様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　３ 年　 ６ 月 １６ 日現在

機関番号：１７１０２
研究種目：挑戦的研究（萌芽）
研究期間：2017～2020
課題番号：１７Ｋ２０００８

研究課題名（和文）異なる大規模データベースで学習された深層畳込ネットを融合した特徴抽出

研究課題名（英文）Extracting Fusion Features of Deep Convolutional Neural Networks Trained on Different Large-scale Datasets

研究代表者
松川　徹（Tetsu, Matsukawa）

九州大学・システム情報科学研究院・助教

研究者番号：８０７４７２１２

交付決定額（研究期間全体）：（直接経費）　　4,800,000円

研究成果の概要（和文）：複数の学習済みCNNから得られる特徴を融合し，異なる画像認識タスクに転用する手法を開発した．まず，全結合層の重みの正規直交化により，CNNを学習したデータベースとは異なるタスクに特徴を転用した場合に性能が向上することを示した．この正規直交化した全結合層を画素へ局在化した特徴マップを2つのCNNから取得し，双線形プーリングにより融合を行った．この融合特徴は，最上位畳込層の特徴チャネルを融合するよりも少ない計算量で得られるが，それと同等以上の認識性能を示した．また，対象タスクの学習データを用いて，畳み込み層の特徴を判別的に集積する方法を開発した．

研究成果の学術的意義や社会的意義
クラス代表ベクトルの正規直交化は，対象タスクの学習データを利用せず学習済みのモデルのみから実行可能であるが，全結合層の特徴をそのまま利用した場合よりも対象タスクでの認識性能が高く，最上位畳み込み層よりも出力される次元数が少ない特徴抽出を可能とする．これを特徴融合に利用することで，計算量が削減された高性能な融合特徴を抽出できる.また，判別的特徴集積は，小サンプルデータにおいて高速に学習でき，CNNをファインチューニングするよりも有効であった．このように本研究成果は，学習済みの巨大な深層学習のモデルを，限られた計算資源で小サンプルデータ問題に適用する場合に有用である．

研究成果の概要（英文）：We have developed feature extraction methods from multiple pre-trained CNNs for transferring the learned representation to various image recognition tasks. First, we have shown that the orthonormalization of the weights of a fully connected layer improves the performance when transferring them to various tasks. We have extracted the localized feature maps of the orthonormalized fully connect layer from two CNNs, and fused them with bilinear pooling. The proposed fusion features achieved comparable or better performance compared with the fusion features of the top convolutional layers with a smaller computational cost. We have also developed a method, which conducts discriminative pooling of the convolutional features with training data of a target task.

研究分野：パターン認識

キーワード：畳込ネット　学習済みモデル　特徴転移　正規直交化　融合　双線形プーリング　画像識別　カメラ間人物照合

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

近年，深層畳込ニューラルネット(CNN)が画像認識分野で注目を集めている．大量の学習データにより学習された CNN は，画像中に写る物体のカテゴリ名称の認識を極めて正確に行え，画像から物体や背景などの関係性の説明文を自動生成するなど，広範囲な問題に対して適用され始めている.しかしながら，CNN が高い性能を発揮するためには，時に百万枚を超える膨大な量のラベル付き学習データを必要とし，その入手には相当に高い費用と労力を要する．

この弱点を克服するため，ImageNet などの大規模なデータベース(DB)で学習された CNN の中間層の出力を他のタスクへの特徴量として転用する CNN 特徴転移というアプローチが提案された．このアプローチでは，CNN の学習には膨大な量のラベル付データを要するが，転用先のタスクでは，CNN より抽出する特徴に比較的少量の学習サンプルを用いて学習した識別器を適用するのみで，画像認識を行う．学習タスクと転用先のタスクの性質が類似する場合には特に高い性能を発揮するが，大きく異なる場合には期待通りの性能が得られない．

一般に画像認識の認識対象は，複数の意味概念の集まりであることが多い．例えば，画像からその内容を説明する文章を自動生成するタスクであれば，複数の物体間や背景などとその関係性が認識対象である．また，動物の種別認識などの単一物体の認識においても，物体は顔や胴体などのパーツやその模様の特性などの複数の意味概念から構成される．このように多様な意味概念から構成される対象の高精度な認識には，単一の DB から学習された CNN 特徴のみでは不十分であると考えられる．

この問題に対処するため，可能な限り豊富な概念を含む巨大な学習画像 DB を用いることで汎用性を向上させることが考えられる．しかし，CNN は 1000 種類程度のカテゴリであっても 100 万枚単位のラベル付画像群と数日の学習時間が必要であり，1 つの CNN で多数の概念を学習することは，計算時間の面で現実的ではない．また，類似した認識カテゴリの識別に特化した CNN は認識精度を向上させるという報告もある．従って，認識タスク毎の複数の大規模 DB で個別に CNN を学習し，それらを対象タスクに転用することが，学習効率，精度面で望ましい．

しかしながら従来の研究では，一つの大規模 DB で学習された CNN を他の認識タスクに転用することに留り，CNN 特徴の連結や学習 DB の結合といった単純な統合方法を除き，複数の DB で学習された CNN の特徴転移はほとんど検討されていない．

２．研究の目的

本研究では，複数の大規模 DB で学習された CNN を融合し，異なるタスクの特徴に転用する手法を開発する．画像カテゴリ認識，同一人物検索などの広範なタスクへ転用可能な汎用的な融合特徴抽出手法を目指す．

３．研究の方法

本研究期間では，複数 CNN の融合に必要となる以下の 3 点を検討する．

(1) 各 CNN から取得する特徴の評価

①　有効な層の組み合わせ

CNN の中間層は，一般に低層ほど低次のパターン（エッジや模様など），上位層ほど高次のパターン（顔や目などの意味概念）を抽出する．複数の CNN から特定の中間層の特徴を抽出する場合，個々の CNN で抽出する特徴の層はそれぞれ可変となる.転用タスクの認識精度を向上させる組み合わせを調査する．

② 局在化した全結合層の有効性

CNN のモデルでは，最終の数層に全結合層を採用するものが多い．全結合層は認識対象カテゴリの意味概念の情報を反映しやすい．この全結合層の出力を画像中に局在化して対応づける技術により，局在化された認識対象カテゴリの意味概念を特徴に用いた場合の性能を評価する．

(2) 融合方式の開発

2 つの CNN から抽出する，特徴チャネルの相関関係を記述する融合方式を開発し，対象タスクでの認識性能の向上を行う．以下の A,B の 2 種類の相関関係を評価する．

A .画像中の同一位置での相関

2 つの特徴マップからこれを記述可能な双線形プーリング(BP)[1]を適用する．

B.画面全体での相関

画面全体で特徴を集積した後，BP へ入力したものを評価する．

(3) 転用先のタスクの学習データに基づく特徴抽出の最適化

転用先タスクでの学習データを用いることで，対象タスクに対してより有効な特徴抽出方法を開発する．

４．研究成果

(1) CNN から取得する特徴の評価

① 有効な層の組み合わせ

以下の A.B. のタスクにおいて有効な層の組み合わせの評価を行った.

A. 画像識別タスク

人物の活動に関する複合的なクラスを含む DB として,99 種類の文化的な行事と 1 種類の背景のクラスを含む Cultural, 61 種類の日常的な出来事のクラスを含む WIDER, 40 種類の人物行動のクラスを含む Action, 17 種類の人物や動物の視覚的関係のクラスを含む Phrase を用いた. 詳細物体の DB として, 200 種類の鳥のクラスを含む CUB を用いた. 場所を対象とした DB として, 67 種類の屋内クラスを含む Indoor を用いた. 2 つの大規模 DB ImageNet 及び Places で各々学習された CNN から特徴を抽出した. 抽出した特徴を線形 SVM により識別した. 数種類の CNN 構造を評価した内, 代表的な構造の ResNet50 の結果を示す.

B. カメラ間人物照合タスク

125-500 人以下の人物の画像から同一人物を同定する 4 つのカメラ間人物照合 DB を評価用 DB に用いた. 現在の最大規模の人物照合 DB である MSMT17 で学習した ResNet50 に基づくパーツベースモデル (PCB)を用いた. PCB から抽出した特徴量に対して, 距離計量学習手法 XQDA を適用して評価した. この距離計量学習の正当性を確認するための理論解析を行い, 国際会議 MVA2019 で成果発表を行った.

図 1 にタスク A の結果を示す. CNN の各層の特徴を大域的平均プーリング(GAP)により集積した後, ImageNet と Places DB で学習された特徴を連結している. Action DB では,最上位畳み込み層(Conv5_x)の連結が最良の結果である.鳥類の詳細カテゴリを認識する CUB DB では, ImageNet における Conv5_x と Places における低層(Conv2_x-Conv_4_x)の組み合わせが最良である. 他の 4 つの DB では, Action DB と同様の結果となった. これらの結果より, 人物の活動に関するクラスを識別するには,場所の意味情報が手がかりとなるが, 同一概念の詳細なカテゴリの認識では, 場所の高レベル情報よりも詳細な見た目の差異を分類するための低レベル特徴が有効と思われる.このようなタスク間での場所の高レベル情報の有効性の差異により, 層の組み合わせに有効性の違いが生じたと考えられる.

タスク B では詳細な見た目の差異への着目が有効と考えられ, 一つの CNN から抽出した低層と高層の特徴の組み合わせの有効性を確認した. 国際会議 ICPR2020 の成果発表では, 抽出した全ての層の連結が最も有効な層の性能を上回ることを示した.

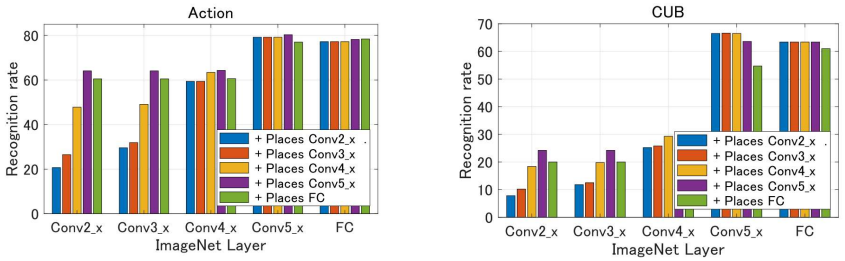

図 1 層の組み合わせの評価. 横軸に示す ImageNet DB で学習された CNN の各層の特徴に対して各棒グラフで示す Places DB で学習された各層の特徴を連結した特徴量の認識率.

A. 局在化した全結合層の有効性

全結合層により予測されるクラスの意味概念の情報を画素に局在化させるクラス活性化マップ(CAM)[2]を特徴として利用した場合の性能をタスク A で評価した. CAM の特徴チャネルの次元数は CNN の学習クラス数に対応し, 多くの CNN 構造で最上位畳み込み層の特徴チャネルの次元数よりも少ない.例えば, ResNet50 の Conv5_x 層の特徴チャネル数 2048 に対して, ImageNet のクラス数は 1000 である.

最上位全結合層の特徴は学習タスクのクラス代表ベクトルに対応するが, 異なるクラスは共通の概念を含み得るため, これらは互いに相関を持つ.この傾向により, CAM で抽出される特徴は,クラス代表ベクトルの少数の次元のみが重視して得られ, 異なるタスクで特徴抽出を行う際の汎用性が損なわれている.この問題を解決するため, 事前学習クラスの代表ベクトルを特異値分解(SVD)により正規直交化を施した特徴を開発した.画像の認識と理解技術に関する会議 MIRU2021 では, この成果を報告する.

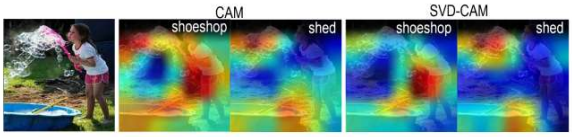

図 2 同一クラス(shoeshop/shed)に対する CAM

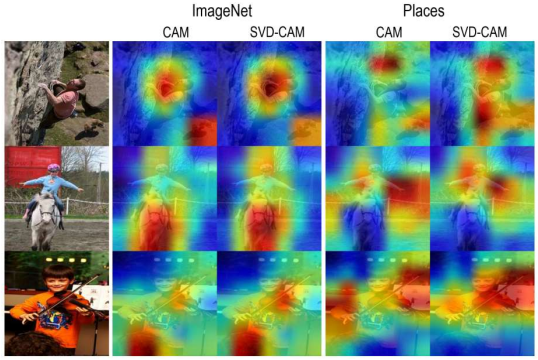

図 3 CAM の全クラスの絶対値の平均

図 2 に Places DB で学習された CAM の例を示す. この例では, SVD-CAM は多くの場所クラスで共通すると思われる背景部分での中程度の大きさを持つ値を減少させ, 人とそれが持つ物体の部位などへは大きな値を残している. 全クラスを通した傾向を解析するため, CAM の絶対値を全クラスで平均化したものを図 3 へ示す. SVD-CAM は, 画像内で高い値を持つ領域が CAM よりも広い. 特に人物内やその周辺で高い値を持つ領域が増加し, また, 背景などの影響も抑えられている. この傾向は, Places で学習されたモデルで顕著であり, 場所を認識するタスクで学習したモデルであってもクラス代表ベクトルの正規直交化が, 人物や物体をより重視して特徴を抽出する変更を行ったことが示唆される.

表 1 に各層より得られる特徴マップのチャネル情報を大域的平均プーリング(GAP)を用いて集積した特徴を用いたタスク A の認識結果を示す. GAP の線形性により, CAM を GAP で集積したものが全結合層(FC)層に対応し, 同様に SVD-CAM に GAP を適用して集積したものが, 正規直交化した全結合層(SVD-FC)へ対応する. FC 層と比較して SVD-FC 層の特徴は全ての DB で高い認識性能を示している. FC 層の重みベクトルの特異値を解析したところ, いずれのモデルも 50%の次元で 80%以上の累積寄与率に到達しており, 少ない次元に特異値が集中していた. これより正規直交化を行っていない場合, 少数の基底を重視した特徴であったため, これを改善した SVD-FC は FC よりも高い性能を示したと考えられる. 最上位畳み込み層(Conv5_x) の特徴と比較して, SVD-FC は I+P の場合, 次元数が 1/3 以下である. SVD-FC は Conv5_x に対して, やや低い性能であるが, その差は, 最大の性能を達成しているモデルに対して, 全ての DB で 0.7%以下であった.

表 1 ResNet50 の各層の特徴を GAP により集積した特徴の認識率. I, P はそれぞれ ImageNet, Places DB で事前学習されたモデルから抽出した特徴, P+I はそれらを連結したものを示す. 赤字, 青字はそれぞれ 1 番目, 2 番目に最良の層, **太字*** は各層で最良のモデルを示す.

Layer	Cultural			WIDER			Action			Phrase			CUB			Indoor		
	I	P	I+P	I	P	I+P	I	P	I+P	I	P	I+P	I	P	I+P	I	P	I+P
Conv2_x	23.0	21.4	**22.3***	22.5	21.8	**25.2***	20.6	20.0	**22.6***	18.6	18.3	**20.7***	7.7	7.2	**18.8***	30.6	28.7	**31.0***
Conv3_x	40.2	35.9	**39.6***	30.7	28.7	**42.6***	33.3	29.0	**30.7***	**33.6***	30.4	32.0	12.9	10.8	**33.2***	48.0	42.8	**46.8***
Conv4_x	66.3	62.6	**67.6***	45.0	42.4	**45.9***	65.9	53.7	**63.4***	**67.0***	56.2	64.5	**34.4***	21.1	29.3	69.7	67.3	**71.3***
Conv5_x	67.5	64.5	**71.8***	49.0	48.3	**52.3***	79.3	64.1	**80.1***	77.5	67.8	**78.8***	**66.5***	24.1	63.6	76.0	84.9	**85.4***
FC	63.9	57.2	**68.4***	46.7	44.9	**50.3***	77.5	60.7	**78.5***	76.5	63.6	**77.3***	**63.4***	19.9	60.9	73.8	**84.3***	83.5
SVD-FC	67.0	61.7	**71.1***	48.5	46.5	**51.7***	79.0	63.1	**80.0***	77.4	67.0	**78.6***	**66.1***	22.2	63.4	75.7	**85.1***	**85.1***

(2) 融合方式の開発

双線形プーリング(BP)[1]に対して行列パワー正規化(MPN)[3]を適用した MPN-BP は CNN 特徴の有効な特徴チャネル融合法であることが知られる. MPN の有効性をタスク B における手動設計特徴で評価し, IEEE Trans. on PAMI で発表した論文の Appendix に示した.

BP の出力次元数は入力特徴チャネルの次元数の 2 乗のオーダーで増加する. そこで, 部分行列平方根[4]を用い, BP の精度を保ち出力次元数の削減が行えるコンパクト型 BP[5](CBP)に対して MPN を適用可能とした MPN-CBP を特徴融合に用いる. MPN-CBP の出力次元数はユーザーが設定でき, これを全ての条件で 5000 と設定するが, 特徴抽出の計算量は, MPN-CBP に入力する特徴マップの特徴チャネル次元数に依存する.

そこで，前節の SVD-CAM を MPN-CBP の入力特徴マップとして用い，入力特徴チャネル次元数を低く抑える．各層から得られた特徴を MPN-CBP で融合した特徴を用いたタスク A の認識結果を表 2 に示す．表 1 の結果と比較することで，MPN-CBP による I+P の融合が，GAP により得られる特徴の単純連結よりも有効な特徴融合法であることが確認できる．MPN-CBP に入力する層として SVD-CAM は，Conv5_x よりも高い性能か同等である．なお Conv5_x の特徴に対して 1/3 程度の次元数である I+P の条件では，SVD-CAM を用いた MPN-CBP は，Conv5_x を用いた場合に対して 20%程度特徴抽出時間が短かった．

SVD-CAM を MPN-CBP の入力に直接用いる場合，3.(2)A.画像中の同一の位置における相関に対応し，SVD-CAM に GAP を適用した SVD-FC に対して MPN-CBP を適用する場合，3.(2)B.画面全体での相関に対応する．SVD-FC は，Cultural と WIDER で有効であった．これらは，青森ねぶた祭や交通事故などの比較的，高レベルな概念を認識するタスクであるため，画面内で存在している意味情報の共起関係が有効であるためと考えられる．その他のデータでは，SVD-CAM が有効であり，これは特に CUB で有効であった．画面内に小さく映る鳥の詳細な種類を識別するタスクである CUB は，同一の位置に着目した詳細な見た目情報が，有効であるためと考えられる．

表 2 ResNet50 の各層の特徴を MPN-CBP により集積した特徴の認識率．P+I は，P,I の各モデルから抽出した特徴チャネルを連結したマップを MPN-CBP の入力としている．

Layer	Cultural			WIDER			Action			Phrase			CUB			Indoor		
	I	P	I+P	I	P	I+P	I	P	I+P	I	P	I+P	I	P	I+P	I	P	I+P
Conv2_x	49.7	46.8	**50.0***	33.3	32.6	**34.1***	34.7	33.4	**35.4***	33.7	32.5	**34.5***	21.6	19.9	**22.1***	48.8	46.1	**49.4***
Conv3_x	61.5	58.4	**60.2***	39.4	38.6	**38.8***	48.6	44.7	**47.0***	**48.8***	45.1	48.2	**34.9***	29.2	33.4	60.8	59.3	**62.1***
Conv4_x	68.4	65.2	**67.4***	43.5	41.8	**42.4***	72.6	59.7	**68.4***	**73.8***	62.7	69.3	**61.8***	38.3	52.4	73.5	71.6	**73.8***
Conv5_x	69.9	67.6	**72.6***	48.2	48.2	**50.8***	80.5	67.2	**81.3***	78.4	70.8	**79.3***	**72.8***	33.1	70.4	75.8	**85.1***	84.5
CAM	68.7	64.1	**71.9***	48.4	47.9	**50.9***	79.7	64.7	**80.1***	**79.2***	68.6	79.1	**71.2***	29.2	69.4	74.5	**84.7***	84.0
SVD-CAM	70.2	67.3	**73.1***	48.4	48.8	**51.4***	80.4	67.0	**81.3***	78.6	71.0	**80.0***	**72.4***	32.9	71.4	75.4	85.3	**85.7***
FC	67.8	62.2	**71.8***	50.2	48.6	**52.9***	77.3	59.7	**77.6***	**79.1***	64.5	78.9	**64.5***	20.5	60.7	75.9	**84.7***	83.6
SVD-FC	69.8	66.0	**73.5***	51.0	49.8	**53.6***	79.3	63.3	**79.6***	79.2	68.5	**79.6***	**67.0***	23.7	63.9	76.8	85.4	**85.2***

(3) 転用先のタスクの学習データを利用した特徴抽出の最適化

転用先のタスクの学習データを用いて，学習済みモデルから畳み込み層の特徴を判別的に集積する方法を開発し，タスク B へ適用した．判別的な重みマップ学習法にランダム射影を導入し，学習の 30 倍以上の高速化を実現した．これに加え，距離計量学習 XQDA と複数の層の特徴の利用により，学習済みモデルを小サンプルの学習データを用いてファインチューニングを行うよりも高精度な人物照合率を達成可能な特徴量を高速に学習できた．この研究成果を国際会議 ICPR2020 で発表した．

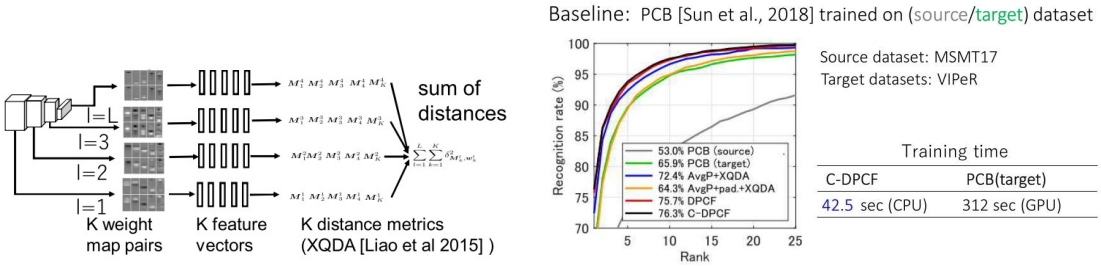

図４　左：提案モデル(C-DPCF)．　右：ファインチューニング(PCB(target))との比較．

<引用文献>

[1] T.-Y.Lin, A.RoyChowdhury, S.Maji, Bilinear convolutional neural networks for fine-grained visual recognition, IEEE Trans. on PAMI, vol.40, no.6, pp.1309-1322, 2018.

[2] B.Zhoug, A.Khosla, A.Lapedriza, A.Oliva, A.Torralba, Learning deep features for discriminative localization, In CVPR2016.

[3] P.Li, J.Xie, Q.Wang, W.Zhuo, Is second-order information helpful for large-scale visual recognition, In ICCV2017.

[4] M.Gou, F.Xiong, O.Camps, M.Szanier, MoNet: Moments embedding network, CVPR2018.

[5] Y.Gao, O.Beijbom, N.Zhang, T.Darrell, Compact bilinear pooling, In CVPR2016.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, Yoichi Sato	４．巻 42
２．論文標題 Hierarchical Gaussian Descriptors with Application to Person Re-Identification	５．発行年 2020年
３．雑誌名 IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)	６．最初と最後の頁 2179～2194
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TPAMI.2019.2914686	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Tetsu Matsukawa, Einoshin Suzuki
２．発表標題 Kernelized Cross-View Quadratic Discriminant Analysis for Person Re-Identification
３．学会等名 Sixteenth International Conference on Machine Vision Applications (MVA 2019)（国際学会）
４．発表年 2019年

１．発表者名 Tetsu Matsukawa, Einoshin Suzuki
２．発表標題 Convolutional Feature Transfer via Camera-specific Discriminative Pooling for Person Re-Identification
３．学会等名 in 25th International Conference on Pattern Recognition (ICPR2020)（国際学会）
４．発表年 2021年

１．発表者名 松川徹，鈴木英之進
２．発表標題 正規直交クラス代表ベクトルとの内積表現によるConvNet特徴転移
３．学会等名 画像の認識・理解シンポジウム (MIRU2021)
４．発表年 2021年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関