるメディアデータRの領域を $R(l)$ と表すとすると、目標領域は $R(l_0)$ により表すことができる。また、繰り返し回数 t における評価対象の候補領域は $R(l_t)$ により表す。以下に説明するように、検索装置100は、より小さい繰り返し回数 T で目標領域が位置特定できるように、学習済みモデルを用いて検索パス $\{l_t\}_{t=0}^T$ （すなわち、候補領域の位置の順序）を決定する。ただし、 T は最大繰り返し回数である。

[0023] 検索装置100は、第1の特徴抽出部120と、第2の特徴抽出部130と、位置特定部140と、制御部150とを含む。検索装置100は、初期領域予測部110を更に含んでもよい。

[0024] 初期領域予測部110は、メディアデータRを入力とし、評価対象の初期領域 $R(l_0)$ 又はその位置 l_0 を出力とするニューラルネットワークである。ニューラルネットワークは畳み込みニューラルネットワーク（CNN：convolutional neural network）でもよく、他のニューラルネットワークでもよい。音響信号の場合、再帰型ニューラルネットワーク（RNN：recurrent neural network）又は長短期記憶ネットワーク（LSTM：long-short term memory）が用いられてもよい。初期領域予測部110は、メディアデータRをダウンサンプリング後のメディアデータ R_{coarse} にダウンサンプリングし、ダウンサ

ンプリング後のメディアデータ R_{coarse} に基づいて初期領域 $R(l_0)$ を取得する。初期領域 $R(l_0)$ は初期位置 l_0 においてメディアデータから抽出される領域であり、初期位置 l_0 はダウンサンプリング後のメディアデータ R_{coarse} の特徴量を位置ベクトルに線形射影することにより取得されてもよい。初期領域 $R(l_0)$ は、第2の特徴抽出部130により評価される最初の候補領域である。

[0025] なお、初期領域予測部110は検索装置100に含まなくてもよく、初期位置 l_0 は任意に決定されてもよい。

[0026] 第1の特徴抽出部120は、クエリデータ Q を入力とし、クエリデータ Q の特徴ベクトル $f(Q)$ を出力とするニューラルネットワークである。ニューラルネットワークはCNNでもよく、他のニューラルネットワークでもよい。音響信号の場合、RNN又はLSTMが用いられてもよい。第1の特徴抽出部120は、クエリデータ Q から特徴ベクトル $f(Q)$ を抽出する。

[0027] 第2の特徴抽出部130は、メディアデータ R 及び候補領域 $R(l_t)$ 又はその位置 l_t ($t=0$ の場合には初期領域 $R(l_0)$ 又はその位置 l_0) を入力とし、メディアデータ R の候補領域の特徴ベクトル $f(R(l_t))$ を出力とするニューラルネットワークである。位置 l_t が第2の特徴抽出部130に入力された場合、第2の特徴抽出部130は、メディアデータ R から位置 l_t における候補領域 $R(l_t)$ を抽出する。第2の特徴抽出部130のニューラルネットワークは、第1の特徴抽出部120のニューラルネットワークと同じであり、第1の特徴抽出部120のニューラルネットワークと同じパラメータを共有する。例えば、双方のニューラルネットワークがCNNである場合、フィルタのサイズ、ストライド等のような同じパラメータを用いる。第2の特徴抽出部130は、候補領域 $R(l_t)$ を取得し、候補領域 $R(l_t)$ から特徴ベクトル $f(R(l_t))$ を抽出する。第2の特徴抽出部130のニューラルネットワークは第1の特徴抽出部120のニューラルネットワークと同じパラメータを共有するため、候補領域 $R(l_t)$ のサイズは、クエリデータ Q のサイズと同じである必要がある。

[0028] 位置特定部140は、特徴ベクトル $f(Q)$ 、特徴ベクトル $f(R(l_t))$ 及び候補領域 $R(l_t)$ 又はその位置 l_t ($t=0$ の場合には初期領域 $R(l_0)$ 又はその位置 l_0) を

入力とし、次の候補領域 $R(l_{t+1})$ 又はその位置 l_{t+1} ($t < T$)、或いは最終結果 ($t = T$ の場合の候補領域 $R(l_T)$ 又はその位置 l_T 、或いはクエリデータ Q に一致する目標領域又はその位置) を出力するニューラルネットワークである。位置特定部 140 のニューラルネットワークは LSTM でもよく、他のニューラルネットワークでもよい。位置特定部 140 は、特徴ベクトル $f(Q)$ と、特徴ベクトル $f(R(l_t))$ と、候補領域 $R(l_t)$ 又はその位置 l_t とに基づいて、次の候補領域 $R(l_{t+1})$ を決定する。より具体的には、位置特定部 140 は、特徴ベクトル $f(Q)$ と特徴ベクトル $f(R(l_t))$ とを単一のベクトルに結合し、その後、結合されたベクトルと、現在の候補領域 $R(l_t)$ 又はその位置 l_t と、位置特定部 140 の現在の内部状態（「隠れ状態」とも呼ばれる）とに基づいて、次の候補位置 $R(l_{t+1})$ 又はその位置 l_{t+1} を決定する。以下に説明するように、最大繰り返し回数 T で制御部 150 が繰り返しを終了した場合、位置特定部 140 は、候補領域 $R(l_T)$ 又はその位置 l_T を出力する。また、候補領域 $R(l_t)$ がクエリデータ Q に一致した場合、位置特定部 140 は、候補領域 $R(l_T)$ がクエリデータ Q に一致する目標領域であると判断し、目標領域又はその位置を出力する。

[0029] 制御部 150 は、次の候補領域 $R(l_{t+1})$ 又はその位置 l_{t+1} 、或いは最終結果を入力とし、繰り返しを終了するか否かを判断する処理部である。制御部 150 は、次の候補領域 $R(l_{t+1})$ 又はその位置 l_{t+1} を第 2 の特徴抽出部 130 に入力し、所定の条件が満たされるまで、第 2 の特徴抽出部 130 及び位置特定部 140 の動作を繰り返す。例えば、制御部 150 は、繰り返し毎に繰り返し回数を 1 だけ増加させ、繰り返し回数 t が所定の制限値 T になった場合、繰り返しを終了する。また、例えば、制御部 150 は、位置特定部 140 によって、候補領域 $R(l_t)$ がクエリデータに一致する目標領域であると判断された場合、繰り返しを終了してもよい。

[0030] 以下に説明するように、第 1 の特徴抽出部 120、第 2 の特徴抽出部 130 及び位置特定部 140（及び初期領域予測部 110 が含まれる場合には初期領域予測部 110）のニューラルネットワークは、学習用クエリデータ及び学習用メディアデータを用いて学習されている。学習用クエリデータが第

1 の特徴抽出部 120 に入力され、学習用メディアデータが第 2 の特徴抽出部 130（又は初期領域予測部 110 が含まれる場合には初期領域予測部 110）に入力された場合、ニューラルネットワークのパラメータは、位置特定部 140 により決定された候補領域 $R(l_{t+1})$ が目標領域を捉えたか否かを判定した結果に基づいて学習されている。また、ニューラルネットワークは、第 1 の特徴抽出部 120 により抽出された特徴ベクトル $f(Q)$ と、第 2 の特徴抽出部 130 により抽出された特徴ベクトル $f(R(l_t))$ との類似度が大きくなるように学習されている。

[0031] [学習装置]

図 2 は、本発明の第 1 の実施例に係る学習装置 200 の機能構成を示す図である。学習装置 200 の目的は、目標領域を特定しつつ、より小さい繰り返し回数 T で検索パス $\{l_t\}_{t=0}^T$ を決定することである。

[0032] 学習装置 200 は、第 1 の特徴抽出部 220 と、第 2 の特徴抽出部 230 と、位置特定部 240 と、制御部 250 と、学習部 260 とを含む。学習装置 200 は、初期領域予測部 210 を更に含んでもよい。初期領域予測部 210、第 1 の特徴抽出部 220、第 2 の特徴抽出部 230、位置特定部 240 及び制御部 250 は、検索装置 100 における初期領域予測部 110、第 1 の特徴抽出部 120、第 2 の特徴抽出部 130、位置特定部 140 及び制御部 150 とそれぞれ同じである。

[0033] 学習装置 200 は、入力として学習用メディアデータ R 及び学習用クエリデータ Q を含む学習用データ（「クエリ参照ペア」とも呼ばれる）を用いる。学習用クエリデータ Q は学習用メディアデータ R の一部でもよく、学習用メディアデータ R の一部に類似したデータでもよい。学習用メディアデータ R 中の学習用クエリデータ Q の正確な位置は必ずしも与えられる必要はない。

[0034] 入力として学習用メディアデータ R 及び学習用クエリデータ Q を用いることにより、特徴ベクトル $f(Q)$ が第 1 の特徴抽出部 220 により取得でき、特徴ベクトル $f(R(l_t))$ が第 2 の特徴抽出部 230 により取得でき、次の候補領

域 $R(l_{t+1})$ 又はその位置 l_{t+1} 、或いは最終結果が位置特定部 240 により取得できる。

[0035] 学習部 260 は、特徴ベクトル $f(Q)$ 、特徴ベクトル $f(R(l_t))$ 及び次の候補領域 $R(l_{t+1})$ 又はその位置 l_{t+1} 、或いは最終結果を入力とし、第 1 の特徴抽出部 220、第 2 の特徴抽出部 230 及び位置特定部 240（及び初期領域予測部 210 が含まれる場合には初期領域予測部 210）のニューラルネットワークのパラメータを出力する処理部である。学習部 260 は、位置特定部 240 により決定された候補領域 $R(l_{t+1})$ が目標領域を捉えたか否かを判定した結果に基づいて、ニューラルネットワークのパラメータを更新する。また、学習部 260 は、特徴ベクトル $f(Q)$ と特徴ベクトル $f(R(l_t))$ との類似度を計算し、特徴ベクトル $f(Q)$ と特徴ベクトル $f(R(l_t))$ との類似度に基づいて、ニューラルネットワークのパラメータを更新する。さらに、学習部 260 は、位置の系列 $\{l_{\tau}\}_{\tau=1}^{t-1}$ を用いて報酬（reward）を計算し、報酬に基づいてニューラルネットワークのパラメータを更新してもよい。

[0036] <第 2 の実施例>

第 2 の実施例では、第 1 の実施例の概念を用いた画像検索方法について説明する。第 2 の実施例では、メディアデータは参照画像 R であり、クエリデータはクエリ画像 Q である。

[0037] [検索装置]

図 3 は、第 2 の実施例に係る検索装置 100 の概念図である。

[0038] 初期領域予測部 110 において、参照画像 R は、低解像度の参照画像 R_{coarse} にダウンサンプリングされ、特徴ベクトル $f(R_{\text{coarse}})$ が CNN によって低解像度の参照画像 R_{coarse} から抽出され、特徴ベクトル $f(R_{\text{coarse}})$ が位置 l_0 に線形射影される。

[0039] 第 1 の特徴抽出部 120 において、画像特徴量を表す特徴ベクトル $f(Q)$ は CNN によってクエリ画像 Q から抽出される。第 2 の特徴抽出部 130 において、画像特徴量を表す特徴ベクトル $f(R(l_0))$ は CNN によって位置 l_0 における参照画像 R から抽出される。

- [0040] そして、位置特定部 140 において、現在位置 l_0 が特徴ベクトル $f(R_{\text{coarse}})$ 又は $f(R(l_0))$ と同じ次元を有する位置ベクトルに線形射影され、次の隠れ状態が、特徴ベクトル $f(R_{\text{coarse}})$ 及び $f(R(l_0))$ の組み合わせと、位置特定部 140 の現在の隠れ状態と、位置ベクトルとに基づいて LSTM によって決定され、次の隠れ状態が次の位置 l_1 に線形射影される。
- [0041] そして、次の位置 l_1 が与えられると、制御部 150（図 3 に図示せず）は、繰り返し回数が所定の制限値 T になるまで、第 2 の特徴抽出部 130 及び位置特定部 140 の動作を繰り返す。制限値になった場合、画像検索処理は終了し、画像検索結果が出力される。
- [0042] 以下、検索装置 100 により実行される画像検索方法の各ステップについて、図 4 を参照して詳細に説明する。
- [0043] ステップ S101～S103 は初期化ステップに関し、初期化ステップでは、参照画像 R を入力とし、初期位置 l_0 を出力する。
- [0044] ステップ S101 において、初期領域予測部 110 は、参照画像 R を低解像度の参照画像 R_{coarse} にダウンサンプリングする。参照画像 R のダウンサンプリングは、3 のスケーリング係数を用いて行われてもよい。
- [0045] ステップ 102 については、図 5 を参照して詳細に説明する。図 5 は、初期領域予測部 110 における CNN の詳細図である。
- [0046] 初期領域予測部 110 は、ダウンサンプリング後の参照画像 R_{coarse} を特徴ベクトル $f(R_{\text{coarse}})$ にマッピングする 3 つの畳み込み層を使用する。特徴ベクトル $f(R_{\text{coarse}})$ はダウンサンプリング後の参照画像 R_{coarse} から取得されるため、所与の参照画像 R の中で潜在的に関心のある領域がどこにあるかという示唆を効果的に与える。より具体的には、第 1 の畳み込み層は、 R_{coarse} を入力とし、サイズ 7×7 の 32 個の 2D 畳み込みフィルタに続いて、ストライド 2 の最大プーリング層を適用する。第 2 及び第 3 の畳み込み層のそれぞれは、同じサイズ 3×3 の 32 個の 2D 畳み込みフィルタに続いて、同じストライド 2 の最大プーリング層で構成される。最後に、第 3 の畳み込み層の出力を受け取り、長さ 256 の固定長の特徴ベクトルを生成する全結合（FC：fully-connected

）層が存在する。

[0047] ステップS 1 0 3において、ステップS 1 0 2において取得された特徴ベクトル $f(R_{\text{coarse}})$ に線形射影が適用され、長さ256の特徴ベクトル $f(R_{\text{coarse}})$ を長さ2の位置ベクトルに変換する。さらに、位置ベクトルは-1から1の範囲に正規化される。

[0048] ステップS 1 0 4～S 1 0 6は特徴抽出ステップに関する。

[0049] ステップS 1 0 4において、ステップS 1 0 3の位置ベクトルを用いて、当該位置ベクトルにより表される位置 l_0 において画像領域が参照画像Rから抽出される。

[0050] ステップS 1 0 5及びS 1 0 6については、図6を参照して詳細に説明する。図6は、第1の特徴抽出部120及び第2の特徴抽出部130におけるCNNの詳細図である。位置 l_t は制御部150の制御で繰り返し決定されるため、以下の説明において、一般的に繰り返し回数を t として示す。なお、繰り返し回数が0であるときに、初期位置 l_0 が用いられる。

[0051] 第1の特徴抽出部120及び第2の特徴抽出部130は、クエリ画像Q及び抽出された画像領域 $R(l_t)$ をそれぞれ特徴ベクトル $f(Q)$ 及び $f(R(l_t))$ にマッピングする5個の畳み込み層を使用する。より具体的には、第1の特徴抽出部120及び第2の特徴抽出部130は、一連の5つの畳み込み正規化線形ユニット（Conv-ReLU: convolutional-rectified linear unit）層（2D畳み込みの後のReLU活性化）に続いて、大域的平均プーリング（GAP: global average pooling）層を有するように設計された、同じパラメータを有するCNNである。これは全層畳み込み（fully-convolutional）であり、全結合層を有さない。この構成の利点は、任意のサイズの入力画像から同じ長さ（128）の特徴ベクトルを抽出することができることにある。各畳み込み層の仕様は、第1層：32個の2D畳み込みフィルタ（フィルタ：7×7及びストライド：1×1）、第2層：64個の2D畳み込みフィルタ（フィルタ：5×5及びストライド：1×1）、第3層：128個の2D畳み込みフィルタ（フィルタ：3×3及びストライド：1×1）、第4層：256個の2

D 畳み込みフィルタ（フィルタ：1×1及びストライド：1×1）、第5層：128個の2D畳み込みフィルタ（フィルタ：1×1及びストライド1×1）である。

[0052] ステップS107において、第1の特徴抽出部120及び第2の特徴抽出部130から取得された固定長の特徴ベクトル $f(Q)$ 及び $f(R(l_t))$ は、長さ256の単一のベクトルに結合される。

[0053] ステップS108～S110は位置特定ステップに関し、位置特定部140は、画像の特徴ベクトル $f(Q)$ 及び $f(R(l_t))$ と、現在位置 l_t と、LSTMの現在状態 h_t とを含む3つの入力に基づいて、LSTMによって次の位置 l_{t+1} を順に予測する。

[0054] 位置 l_t は特徴ベクトル（256）に比べて次元が低いため（2Dであるため）、結果のベクトルは特徴ベクトルの要素に支配され、有用な位置情報を取得することが困難になる。これを回避するため、ステップS108において、まず、現在位置 l_t は、線形射影によって特徴ベクトルと同じ次元を有する位置ベクトルに符号化され、次に、特徴ベクトルと組み合わせて処理される。

[0055] ステップS109において、 $t=0$ の場合、LSTMの初期の隠れ状態 h_0 は、ステップS101において取得されたダウダンプリング後の低解像度の参照画像の特徴ベクトル $f(R_{\text{coarse}})$ により決定される。次に、結合されたベクトルと、現在位置 l_t と、現在の隠れ状態 h_t とを含む3つの入力組み合わせられ、単一のベクトルを形成してLSTMに入力される。位置特定部140の出力は、次の隠れ状態 h_{t+1} の固定長のベクトル（256）である。

[0056] ステップS110において、LSTMの結果の次の隠れ状態 h_{t+1} は、線形射影により次の領域の予測位置の期待値

[0057] [数1]

$$\hat{l}_{t+1}$$

に変換される。次の領域の位置 l_{t+1} が実際にはガウス分布に従う確率変数であり、

[0058] [数2]

$$\hat{l}_{t+1}$$

が平均ベクトルとなることを仮定する。具体的には、 l_{t+1} は

[0059] [数3]

$$l_{t+1} \sim \mathcal{N}(\hat{l}_{t+1}; I)$$

として、分布

[0060] [数4]

$$\mathcal{N}(\hat{l}_{t+1}; \lambda I)$$

からのサンプルとして取得される。ただし、 I は単位行列であり、 λ はネットワークのハイパーパラメータである。

[0061] 次にステップS 1 1 1において、制御部1 5 0は繰り返し回数 t を1だけ増加させ、最大繰り返し回数 T になるまで、ステップS 1 0 4及びS 1 0 6～1 1 0を繰り返す。最大繰り返し回数 T になった場合、位置 l_t が出力される。簡単にするため、最大繰り返し回数 T は6に固定されてもよい。また、 T は適応的に決定されてもよい。

[0062] [学習装置]

以下、学習装置2 0 0により実行される学習方法の各ステップについて、図7を参照して詳細に説明する。

[0063] 目標領域は順に特定されるため（すなわち、現在の決定は全ての過去の決定に依存して行われるため）、各繰り返し回数における決定は独立して評価できない。したがって、サンプルが独立同分布であると仮定する一般的な教師あり学習法又は教師なし学習法は適用できない。過去の決定における依存性に対処するために、決定過程は部分可観測マルコフ決定過程（POMDP : partially observable Markov decision process）としてモデル化され、学習は強化学習方式で行うことができる。具体的には、ニューラルネットワークを学習するために、方策勾配法が用いられる。

[0064] この問題では、方策（policy）は、マッチングさせる次の位置をどのように選択するかを決定することである。この問題を解決するために、次の位置が平均値

[0065] [数5]

$$\hat{l}_{t+1}$$

及び標準偏差 λ のガウス分布に従う確率変数であると仮定する。また、平均値

[0066] [数6]

$$\hat{l}_{t+1}$$

のみが学習対象であり、 λ はハイパーパラメータであると仮定する。

[0067] 学習処理は、学習用データ（クエリ参照ペア）を初期領域予測部 210 に入力することから始まる。

[0068] ステップ S201～S203において、初期領域予測部 210 は、ステップ S101～S103において説明したように、初期位置 l_0 を出力する。

- [0069] ステップS204～S206において、第1の特徴抽出部220及び第2の特徴抽出部230は、ステップS204～S206において説明したように、特徴ベクトルを抽出する。
- [0070] ステップS207～S210において、ランダムな確率過程（ガウス分布）が位置特定部240の出力に適用され、結果（すなわち、次の予測位置 l_t ）を生成する。
- [0071] ステップS211において、学習部260は、予測位置 l_t の精度に基づいて報酬を計算する。さらに、ステップS212において、学習部260は、特徴ベクトル $f(Q)$ と特徴ベクトル $f(R(l_t))$ との類似度を計算する。
- [0072] ステップS213において、計算された報酬及び類似度を用いて、学習部260は、ニューラルネットワークのパラメータを更新するためにバックプロパゲーションを開始する。バックプロパゲーションはステップS210の確率過程を迂回し、位置特定部240、第2の特徴抽出部230、第1の特徴抽出部220及び初期領域予測部210のニューラルネットワークのパラメータを更新する。
- [0073] 概念的には、ニューラルネットワークのバックプロパゲーションは、より多くの報酬が将来的に与えられるよう、ニューラルネットワークのパラメータが更新され则认为られる。
- [0074] 要するに、この学習方法は、画像の特徴量と検索パス（すなわち、検索対象の位置の順序）とを統一した枠組みで一緒に学習することができる。非特許文献5、非特許文献6及び非特許文献7において提案されているように、いくつかの再帰型アテンションモデル（recurrent attention model）が存在する。これらの非特許文献とは異なり、第2の実施例の学習方法は、画像検索タスクにカスタマイズされており、類似検索のための画像特徴量を効果的に学習するように設計されている。さらに、全体のモデルは教師なしで学習できるようにされており、すなわち、上記のモデルとは異なり、学習用のクラスラベルは必要とされない。
- [0075] 以下、学習方法について詳細に更に説明する。

[0076] $\Theta = \{\theta_f, \theta_l\}$ が全体のモデルの一式のパラメータであり、 θ_f が第1の特徴抽出部220及び第2の特徴抽出部230の一式のパラメータであり、 θ_l が位置特定部240の一式のパラメータであると仮定する。或いは、 $\Theta = \{\theta_f, \theta_l, \theta_i\}$ が定義されてもよく、 θ_i は初期領域予測部210の一式のパラメータである。 Θ を調整するために、強化学習が用いられる。

[0077] 上記のように位置 l_t は順に決定されるため、 l_t は全ての過去の位置を条件として決定される。表記を簡潔にするために、 $s_{t-1} = \{\{l_\tau\}_{\tau=1}^{t-1}, Q, R\}$ と定義する。このモデルの方策は、条件付分布 $\pi(l_t | s_{t-1}; \Theta)$ として表すことができる。学習の目的は、 Θ に関して報酬 $R = \sum_{t=1}^T r_t$ を最大化することである。ただし、 r_t は報酬関数である。一般的には、報酬関数 r_t は、繰り返し回数 t における検索の成功又は失敗に基づくものとすることができる。すなわち、繰り返し回数 t におけるウィンドウが位置 l_g を正確に捉えた場合、 $r_t=1$ であり、そうでない場合には $r_t=0$ である。ウィンドウが位置 l_g を正確に捉えるか否かは、ウィンドウと領域 $R(l_g)$ とのIoU(intersection over union)に基づいて決定されてもよい。全体の報酬の期待値は以下の式(1)として与えられる。このように、学習においては、目標位置の候補である繰り返し回数 t におけるウィンドウが目標位置である位置 l_t を正確に捉えたか否か、言い換えれば目標位置の候補である繰り返し回数 t におけるウィンドウが目標位置である位置 l_t と整合するか否かに基づいて、学習のパラメータを決定する。

[0078] [数7]

$$J(\Theta) = \mathbb{E}_{p(s_T; \Theta)}[R] \quad (1)$$

ただし、 $p(s_T; \Theta)$ は、方策 π に依存する s_T の確率分布である。上記の期待値の定義を用いることにより、位置 l_t に対する全体の期待報酬は以下の式(2)のように書き直すことができる。

[0079] [数8]

$$J(\Theta) = \sum_{l_t} \pi(l_t | s_{t-1}; \Theta) R \quad (2)$$

繰り返し回数 t における期待報酬の勾配は以下の式(3)のように定義できる

。

[0080] [数9]

$$\nabla_{\Theta} J(\Theta) = \nabla_{\Theta} \sum_{l_t} \pi(l_t | s_{t-1}; \Theta) R \quad (3)$$

式(3)における加算と勾配とを入れ替え、方策で乗算及び除算することにより、式(3)は以下の式(4)のように書き直すことができる。

[0081] [数10]

$$\nabla_{\Theta} J(\Theta) = \sum_{l_t} \pi(l_t | s_{t-1}; \Theta) \frac{\nabla_{\Theta} \pi(l_t | s_{t-1}; \Theta)}{\pi(l_t | s_{t-1}; \Theta)} R \quad (4)$$

以下の式(5)及び(6)は、

[0082]

[数11]

$$\nabla_{\theta} \log z = \frac{1}{z} \nabla_{\theta} z$$

であるという事実を単に使用することにより、式(4)から導かれる。

[0083] [数12]

$$\nabla_{\Theta} J(\Theta) = \sum_{l_t} \pi(l_t | s_{t-1}; \Theta) \nabla_{\Theta} \log \pi(l_t | s_{t-1}; \Theta) R \quad (5)$$

$$= \mathbb{E}_{p(s_T; \Theta)} [\nabla_{\Theta} \log \pi(l_t | s_{t-1}; \Theta) R] \quad (6)$$

式(6)の左辺は、将来的に報酬を高くするために、現在位置 l_t においてパラメータが報酬関数の勾配の方向に更新されるべきであることを意味しており、この勾配の方向は、報酬 R を用いて位置 l_t の尤度 $\log \pi(l_t | s_{t-1}; \Theta)$ を最大化する方向（式(6)の右辺）と同じである。

[0084] Θ に関する勾配の計算は簡単ではないが、モンテカルロ法と同様に、方策から $\{l_t; s_{t-1}\}_{t=1}^T$ の系列をサンプリングすることにより近似的に計算することができ、以下の式(7)のようになる。

[0085] [数13]

$$\nabla_{\Theta} J(\Theta) \approx \frac{1}{M} \sum_{i=1}^M \sum_{t=1}^T \nabla_{\Theta} \log \pi(l_t^i | s_{t-1}^i; \Theta) R^i \quad (7)$$

ただし、 M はサンプル系列の数である。式(7)を用いることにより、 Θ は勾配

が上昇する方向に繰り返し更新できる。

[0086] 強化学習のみでのモデルの学習は、しばしば予測結果を不安定にする。したがって、第2の実施例では、学習処理の安定性を向上させるために、他のロス関数を取り入れられる。具体的には、第1の特徴抽出部220及び第2の特徴抽出部230からそれぞれ抽出された画像の特徴ベクトル $f(Q)$ 及び $f(R(I_t))$ が、 $f(Q)$ と $f(R(I_t))$ との類似度を測定するために用いられる。一致するペアの場合、2つの特徴ベクトルの距離は小さくなり、不一致のペアの場合、2つの特徴ベクトルの距離は大きくなるはずである。

[0087] 一致するペア及び不一致のペアについての真値ラベルを取得するために、教師ありの情報を用いる必要はない。真値ラベルは、ニューラルネットワークの学習中にそれぞれのペアに対して与えられる報酬から直接的に推測できる。報酬が1である場合、ペアは一致し、そうでない場合、不一致のペアであるとして扱われる。このため、広く用いられているシムネットワーク (Siamese network) の対比ロス関数 (contrastive loss function) と同様に、第1の特徴抽出部220及び第2の特徴抽出部230から以下のロス関数(8)が取り入れられる。

[0088] [数14]

$$L(\theta_f) = \sum_{t=1}^T r_t d^2 + (1 - r_t) \max\{0, m - d\}^2 \quad (8)$$

ただし、 $d = ||f(Q) - f(R(I_t))||$ であり、 m はマージンである。これは θ_f に関して区分的微分可能であるため、勾配降下によって最適化できる。

[0089] <評価結果>

本発明の実施例に記載の画像検索方法を評価するために、MNIST (<http://yann.lecun.com/exdb/mnist/>) 及びFlickrLogos-32 (<http://www.multimedia-computing.de/flickrlogos/>) という2つのベンチマーク用データセットを用

いた。

[0090] [データセット]

MNISTに関して、3つのデータセットをMNISTデータセットに基づいて生成した。第1のデータセットは「Translated MNIST」と呼ばれ、各参照画像は、 28×28 の数字の画像（ 28×28 画素の画像）を 100×100 のブランク画像のランダムな位置に配置することにより生成された。より具体的には、位置座標は、0からブランク画像と数字の画像とのサイズ差までの範囲内の乱数である。この範囲は、ブランク画像の境界位置に数字の画像を配置することを回避するために設定されたものである。

[0091] 第2のデータセットは「Cluttered MNIST」と呼ばれ、複雑背景に関して画像検索方法のロバスト性を評価するために用いられるものである。ここで、各参照画像は、他のランダムな数字の画像からのランダムな 9×9 の部分画像（sub-patch）を、Translated MNIST参照画像のランダムな位置に追加することにより生成された。具体的には、まず、 28×28 の数字の画像がランダムに選択され、次に、他の数字の画像の 9×9 の画素の部分画像がランダムな位置で切り取られ、最後に、 9×9 の画素の部分画像が 100×100 のTranslated MNIST参照画像のランダムに選択された位置に埋め込まれた。部分画像は、既存の数字の画像に重ならないように埋め込まれた。クラッターは、挿入される部分画像の総数を固定することで制御された。

[0092] 第3のデータセットは「Mixed MNIST」と呼ばれる。目標の数字の画像とは異なるランダムに選択された 28×28 の数字の画像が、各Cluttered MNIST参照画像のランダムな位置に配置された。Translated MNISTと同様に、位置座標は、境界位置を回避し、且つ、既存の数字の画像と重ならないように選択された。

[0093] それぞれの 100×100 の画像を参照画像とし、0~9の10個の数字が中心にあるクリーンな画像のマスターセットから、目標の数字の画像と同じ数字を有する 28×28 のクエリ画像を選択することにより、上記の3種類のMNISTに関するデータセットについて全てのクエリ参照ペアが用意された。MNISTの標準

的な意図に従って、テスト用に10,000個のクエリ参照ペアが用いられ、学習用に60,000個のクエリ参照ペアが用いられた。

[0094] FlickrLogos-32に関して、このデータセットは、元々は8,240個の画像で構成されており、学習用セットの320個のロゴ画像と、検証セットの3,260個の画像（3,000個の非ロゴ画像）と、テストセットの3,960個の画像（3,000個の非ロゴ画像）と分けられる。ここで、ロゴ毎に70個の画像を有する32個の異なるロゴのロゴ画像（2,240個）のみを検討した。学習用セット及びテストセットは、それぞれ2,000個及び240個のクエリ参照ペアで構成された。

[0095] 各ペアは、データセット内のロゴ画像を参照画像として考え、クエリ画像と同じブランド名のちょうど切り取られたロゴをクエリ画像として考えることにより生成された。個々のロゴにそれぞれ対応して、合計で32個のクエリ画像を生成した。全ての参照画像は、元のサイズの半分にリサイズされ、各クエリ画像は、参照画像のロゴと同じサイズにリサイズされた。

[0096] [評価の設定]

性能指標に関して、画像検索方法を精度及び速度の観点で評価した。クエリ画像及び参照画像を与えて、参照画像の中の領域に対応する予測ウィンドウを出力し、精度を評価するためにこの予測ウィンドウを用いた。特に、画像検索は、物体検出手法と同じ基準に従って、予測ウィンドウと真値ウィンドウとのIoU（intersection over union）が0.5より大きい場合に成功であると考えられるものとした。成功率は、全てのペアに対する正確に一致した画像ペアの数の比である。

[0097] 効率面は2つの指標の観点で評価された。一方は評価されたウィンドウの数であり、他方は各クエリ参照ペアを処理するために必要な実行時間である。具体的には、実行時間は、テストセットのクエリ参照ペアをマッチングするために必要な合計時間の平均を取ることで決定した。

[0098] 実施例に係る画像検索方法は、BBS（非特許文献2）及びMTM（非特許文献8）の2つの既存の画像検索方法と比較して評価された。各方法について同じハードウェア環境が用いられ、これらの方法のハイパーパラメータは注意

深く調整された。

[0099] 実施例に係るモデルは、MNISTでは64のバッチサイズ及びFlickrLogos-32では1のバッチサイズを有するAdamを用いて始めから学習された。学習レートは、指数関数的減衰によって範囲 $[10_{-4}, 10_{-3}]$ に保持された。学習用セットにおいて、MNISTデータセットの結果は3エポック数だけ繰り返され、FlickrLogos-32の結果は4 5エポック数だけ繰り返された。次の位置をサンプリングするために用いられるガウス分布のハイパーパラメータ λ は0.22に固定した。

[0100] 対比ロス関数のハイパーパラメータのマージンは0.2に固定した。マージンを高く設定しすぎると、ネットワークは「不一致」についてのみ考慮するようになり、マージンを低く設定しすぎると、ネットワークは「不一致」について何も学習しなくなるというように、マージンの選択においてトレードオフが存在する。この評価において、適切なマージンは手動で決定した。

[0101] 強化学習により記述される確率的なネットワークの勾配の更新（式(7)）を実現するために、勾配をニューラルネットワークに直接渡すための勾配停止関数を用いることにより、ニューラルネットワークを近似的にのみ学習させるためのロス関数が用いられた。具体的には、ニューラルネットワークのパラメータは、ニューラルネットワークから所望の平均値が出力されるように更新された。

[0102] [結果]

全てのデータセットについて、成功率、評価されたウィンドウの数及び実行時間を、それぞれ図8～図10に示す。図8～10において、「Joint-Training」は本発明の実施例の結果である。本発明の実施例に従ってクエリ画像を特定するための検索パスは図11に示されている。

[0103] Translated MNISTについての結果に関して、図8に示すように、実施例の方法の成功率は、全ての画像検索方法の中で最高である。実施例の最大利得はBBSに対して0.25にもなり、MTMに対して0.27にもなる。この結果は、実施例がTranslated MNISTデータセットにおいて検索パスを非常に正確に学習できることを明確に示している。次に、図9に示すように、実施例は、クエリ

画像を特定するために評価された候補ウィンドウの総数の観点でBBS及びMTMより明らかに優れている。実施例の方法は6個の候補ウィンドウのみを評価しているが、他の方法は数千以上ものウィンドウを評価している。数ウィンドウのみを処理する利点は、実行時間に反映されている。各方法は画素毎に異なる計算要件を有するため、実行時間は処理されたウィンドウの数に正比例しない。それにもかかわらず、実施例の方法は、他の2つの方法に対して同等以上に高速であり、一方で、図10に示すように非常に良好な一致精度を有する。

[0104] Cluttered MNIST及びMixed MNISTに関して、図8に示すように、実施例の方法の成功率は、全ての画像検索方法の中で最高である。これは、広範囲に複雑背景のクラッターが存在する場合であっても、実施例の方法が検索パスの学習に成功することを示している。BBSの成功率は3つの方法の中で最低である。これは、BBSにおけるクエリ画像と候補ウィンドウとの一致が画素の分布の整合性に従って評価されているためであり、 (x, y, R, G, B) 空間における画素の分布が似ているときに2つのウィンドウが一致するペアであると判断されるためである。この方法は、ノイズが目標画像と同じ分布を有し得るCluttered MNIST及びMixed MNISTでは有効ではない。図9に示すように、実施例の方法は、Cluttered MNIST及びMixed MNISTの双方について、8個の候補ウィンドウだけでクエリ画像を正確に特定することができる。また、実施例の方法は、図10に示すように実行時間の観点で同等以上に優れている。

[0105] FlickrLogos-32についての結果に関して、様々な縮尺の変化、視点の変化及びゆがみ等を起因として目標画像がクエリ画像とは異なっているため、このデータセットは最も処理が困難である。図8は、実施例の方法が精度に関して全ての他の方法より優れていることを示している。実行時間に関して、図10に示すように、実施例の方法の利得はMNISTデータセットの場合に比べて大きい。これは、参照画像のサイズがMNISTのときより大きく、BBS及びMTMの実行時間は参照画像のサイズにほぼ線形性があるからである。実施例の方法の実行時間は、評価対象のウィンドウの数のみに依存し、図9に示すよう

に、2つの基準となる方法よりかなり小さい。これは、より現実的で大きいサイズの画像に適用されるときに、実施例の方法の方が効率的であることを示している。

[0106] 定性的な結果に関して、図11は、クエリ画像を特定するための検索パスを示している。実施例の方法は、検索パスを学習する優れた能力を有することが分かる。MNISTデータセットにおける結果は、クラッターのため検索レベルが困難になった場合であっても、ほぼ同じ数の評価される候補ウィンドウでクエリ画像の特定に成功することができることを示している。これは、実施例の方法が、マッチングのための検索パスと有効な特徴量とを一緒に学習するためである。

[0107] <実施例の効果>

上記のように、本発明の実施例は、参照画像の中の領域に対してクエリ画像をマッチングする問題に対処する。当該方法は、各繰り返し回数において目標領域に向かって次の位置を順に出力するニューラルネットワーク（例えば、CNN及びLSTMの組み合わせ）に基づいている。より具体的には、本発明の実施例は、参照画像のどこで次の領域を抽出するかを決定するために位置特定部を用いる。位置特定部の性能を最大化するために、本発明の実施例は、次の位置を予測するための強化学習に基づく技術を取り入れる。したがって、参照画像の関係する領域に着目することが可能になり、クエリ画像を特定するために必要なウィンドウ（候補領域）の数がかなり減少し、特に大きい参照画像の場合には、より高速な画像検索をもたらす。

[0108] クエリ画像と参照画像との類似度に基づいて、検索パスと有効な特徴量とを一緒に学習することができるため、第1に、クエリ画像を特定するために処理される候補ウィンドウの数を既存の方法よりかなり小さくすることができる。特に大きい参照画像の場合にはより高速な画像検索をもたらす。第2に、評価結果から分かるように、相当な複雑背景のクラッターを有する参照画像であっても、クエリ画像を正確に特定することができる。

[0109] <ハードウェア構成>

図１２に、本発明の実施例における各装置（検索装置１００又は学習装置２００）のハードウェア構成例を示す。各装置は、ＣＰＵ（Central Processing Unit）１５１等のプロセッサ、ＲＡＭ（Random Access Memory）やＲＯＭ（Read Only Memory）等のメモリ装置１５２、ハードディスク等の記憶装置１５３等から構成されたコンピュータでもよい。例えば、各装置の機能及び処理は、記憶装置１５３又はメモリ装置１５２に格納されているデータやプログラムをＣＰＵ１５１が実行することによって実現される。また、各装置に必要な情報は、入出力インタフェース装置１５４から入力され、各装置において求められた結果は、入出力インタフェース装置１５４から出力されてもよい。

[0110] <補足>

説明の便宜上、本発明の実施例に係る各装置（検索装置１００又は学習装置２００）は機能的なブロック図を用いて説明しているが、各装置は、ハードウェア、ソフトウェア又はそれらの組み合わせで実現されてもよい。例えば、本発明の実施例は、コンピュータに対して本発明の実施例に係る各装置の機能を実現させるプログラム、コンピュータに対して本発明の実施例に係る方法の各手順を実行させるプログラム等により、実現されてもよい。また、各機能部が必要に応じて組み合わせて使用されてもよい。

[0111] 以上、メディアデータにおける評価対象の候補領域の数を減少させ、メディアデータからクエリデータに一致する目標領域をより効率的に見つけることができるための手法について説明したが、本発明は、上記の実施例に限定されることなく、請求の範囲内において、種々の変更・応用が可能である。

[0112] 本国際出願は２０１９年３月２６日に出願した日本国特許出願２０１９－０５９４３７号に基づく優先権を主張するものであり、２０１９－０５９４３７号の全内容を本国際出願に援用する。

符号の説明

[0113] １００ 検索装置

- 1 1 0 初期領域予測部
- 1 2 0 第 1 の特徴抽出部
- 1 3 0 第 2 の特徴抽出部
- 1 4 0 位置特定部
- 1 5 0 制御部
- 2 0 0 学習装置
- 2 1 0 初期領域予測部
- 2 2 0 第 1 の特徴抽出部
- 2 3 0 第 2 の特徴抽出部
- 2 4 0 位置特定部
- 2 5 0 制御部
- 2 6 0 学習部

請求の範囲

- [請求項1] クエリデータに一致する目標領域を求めてメディアデータを検索する検索装置であって、
- 第1の学習済みニューラルネットワークを用いて、前記クエリデータから第1の特徴ベクトルを抽出する第1の特徴抽出部と、
- 前記メディアデータから第1の領域を取得し、第2の学習済みニューラルネットワークを用いて、前記第1の領域から第2の特徴ベクトルを抽出する第2の特徴抽出部と、
- 第3の学習済みニューラルネットワークを用いて、前記第1の特徴ベクトルと、前記第2の特徴ベクトルと、前記第1の領域又は前記第1の領域の位置とに基づいて、前記目標領域の候補を決定する位置特定部と、
- 前記決定された目標領域の候補を、前記第2の特徴抽出部により用いられる前記第1の領域として使用することにより、所定の条件が満たされるまで、前記第2の特徴抽出部及び前記位置特定部の動作を繰り返す制御部と、
- を有する検索装置。
- [請求項2] 学習用クエリデータが前記第1の特徴抽出部に入力され、学習用メディアデータが前記第2の特徴抽出部に入力された場合、前記第1の学習済みニューラルネットワーク、前記第2の学習済みニューラルネットワーク及び前記第3の学習済みニューラルネットワークは、前記位置特定部により決定された前記目標領域の候補が、前記目標領域を捉えたか否かを判定した結果に基づいて、学習用クエリデータ及び学習用メディアデータを用いて学習されている、請求項1に記載の検索装置。
- [請求項3] 前記第1の学習済みニューラルネットワークのパラメータは、前記第2の学習済みニューラルネットワークのパラメータと同じである、請求項1又は2に記載の検索装置。

[請求項4] 前記メディアデータをダウンサンプリング後のメディアデータにダウンサンプリングし、第4の学習済みニューラルネットワークを用いて、前記ダウンサンプリング後のメディアデータに基づいて前記第1の領域を取得する初期領域予測部を更に有する、請求項1乃至3のうちいずれか1項に記載の検索装置。

[請求項5] クエリデータに一致する目標領域を求めてメディアデータを検索するために用いられるニューラルネットワークを学習する学習装置であって、

第1のニューラルネットワークを用いて、学習用クエリデータから第1の特徴ベクトルを抽出する第1の特徴抽出部と、

学習用メディアデータから第1の領域を取得し、第2のニューラルネットワークを用いて、前記第1の領域から第2の特徴ベクトルを抽出する第2の特徴抽出部と、

第3のニューラルネットワークを用いて、前記第1の特徴ベクトルと、前記第2の特徴ベクトルと、前記第1の領域又は前記第1の領域の位置とに基づいて、前記目標領域の候補を決定する位置特定部と、

前記決定された目標領域の候補を、前記第2の特徴抽出部により用いられる前記第1の領域として使用することにより、所定の条件が満たされるまで、前記第2の特徴抽出部及び前記位置特定部の動作を繰り返す制御部と、

前記位置特定部により決定された前記目標領域の候補が、前記目標領域を捉えたか否かを判定し、当該判定結果に基づいて、前記第1のニューラルネットワーク、前記第2のニューラルネットワーク及び前記第3のニューラルネットワークのパラメータを更新する学習部と、

を有する学習装置。

[請求項6] 前記学習用メディアデータをダウンサンプリング後の学習用メディアデータにダウンサンプリングし、第4のニューラルネットワークを用いて、前記ダウンサンプリング後の学習用メディアデータに基づい

て前記第1の領域を取得する初期領域予測部を更に有し、

前記第4のニューラルネットワークのパラメータは、前記位置特定部により決定された前記目標領域の候補が、前記目標領域を捉えたか否かを判定し、当該判定結果に基づいて、前記第1のニューラルネットワーク、前記第2のニューラルネットワーク及び前記第3のニューラルネットワークのパラメータと共に更新される、請求項5に記載の学習装置。

[請求項7]

クエリデータに一致する目標領域を求めてメディアデータを検索する検索装置により使用される検索方法であって、

第1の学習済みニューラルネットワークを用いて、前記クエリデータから第1の特徴ベクトルを抽出する第1のステップと、

前記メディアデータから第1の領域を取得し、第2の学習済みニューラルネットワークを用いて、前記第1の領域から第2の特徴ベクトルを抽出する第2のステップと、

第3の学習済みニューラルネットワークを用いて、前記第1の特徴ベクトルと、前記第2の特徴ベクトルと、前記第1の領域又は前記第1の領域の位置とに基づいて、前記目標領域の候補を決定する第3のステップと、

前記決定された目標領域の候補を、前記第2のステップにより用いられる前記第1の領域として使用することにより、所定の条件が満たされるまで、前記第2のステップ及び前記第3のステップを繰り返す第4のステップと、

を有する検索方法。

[請求項8]

クエリデータに一致する目標領域を求めてメディアデータを検索するために用いられるニューラルネットワークを学習する学習装置により使用される学習方法であって、

第1のニューラルネットワークを用いて、学習用クエリデータから第1の特徴ベクトルを抽出する第1のステップと、

学習用メディアデータから第1の領域を取得し、第2のニューラルネットワークを用いて、前記第1の領域から第2の特徴ベクトルを抽出する第2のステップと、

第3のニューラルネットワークを用いて、前記第1の特徴ベクトルと、前記第2の特徴ベクトルと、前記第1の領域又は前記第1の領域の位置とに基づいて、前記目標領域の候補を決定する第3のステップと、

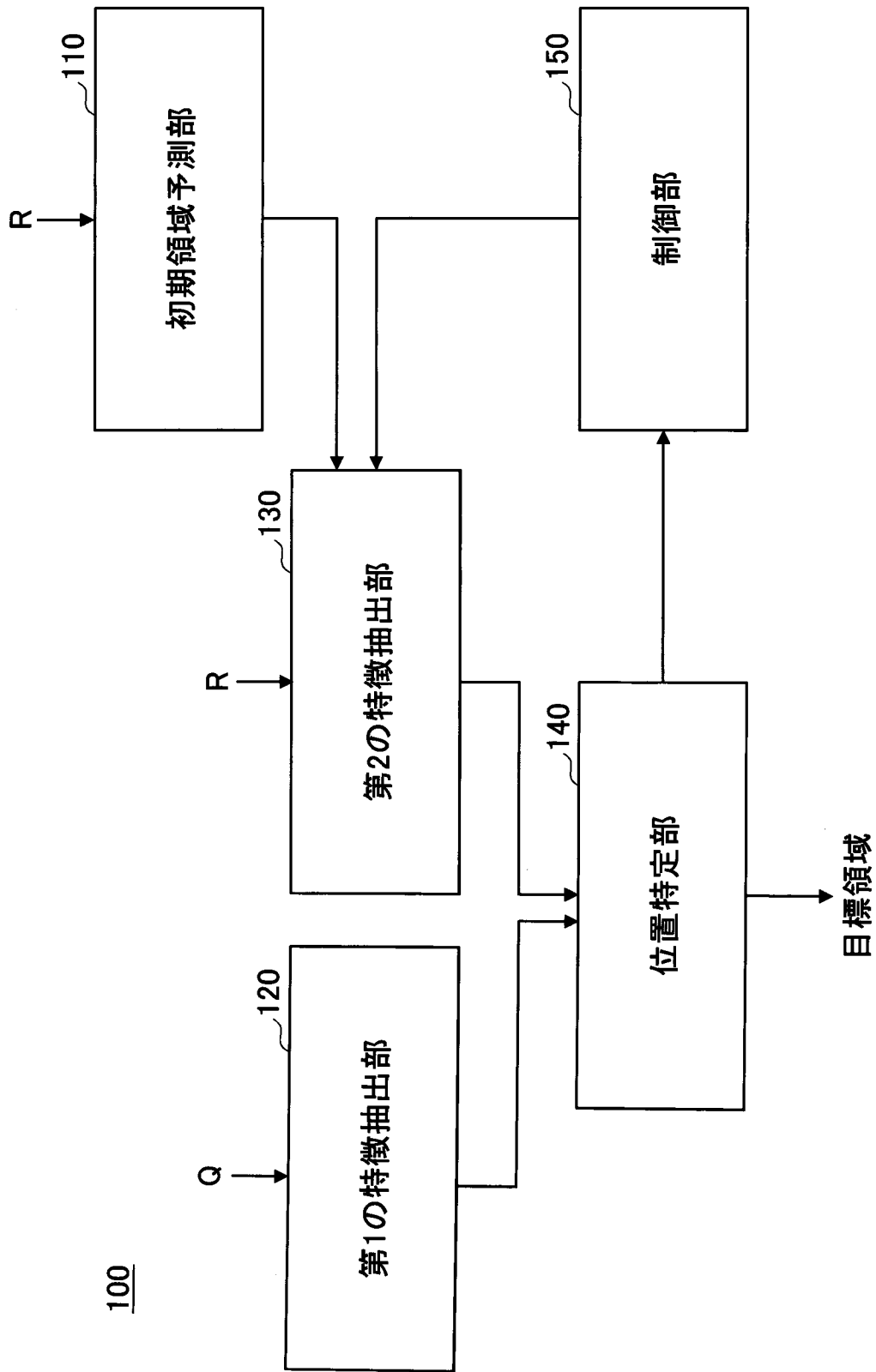
前記決定された目標領域の候補を、前記第2のステップにより用いられる前記第1の領域として使用することにより、所定の条件が満たされるまで、前記第2のステップ及び前記第3のステップを繰り返す第4のステップと、

前記目標領域の候補が、前記目標領域を捉えたか否かを判定し、当該判定結果に基づいて、前記第1のニューラルネットワーク、前記第2のニューラルネットワーク及び前記第3のニューラルネットワークのパラメータを更新する第5のステップと、

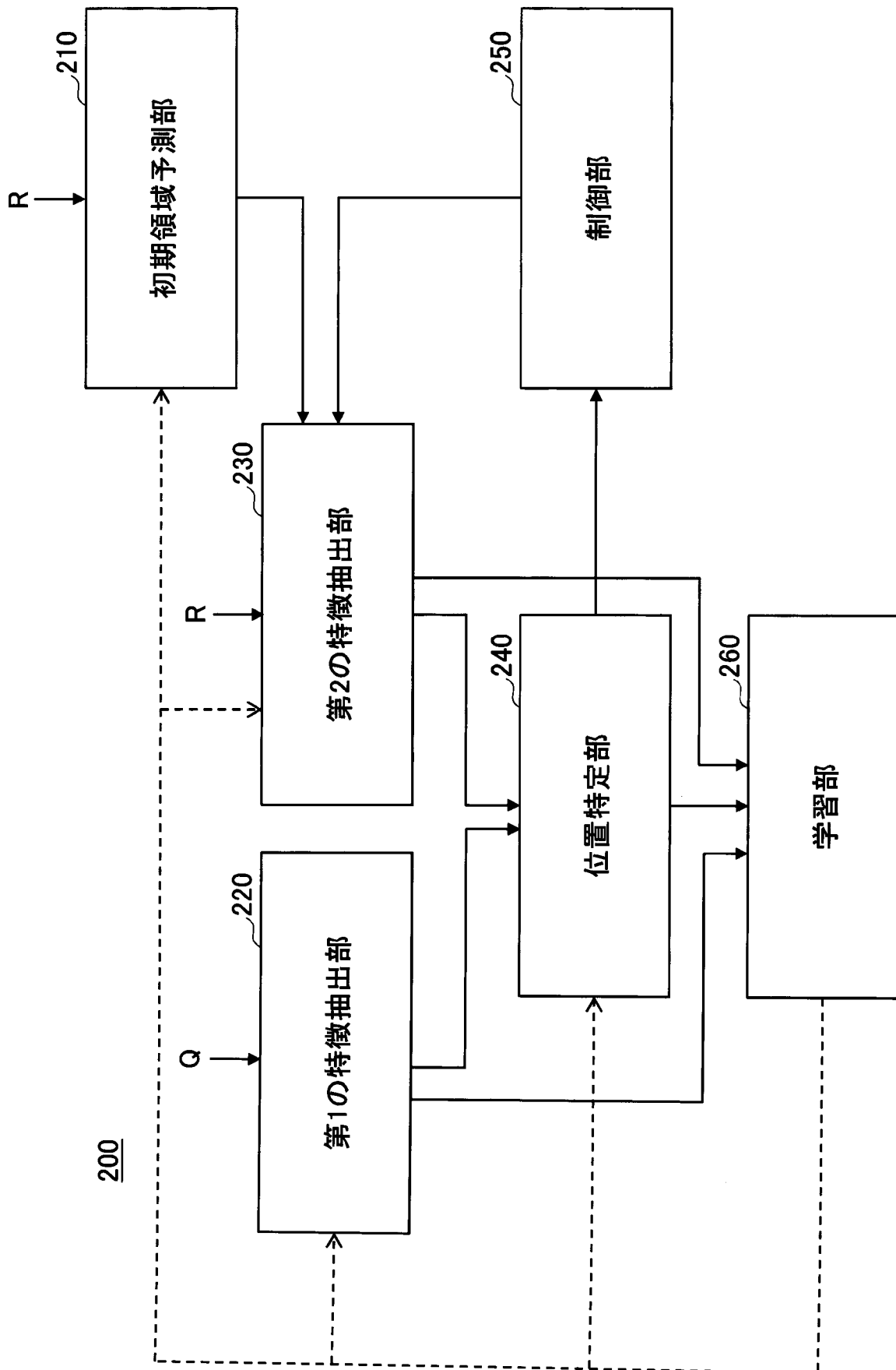
を有する学習方法。

[請求項9] 請求項1乃至6のうちいずれか1項に記載の装置としてコンピュータを機能させるプログラム。

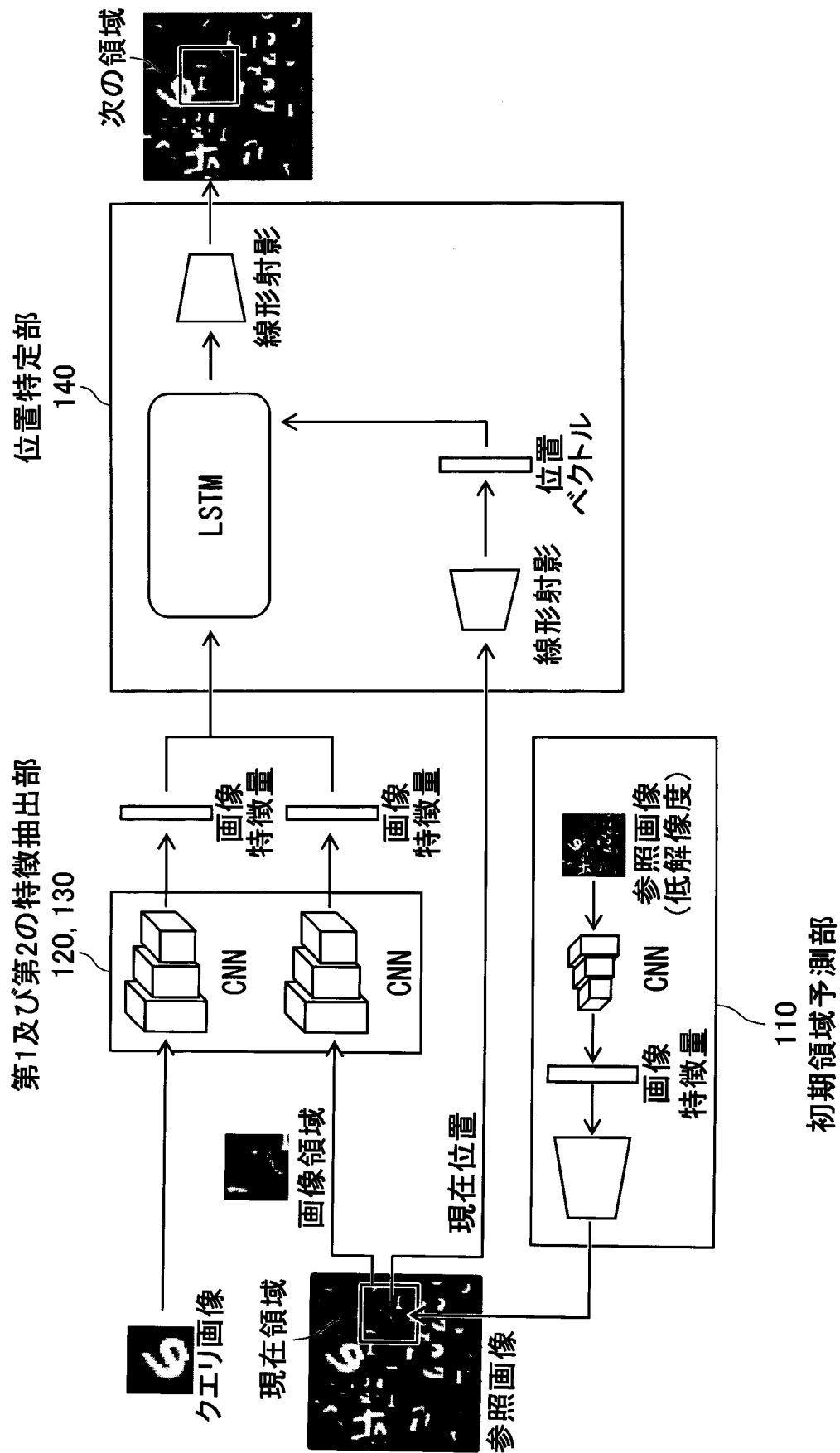
[図1]



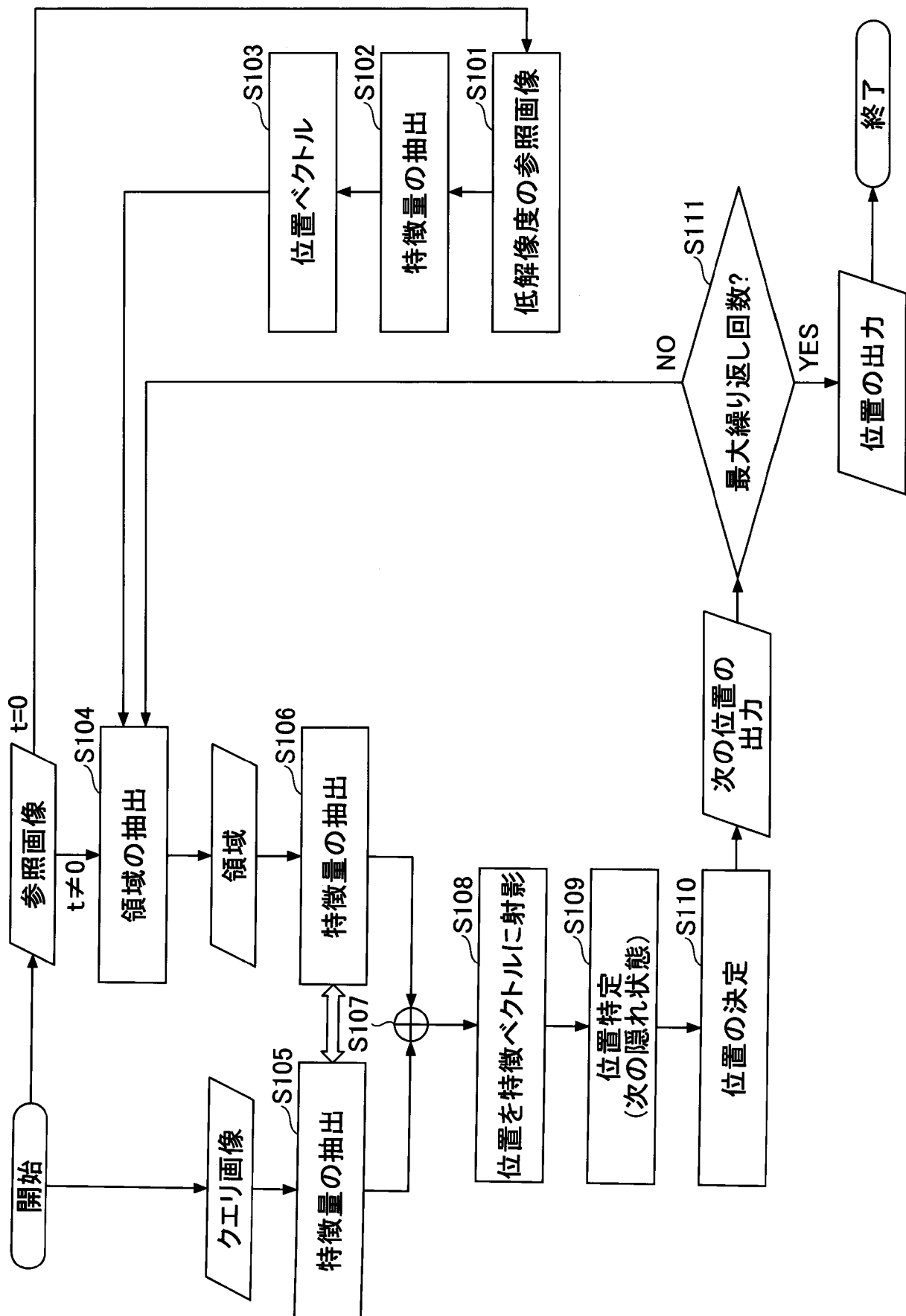
[図2]



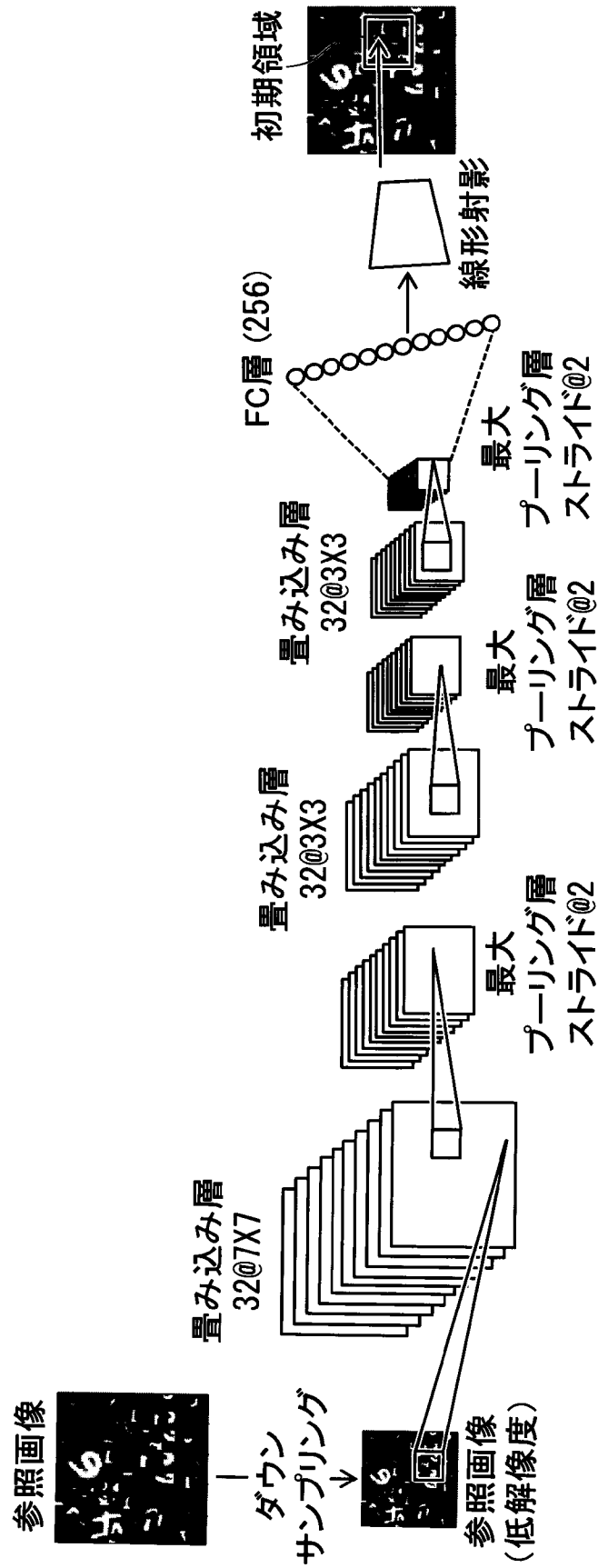
[図3]



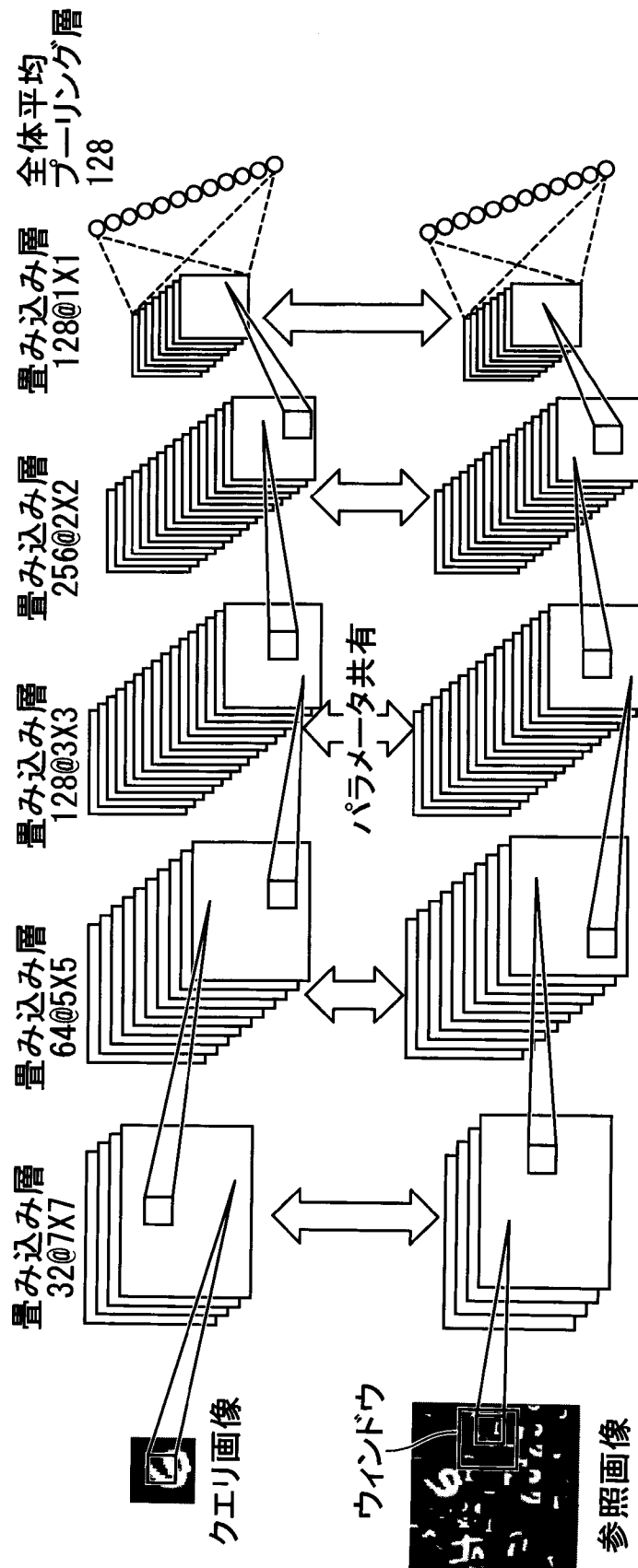
[図4]



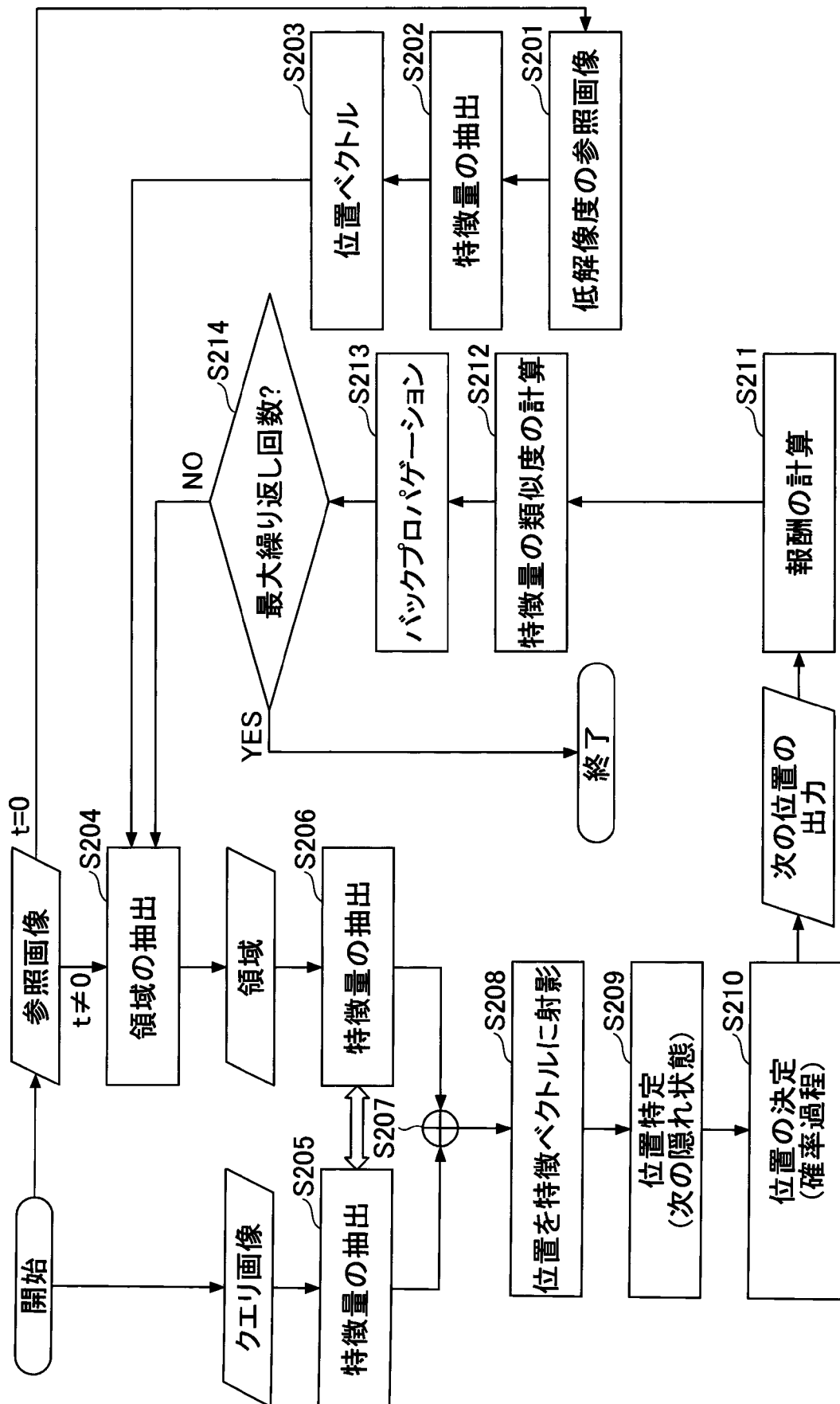
[図5]



[図6]



[図7]



[図8]

データセット	TRANSLATED MNIST	CLUTTERED MNIST	MIXED MNIST	FLICKR LOGOS-32
JOINT-TRAINING	0.95	0.91	0.88	0.39
MTM	0.68	0.20	0.15	0.28
BBS	0.70	0.11	0.08	0.36

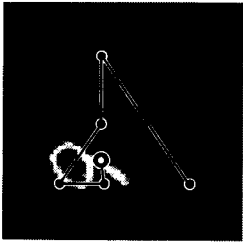
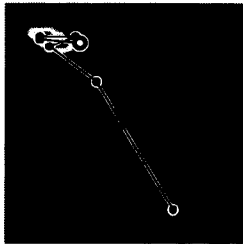
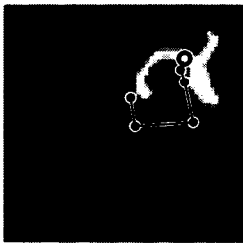
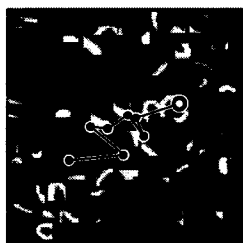
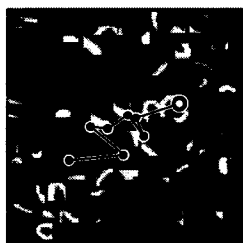
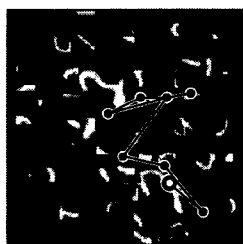
[図9]

データセット	TRANSLATED MNIST	CLUTTERED MNIST	MIXED MNIST	FLICKR LOGOS-32
JOINT-TRAINING	6	8	8	6
MTM	5329	5329	5329	18298
BBS	10000	10000	10000	40112

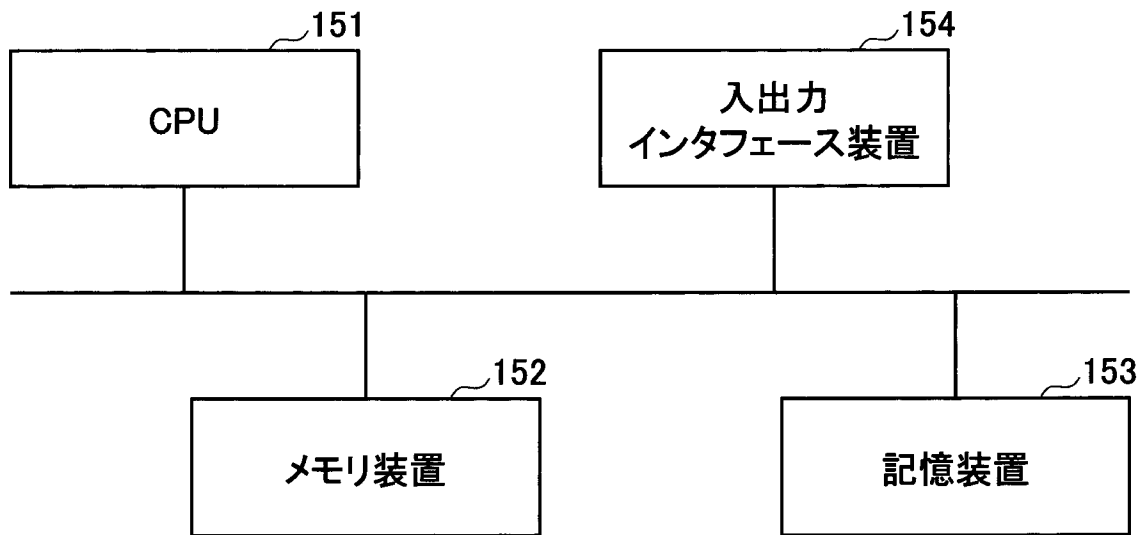
[図10]

データセット	TRANSLATED MNIST	CLUTTERED MNIST	MIXED MNIST	FLICKR LOGOS-32
JOINT-TRAINING	1	3	4	6
MTM	2	3	5	230
BBS	132	141	148	2390

[図11]

	TRANSLATED MNIST				CLUTTERED MNIST 及 MIXED MNIST			
クエリ画像	9	0	2		3	2		
検索パス								
一致領域	9	0	2		3	2		

[図12]



INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/035526

A. CLASSIFICATION OF SUBJECT MATTER

Int. Cl. G06F16/43 (2019.01) 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)

Int. Cl. G06F16/43

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

Published examined utility model applications of Japan 1922-1996

Published unexamined utility model applications of Japan 1971-2019

Registered utility model specifications of Japan 1996-2019

Published registered utility model applications of Japan 1994-2019

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 2006-221525 A (CHUDEN ENGINEERING CONSULTANTS) 24 August 2006, entire text, all drawings (Family: none)	1-9
A	JP 2009-251667 A (TOYOTA MOTOR CORP.) 29 October 2009, entire text, all drawings & US 2009/0245657 A1, entire text, all drawings	1-9
A	JP 2006-338313 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 14 December 2006, entire text, all drawings (Family: none)	1-9
A	JP 2019-28700 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 21 February 2019, entire text, all drawings (Family: none)	1-9



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
14.11.2019

Date of mailing of the international search report
26.11.2019

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

Authorized officer

Telephone No.

INTERNATIONAL SEARCH REPORT

International application No.

PCT/JP2019/035526

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	JP 2018-22390 A (NIPPON TELEGRAPH AND TELEPHONE CORP.) 08 February 2018, entire text, all drawings (Family: none)	1-9

A. 発明の属する分野の分類（国際特許分類（IPC））

Int.Cl. G06F16/43(2019.01)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. G06F16/43

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	1971-2019年
日本国実用新案登録公報	1996-2019年
日本国登録実用新案公報	1994-2019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 2006-221525 A（中電技術コンサルタント株式会社）2006.08.24, 全文, 全図（ファミリーなし）	1-9
A	JP 2009-251667 A（トヨタ自動車株式会社）2009.10.29, 全文, 全図 & US 2009/0245657 A1, 全文, 全図	1-9
A	JP 2006-338313 A（日本電信電話株式会社）2006.12.14, 全文, 全図（ファミリーなし）	1-9

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

14.11.2019

国際調査報告の発送日

26.11.2019

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）

吉田 誠

5 N

3 6 5 9

電話番号 03-3581-1101 内線 3586

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 2019-28700 A (日本電信電話株式会社) 2019.02.21, 全文, 全図 (ファミリーなし)	1 - 9
A	JP 2018-22390 A (日本電信電話株式会社) 2018.02.08, 全文, 全図 (ファミリーなし)	1 - 9