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　６ 年　 ５ 月 １７ 日現在

機関番号：１７１０２
研究種目：基盤研究(C)（一般）
研究期間：2020～2023
課題番号：２０Ｋ１１８９０

研究課題名（和文）特徴チャネルに対する不変性を持つ空間基底に基づく畳み込みネットの注意機構

研究課題名（英文）Attention Mechanism of Convolutional Neural Networks Based on Spatial Bases with Feature Channel Invariance

研究代表者
松川　徹（Matsukawa, Tetsu）

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）　　3,300,000円

研究成果の概要（和文）：人物の向きを，服装の柄，場所や天候など撮影状態の様々な変動に頑健に推定するには，局所的なパーツの空間的配置が，正確な推定を行うための重要な要素となる．そこで畳込ネットの特徴マップの冗長なチャネル活性化を要約するチャネルプーリングに注目する． 本研究では，特徴マップを特徴チャネルの断面で取り出した画像群に対して特異値分解を適用することで得られる空間基底が，特徴チャネル順序の入れ替えに対して不変であり，かつ空間的情報に対して判別的な特徴量となることに着目し，この空間基底の部分空間を特徴表現とするグラスマンチャネルプーリングを提案した．

研究成果の学術的意義や社会的意義
学術的意義は，提案手法と従来の双線形プーリングとグラスマンプーリングとの関係性を示し，撮影状況の変動を模擬するスタイル変換に対する頑健性の論拠を示した点， また，提案手法がスタイル変動への頑健性を高める効果を持つインスタンス正規化と併用した場合でも効果があることを示した点にある．
社会的には，スタイル変動へ頑健な空間概念認識は，様々な撮影変動の存在下での人物方向推定の頑健性を高めることへ寄与する．例えば，天候や照明環境，カメラの故障により，入力される画像の全体的な雰囲気がシステムを学習したデータと異なっていても安定して認識できることは，誤認識による事故を低減する．

研究成果の概要（英文）：For robustly estimating the orientations of a person in the presence of various clothing textures, as well as considering the effects of shooting conditions like location and weather, the spatial arrangement of local parts can be a crucial factor for precise estimation. Therefore, we focus on channel pooling, which summarizes redundant channel activations in the feature maps of convolutional neural networks. Focusing on the fact that the spatial bases obtained by applying singular value decomposition to a set of images extracted from a feature map in the cross-section of feature channels are invariant to the permutation of the order of feature channels and can be discriminative spatial features, we have proposed Grassmann channel pooling, which represents the subspace of the spatial bases as the feature representation.

研究分野：パターン認識

キーワード：畳込ネット　空間基底　不変特徴　頑健性　双線系/グラスマンプーリング　人物方向推定

※科研費による研究は、研究者の自覚と責任において実施するものです。そのため、研究の実施や研究成果の公表等については、国の要請等に基づくものではなく、その研究成果に関する見解や責任は、研究者個人に帰属します。

１．研究開始当初の背景

人物などの姿勢の変動の大きい対象間の照合では，対象の姿勢を服装の柄や色によらず，高精度に認識できることが望ましい．画像認識で用いられる畳み込みネットでは，画像上の各位置に対して，畳み込みフィルタの出力値を特徴チャネルとする特徴マップを構成し，その情報を空間的に集約する処理を階層的に繰り返し，位置変動に頑健な特徴表現を得ている．また，注意機構では，特徴マップ上の各位置に対する重要度を表す注意マップを用いてチャネル値を集約することで，対象の特定の部位に限定された特徴表現を得るものであるが，既存の研究の多くは特徴マップ上の局所的な性質に基づいて重要度を算出していた[1]．

本研究では，特徴マップを特徴チャネルの断面で取り出した画像群に対して主成分分析を適用して得られる空間基底に着目する．このような空間基底は，同一チャネルの異なる位置間で大域的に類似した性質を持つ位置を表す．また，1 枚の人物画像に対して得られる空間基底は，特徴チャネルの順序を入れ替えても，同一のものとなる特徴チャネルに対する不変性を持つ．これは，人物の姿勢や観測方向を服装の模様や色などによらず推定することの手がかりとなる．

２．研究の目的

本研究では「特徴マップの空間基底に基づく不変特徴量の基本的性能の解明と高度化とその効果の定量的な実証」を目的とする．従来の空間位置に対する注意機構モデルの多く[1]は，各位置の重要度を決定する際に特徴マップの局所的な性質のみに着目し，大域的空間情報を活用していない．本研究では，特徴マップの空間基底を用い，大域的に見て類似した性質を示す位置の空間情報を抽出する不変特徴量の基本的な性質を解明する．

３．研究の方法

人物の観測方向推定における畳み込みネットの特徴マップの空間基底に基づく不変特徴表現の性能を明らかとして，その高度化を行う．性能評価には，人物の CG シミュレーションにより得られた，密な観測方向ラベルを含む PersonX データベース[2]における観測方向推定で行う．

４．研究成果

特徴マップの空間基底に基づく不変特徴表現としてグラスマンチャネルプーリング(GCP)を開発した．その効果を PersonX データベース[2]のテストデータにスタイル変動を加えて評価した．スタイル変動への頑健性の評価は当初計画していなかったが，これは天候や照明環境，カメラの故障などにより，撮影環境の変動へのシステムの頑健性へ寄与することより導入した．

4.1 Grassmann Channel Pooling(GCP)

畳み込みネットのプーリングを適用する特徴マップの幅，高さ，特徴チャネル数それぞれを H,W,C とする．特徴マップを形状変更して S (= HW) × C の行列 Xとする．これを $X \approx USV^T$ と，特異値分解する．ここで，左特異値ベクトルUは空間基底，Sは特異値を含む対角行列，Vは特徴基底を表す．この空間基底により張られる部分空間は，基底を特異値の昇順などで並べたときの順序によらないため，特異値の変動へ頑健であると期待される．そこで，GCP は空間基底の張る線形部分空間で画像を表現する．線形部分空間をユークリッド空間上のデータとして処理するため，グラスマン多様体上の射影計量[3]を用いると，各サンプルは$\varphi(U) = UU^T$と表される．これをベクトル化したものをプーリングの出力とする．

GCP と素朴なチャネルプーリング(Avg, Max, Std)との比較を図１に示す．素朴なチャネルプーリングは，局所的なフィルタ出力値をチャネル方向へ集約するため，局所的な情報を抽出している． GCP は特徴マップの大域的な情報を解析して，複数の観点から空間情報を得ている．

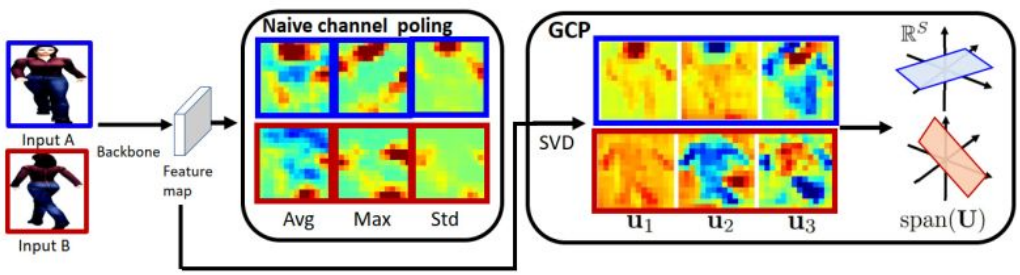


図１　GCP の概要

4.2 従来の双線系・グラスマンプーリングとの関係

従来のグラスマンプーリング[5]は，双線系プーリング[4]の拡張として提案されており，それらは特徴マップのチャネル値を空間方向へ集約する空間プーリングとして用いられていた．それぞれ以下の行列をベクトル化した特徴表現を出力する．

Bilinear Spatial Pooling (BSP)[4]: $G = X^T X = V^T S^2 V$
Grassmann Spatial Pooling (GSP)[5]: $G' = V^T V$

BSP の用いる行列Gは，スタイル変換の研究ではグラム行列と呼ばれ，画像のテクスチャやスタイルをよく表現することが知られる[6]．実際,スタイル変換は画像から抽出する行列Gを変換したいスタイルの画像のGへ近づけることで行われる[6]．よって BSP は，画像のスタイルの影響を受けやすいと考えられる．GSP は，Gから特異値Sを除外している．これによりスタイル変動の影響を抑えることが期待される．これに対して GCP はグラム行列の部分行列SとVの両方を含んでいないため，よりスタイル変動へ頑健であることが期待できる．

4.3 実験結果

4.3.1 実装詳細

ネットワーク構造として姿勢推定へ効果的な構造として知られる High-Resolution Net(HR-Net)[7]を用いる．HR ネットの Stage1-4 の特徴マップをダウンサンプリングとアップサンプリングを行い，空間のサイズを合わせ，それらのチャネルを 1x1 畳込みで混合した (H,W,C) = (14,14,80) の特徴マップへ GCP を適用した．また，スタイル変動への頑健性を増すことが知られる Instance Normalization (IN)[8]をこの特徴マップを取得する際に入力層(Input-IN),低位層(Bottom-IN),高位層(Top-IN)の 3 つへ施した．GCP の出力層に多層パーセプトロン(MLP)を適用し，Biternion 表現に基づく von Mises 損失関数[9]を用いて観測方向を回帰させた．Optimizer として Adadelta[10]を用い，100 エポック学習させた．

4.3.2 実験条件

PersonX データセット[2]は，1266 人に対して 10 度毎の角度(36 方向)の画像を 6 つのカメラで撮影した，1266 (人)×36(方向)×6 (カメラ) = 273,456 枚の画像で構成され，410 人/856 人が学習/評価データとして分割されている．学習データの内，150 人(32,400 枚)を学習データ，150 人 (32,400 枚)を検証データとしてランダムに分割して用いる．テストデータの内，各人物からランダムに 1 枚のカメラを選択した画像(30,816 枚) を評価に用いる．

評価指標として，推定値の誤差が 10 度以下となる確率 ACC_{10°と予測の平均絶対誤差 (MAE)を用いる．スタイル変動への頑健性の評価のため，テスト画像をスタイル変換アルゴリズム[11]により 5 つのスタイルへ変換した Styled データを作成した．Styled データと元のテストデータ(ORG)で予測される Biternion 表現の一致度を Sim として評価する．

4.3.3. 性能比較

結果を表 1 に示す．Full は SC 次元の特徴マップの全要素にプーリングを行わずそのまま特徴表現として用いた場合を表している．w/o IN，w/IN はそれぞれ，IN を適用しない場合，適用する場合である．結果より以下のことが読み取れる．

表 1　Person X データセットにおける性能比較.

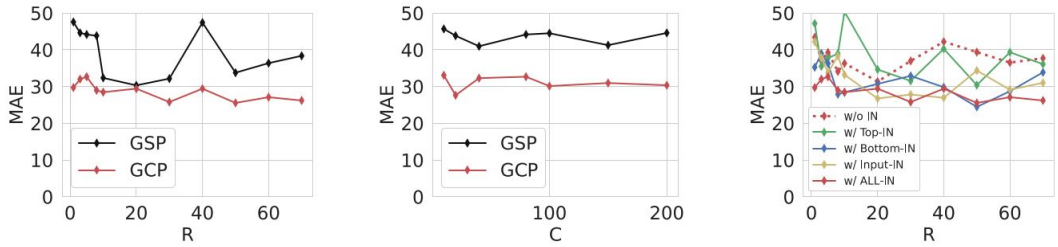
	w/o IN					w/ IN				
	ORG		Styled			ORG		Styled		
Pool	$Acc_{10^\circ}\uparrow$	MAE ↓	$Acc_{10^\circ}\uparrow$	MAE ↓	SIM ↑	$Acc_{10^\circ}\uparrow$	MAE ↓	$Acc_{10^\circ}\uparrow$	MAE ↓	SIM ↑
without pooling										
Full	95.2	3.6	69.3	9.7	97.0	94.9	3.6	84.6	5.9	**99.2**
spatial pooling										
BSP	94.2	4.3	68.9	10.1	96.4	95.8	3.4	84.5	6.3	98.7
GSP	**96.2**	3.3	75.2	8.3	97.6	95.9	**3.3**	86.0	5.9	99.0
channel pooling										
Avg	95.0	3.7	76.8	8.2	97.8	94.0	4.0	82.2	6.8	98.6
Std	95.6	3.5	76.1	8.2	97.8	95.0	3.7	83.4	6.5	98.9
Max	94.6	3.8	78.1	7.6	98.1	95.2	3.6	82.7	6.6	98.8
Avg+Std+Max	95.5	3.4	77.2	7.6	**98.3**	**96.0**	**3.3**	84.9	6.0	98.9
GCP	96.0	**3.2**	**79.4**	**7.4**	98.1	95.6	3.4	**87.2**	**5.6**	**99.2**

(1) チャネルプーリングにおいて GCP は Styled データに対して最良の性能を示し，素朴なチャネルプーリング手法(Avg,Std,Max とそれらの連結) に対しての有効性が確認できる．
(2) 空間プーリングにおいて，GSP は BSP よりも Styled データで高い性能を示している．これにより，特異値を用いないことによるスタイル変動への頑健性が確認できる．
(3) GCP は，GSP よりも Styled データでの性能が高い．これにより，グラム行列における特徴基底Vを除外することによるスタイル変動への頑健性の向上が確認できる．

4.3.2 パラメータ解析

TUD データセット[12]に対してもスタイル変換を行ったテストデータを作成し，そのデータにおいてパラメータ解析を行った．結果を図 2 に示す．

(1) 基底数／入力チャネルの次元数．
図 2(a),(b)は GSP と GCP の性能を基底数 R と入力チャネル数 C を変更して比較した．C は 1x1 畳み込み層の次元を変動させることで行った．結果より GCP がどの基底数/入力チャネル次元数でも GSP よりも高い性能(小さい MAE)が得られていることが解る．
(2) IN の位置．
図 2 (c)は IN を挿入する位置 (Input-IN，Bottom-IN，Top-IN)とそれら全てを用いる場合(ALL-IN)を比較している．結果よりどの位置に IN を挿入しても，スタイル変動のあるテストデータでの性能が向上すること，ALL-IN を用いることで最良の結果が得られる傾向にあることが解る．

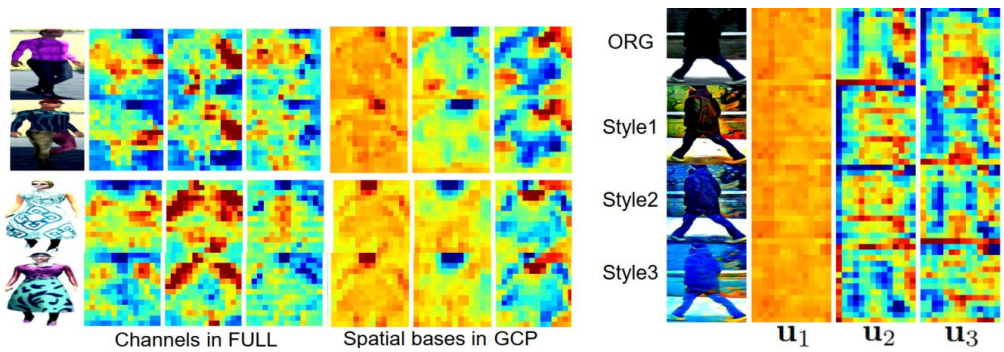


(a) Number of bases (b) Number of channels (c) Places of IN

図 2 パラメータ解析 (TUD データセット)

4.3.4.質的解析

(1) 頑健性の例
図 3(a)は Full の最初の 3 つの特徴チャネルと GCP の最初の 3 基底を比較したものである．この結果は IN を適用しない状態で得たものである．Full においては服装の変動がチャネル活性化の変動を起こしている．一方で GCP の空間基底はこの影響が少ないことが解る．図 3(b)は，スタイル変動により類似した空間基底の順序(特異値の降順)が変動する場合を示している．しかし，GCP はこの順序を無視しているため，この変動の影響を受けない．

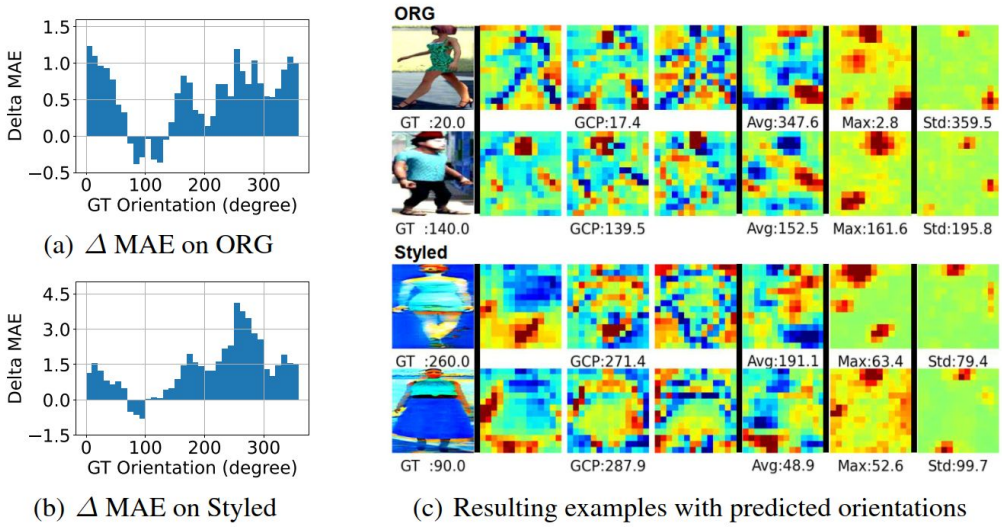


(a) Channel of Full vs. spatial bases (b) Robustness against style changes

図 3 頑健性の例 (PersonX データセット)

(2) 角度毎の解析

図 4 (a),(b)に 10 度毎に算出した GCP の MAE の Avg に対する利得 ΔMAE = MAE(Avg)-MAE(GCP)を示す．ΔMAE が大きいほど，GCP の性能向上が大きいことを示す．ORG データでは 20 ° と 140 ° での GCP の性能向上が大きく，Styled データでは 260 ° での性能向上が大きかった．図 4(c)はこれらの角度での例を示している．GCP は比較手法よりも人物の身体形状をよく表現できている．Styled データにおいて，Max と Std が誤差の大きい背面の画像(260 °)では正面顔に対応するとみなされるチャネル活性化がみられた．図 4(c)の最下部は逆に GCP が正面向き(90 °)を間違えた場合であるが，図 4(b)から読み取れる通り，GCP が比較手法よりも失敗する場合は少なかった．



(a) Δ MAE on ORG

(b) Δ MAE on Styled

(c) Resulting examples with predicted orientations

図 4 質的な結果 (PersonX データセット). (c) において GT は正解の角度，各手法の数値は推定角度を表す.

<引用文献>

[1] S.Woo, J.Park, J.Y.Lee, I.S.Kweon: CBAM: Convolutional block attention module, In: ECCV 2018.
[2] X.Sun, L.Zheng: Dissecting person re-identification from the viewpoint of viewpoint, In: CVPR2019.
[3] Z.Huang, L.V.Cool: Building deep networks on Grassmann manifolds, In: AAAI 2018.
[4] 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.
[5] X.Wei, Y.Zhang, Y.Gong, J.Zhang, N.Zheng, Grassmann pooling as compact homogenous bilinear pooling for fine-grained visual classification, In: ECCV 2018.
[6] L.A.Gatys, A.S.Ecker, M.Bethge: Image style transfer using convolutional neural networks, In: CVPR2016.
[7] J.Wang, K.Sun, T.Cheng, B.Jiang, C.Deng, Y.Zhao, D.Liu, Y.Mu, M.Tan, X.Wang, W.Liu, B.Xiao: Deep high-resolution representation learning for visual recognition, IEEE Trans. on PAMI, vol.43, no.10, pp.3349-3364, 2021.
[8] D.Ulyaanov, A.Vedaldi, V.Lempitsky, Instance normalization: The missing gradient for fast stylization. arXiv:1607.08022, 2016.
[9] L.Beyer A.Hermans, B.Leibe: Biternion Nets: continuous head pose regression from discrete training labels, In: GCPR 2015.
[10] M.D.Zeier: ADADELTA: An Adaptive Learning Rate Method, arXiv:1212.5701, 2012.
[11] J.Wang, H.Yang, J.Fu, T.Yamasaki, B.Guo: Fine-grained image style transfer with visual transformers, In: ACCV 2022.
[12] M.Audriluka, S.Roth, B.Schiele: Monocular 3D pose estimation and tracking by detection, In: CVPR 2010.

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　1件）

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 25th International Conference on Pattern Recognition (ICPR2020)（国際学会）
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