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　２ 年　 ６ 月 １０ 日現在

機関番号：１４４０１
研究種目：挑戦的研究（萌芽）
研究期間：2018～2019
課題番号：１８Ｋ１９８１８

研究課題名（和文）ディープコンピュテーショナルフォトグラフィ

研究課題名（英文）Deep computational photography

研究代表者
長原　一（Nagahara, Hajime）

大阪大学・データビリティフロンティア機構・教授

研究者番号：８０３６２６４８

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）：画像認識の分野ではディープニューラルネットワーク(DNN)が盛んに用いられ，物体認識やシーン理解，画像復元などにおいて，従来のモデルベースの特徴量や学習手法を凌駕している．しかし，従来はデジタル画像として計測された後の画像認識パイプラインのデジタル層にのみ，DNNによる学習が用いられているにすぎなかった．本研究では，特徴量や認識器と共にハードウェア設計も学習により求める新しいフレームワークを提案し，その有効性を実証した．

研究成果の学術的意義や社会的意義
これまでのカメラの設計は，サンプリング理論やノウハウにより設計者の手動により設計されてきた．これに対して，本研究では，データ駆動におる学習アプローチにより，カメラの設計パラメータを最適化することにより，応用に即したハードウェアを設計することで性能向上を行った．このようなアプローチは，Deep sensingやDeep opticsなどと呼ばれ，後追い研究を呼び最近の研究の新しい流れのひとつとなっている．

研究成果の概要（英文）：Deep learning is getting popular in computer vision and it drastically improve the performance of object recognition, scene understanding and image reconstruction etc. However, regular deep learning is applied to the digital domain of a camera pipeline and ignored a physics layer such as optics and sensor of a camera. In this paper, we proposed a framework for modeling the camera pipeline including the physics layer as well as the digital layer and optimize the camera design parameters and task model by learning. We have demonstrated this framework to the compressive video sensing and action recognition tasks and show the effectiveness of the approach.

研究分野：コンピュテーショナルフォトグラフィ

キーワード：コンピュテーショナルフォトグラフィ　ディープラーニング

様　式　C－１９、F－１９－１、Z－１９（共通）

１．研究開始当初の背景

画像認識の分野ではディープラーニング/ディープニューラルネットワーク(DNN)や CNN(コンボリューショナルニューラルネットワーク)が盛んに用いられ，物体認識やシーン理解，画像復元などにおいて，従来のモデルベースの特徴量や学習手法を凌駕している．従来の DNN による画像認識では,図１の内枠内に示す様にデジタル画像を入力とし,特徴抽出を担う下層の CNN 層と認識器を担う上層の全結合ネットワークによる DNN が用いられてきた．このフレームワークにより，従来モデルベースで設計されてきた SIFT や HoG，LBP などの特徴量を使う場合に比べて,目的とする認識タスクに特化した最適な特徴抽出パターンが学習により CNN 層で得られることで,従来の手動による特徴選択の方法にくらべて格段の性能向上を実現している．しかしながら，画像認識における計測や処理の段階プロセス(画像認識パイプライン)を考えた場合，実シーンから発せられる光線の時空間情報は，カメラの光学系によりセンサ上の光学像として形成される．CMOS などの撮像センサが，そのセンサ像の輝度値をグリッド配置された離散ピクセルにより，すべての画素が同期した時間窓関数でサンプリングすることでデジタル画像を得る．従来は，図１に示すデジタル画像として計測された後の画像認識パイプラインのデジタル層にのみ，DNN による学習が用いられているにすぎなかった．一方で，申請者がこれまで牽引してきたコンピュテーショナルフォトグラフィ(CP)では[1]，画像処理や認識のためにどのようなハードウェアで画像をセンシングすべきかをアナログ層も含めて長年議論してきた．ただ，従来の光学設計やセンサ設計は，主に光学や信号処理理論に基づく解析的アプローチにとどまっていた．しかしながら，すべてのシーンや認識タスクが解析のベースとなる理論的条件や背景を満たしているわけではない．

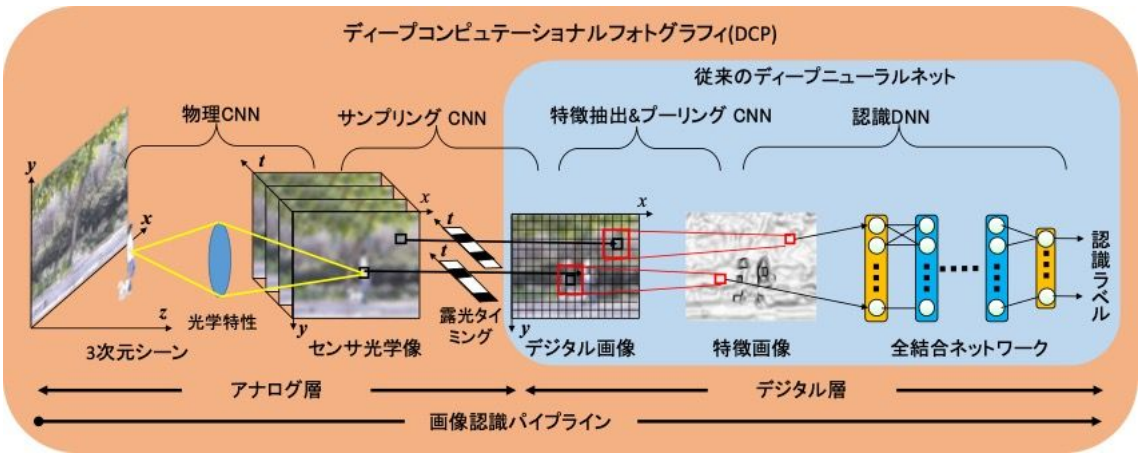

図１：ディープコンピュテーショナルフォトグラフィ(DCP)の概要

２．研究の目的

本研究では，特徴量や認識器と共にハードウェア設計も学習により求める新しいフレームワークを提案し，ディープコンピュテーショナルフォトグラフィ(DCP)と名付ける．DCP では，図の画像認識パイプラインをアナログ層も含めてすべて DNN で表現する．従来の画像特徴抽出のための CNN のさらに下に，画素の時間露光タイミングを表現するサンプリング CNN 層と集光特性を表現する物理 CNN 層をさらに加える．これらの画像認識パイプライン全体を DNN で表現し，シーンとラベルセットにより学習することで従来の認識および画像特徴が学習されると共に，カメラハードウェア設計をサンプリング CNN 層および物理 CNN 層から学習により得る．その学習結果を基にカメラハードウェア試作を行い，学習によるカメラ設計の有用性を検証する．具体的には，圧縮ビデオセンシング(少ないサンプリングからの動画生成)や単一画像からの人の行動認識を事例タスクとして設定し，このタスクに最適化したカメラシステムを実現する．

３．研究の方法

DCP では，図１の画像認識パイプラインをデジタル層だけでなくアナログ層も含めて DNN で表現する．従来の画像特徴抽出のための CNN のさらに下に，画素の時間露光タイミングを表現するサンプリング CNN 層と集光特性を表現する物理 CNN 層をさらに加える．これらの画像認識パイプライン全体を DNN で表現し，シーンとラベルセットにより学習することで従来の認識および画像特徴が学習されると共に，カメラハードウェア設計もサンプリング CNN 層および物理 CNN 層から学習により得る．つまり，提案する新たなフレームワークである DCP は，様々な物理的拘束条件下で人が職人技的に暗黙に最適化を行ってきたハードウェア設計に，学習による最適化を持ち込むことで，タスクに必要な設計を求めるスマートセンサ設計手法であるとも言える．当然ながら，画像認識パイプラインを一貫して最適化しているため，ハードウェア設計が最適化されると共に，その入力画像を用いた最適な特徴抽出や識別器も学習時に同時に得られる．具体的な DCP のタスクとして圧縮ビデオセンシングと単一画像からの行動認識の研究を行った．

(1) 圧縮ビデオセンシング

浜松ホトニクスとの共同研究において，プログラム露光が可能な CMOS センサを用いたビデオ圧縮センシングの研究を行ってきた[2]．圧縮ビデオセンシングとは，近傍画素の露光タイミングを故意にずらすことで複数時間の情報を単一画像として取り込み，この単一画像から動画を生成する手法で撮影自体が圧縮の効果を持つ．この露光タイミングを DCP により最適化することで，圧縮センシングにおける露光パターン設計を行なった．

(2) 単一符号化露光画像からの行動認識

フログラム露光による符号化画像には，隣接ピクセルの露光タイミングのずれから，単一画像であっても複数の時間の情報を含んでいる．本研究では，この符号化露光画像を直接用いた人物の行動認識を実現する．露光タイミングと行動認識モデルの双方をニューラルネットワークでモデル化し，学習により認識モデルのみならず，行動認識に最適な符号化露光パターンを同時最適化により得る．その結果，最小データで行動認識可能な小データ･省電力の IoT センサが実現できる．

４．研究成果

(1) 圧縮ビデオセンシング

図 2 に提案する DNN はハードウェアの制約を満たしながら露光パターンの最適化を行うセンシング層と観測画像から動画を再構成する再構成層の二つの部分より構成されている．このネットワークを用いて露光パターンを再構成層と同時に最適化することで高品質な再構成を行うことができる露光パターンを求めることができる．

露光パターンはハードウェアへの実装上の制約を考慮しなければならない．我々は事前にハードウェアの制約を満たすすべての露光パターンを用意し，その中からネットワークに最適なパターンを選択させることで最適な動画の符号化を行う．制約のある露光パターンはハードウェアの構造から簡単に求めることができる．実験に用いるセンサにおいては，すべての Reset 信号(8bit) と Transfer 信号(8bit) の組を生成する．次に，生成したすべての信号組から生成される露光パターンをシミュレートすることですべての露光パターンを求める．実際の圧縮センシングでは 2 値の露光パターンが用いられるため，ネットワークの訓練での Forward 時には 2 値の重みを用いるが，Backward 時には微分可能とするため連続値に緩和する[3]．次の Forward 時に用いる重みは事前に生成した 2 値の露光パターンの中から連続値の重みと最も近いものを内積により選んだ．再構成層は，圧縮センシング層で学習された露光パターンで圧縮された単一の画像から複数フレームの動画を生成する．この単一の画像から複数フレームの動画への非線形写像を Multi Layer Perceptron (MLP) を用いて学習する．図 2 に示すように，MLP は 4 層の隠れ層を持ち，伝達関数には ReLU を用いる．ネットワークは訓練動画と再構成動画の誤差を小さくするように学習する．損失関数は，再構成動画の評価にピーク信号対雑音比(PSNR) を用いるために，関係の深い平均二乗誤差(MSE) を用いた．

提案するネットワークを学習するための訓練データは 20 本の動画から 16 フレームをランダムに 4 シーンずつ取り出し，それぞれに回転(90◦; 180◦; 270◦) と反転を行ったものを用いた．このようにして用意した 829,440 パッチを用いて，提案する露光パターンと再構成ためのデコーダを同時に最適化するネットワークを end-to-end で学習した．学習はミニバッチサイズ 200 で 500epoch 行った．

露光パターンとデコーダの同時最適化の効果を検証するためシミュレーション実験を行った．各露光パターンで撮影される画像をシミュレートし再構成ネットワークへ入力，動画を再構成することで疑似ランダム露光パターン[2] と最適化した露光パターンの性能を比較した．使用したのは空間解像度 256x256 pixel の 16 フレームの動画 14 本である．再構成品質はピーク信号対雑音比(PSNR) により評価した．図 3 に結果のうち 2 例を示す．図 3 上段(Car)を見ると，手紙のマークの部分が最適化した露光パターンではよりシャープになっていることがわかる．また，図 3 下段(Crushed can) を見ると，最適化した露光パターンではロゴがはっきりし，パッチの境界もより滑らかになっている．再構成品質はシーンごとに異なるが，すべてのシーンで最適化した露光パターンのほうが再構成品質が良かった．

また，実際に画素ごとに露光を制御できるセンサを用いて撮影を行い，実実験を行った．撮影は 15FPS で行い，1 フレームから 16 サブフレームを再構成するため再構成後の動画は 240FPS 相当となる．図 4 に実際に撮影した画像と再構成結果を示す．上段はメトロノームが振れているシーンであるが，メトロノームのおもりが移動している様子が捉えられている．また，中段ではまばたきの様子を，下段では硬貨が水中で落ちる様子を見ることができる．

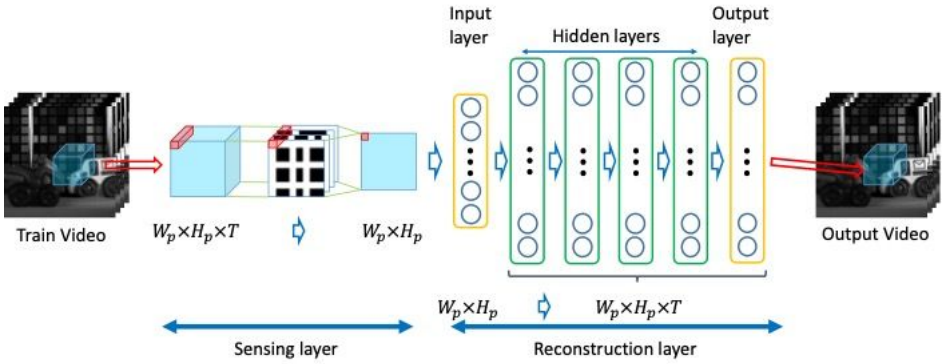

図 2: ビデオ圧縮センシングのための DCP ネットワーク

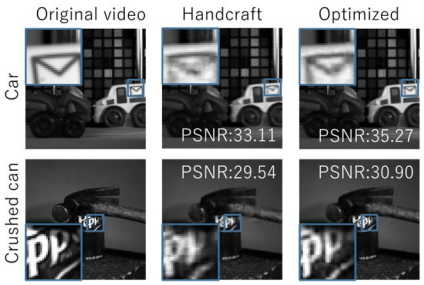

図 3:露光パターンの比較

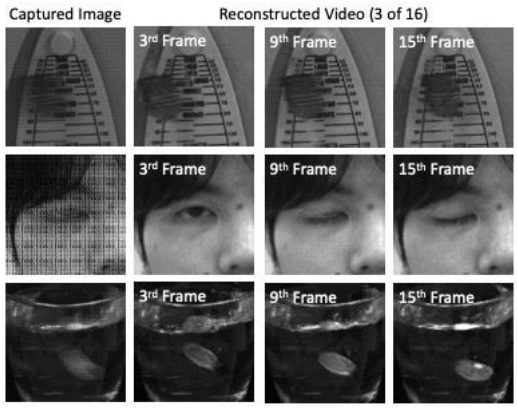

図 4: 圧縮ビデオセンシング実実験結果

(2) 単一符号化露光画像からの行動認識

提案モデルは図 5 に示すように，主に符号化ネットワークと分類ネットワークの 2 つの部分で構成されている．符号化ネットワークは，符号化露光カメラによる符号化露光画像の撮影を表現する．このネットワークは長さ L の p × p のブロックごとの 2 値化 1D CNN として記述される．この 2 値化 1DCNN のパラメータを学習することで最適化された符号化露光パターンを得ることができる．分類ネットワークは，符号化露光画像から行動ラベルを推定する第 1 層に Shift-variant convolution を用いた 12 層からなるニューラルネットワークである．通常の畳み込みカーネルは，隣接する画素は空間的に滑らかであると仮定してシフト不変(shift-invariant) であるのに対し，Shift-variant convolution は，符号化画像の隣接する画素が異なるタイミングで露光される特徴に合わせて，画素位置毎にシフト変位(shift-variant) カーネルでのコンボリューションを実現する．

このネットワーク全体を通常の CNN の学習と同様に動画と行動ラベルのセットを入力として end-to-end で学習する．これにより，行動認識に最適な符号化露光と，その符号化露光画像から行動を分類するモデルを同時に学習することができる．獲得された符号化露光パターンは分類モデルとタスクに最適化され，分類モデルも同様に符号化露光画像の露光パターンに最適化される．したがって，このフレームワークの下，カメラの符号化露光と行動認識の分類モデルという 2 つを同時に最適化することが可能である．実実験では学習により最適化された符号化露光パターンを符号化露光カメラに実装することで実シーンの撮影を行い，撮影画像を行動認識の分類モデルに入力することで行動認識を行う．

Something-Something[4]は人間と物体のやり取りに関する大規模なデータセットであり，時間的な関係性が必要とされる曖昧なカテゴリを含む 174 クラスの行動ラベルがある．このデータセットはテストセットが公開されていないため，検証セットでの結果を表 1 に示す．時間情報が含まれていないため短時間露光画像での認識精度は低かった．長時間露光画像では，時間の経過とともに明るさの変化が積分され，カメラや物体の動きによりモーションブラーが発生する．モーションブラーは，ブラーの形状としてモーション情報を提供するが，物体の形状を不鮮明にする．したがって，モーションブラーにより物体の認識が困難になるため，長時間露光画像の認識精度は低かった．ランダムな符号化露光パターンを使用した C2D での認識精度は前の 2 つの結果よりも高い精度となり，通常のシフト不変の畳み込みを使用した C2D においても符号化露光パターンを最適化することで認識精度を大幅に改善した．さらに，我々の提案手法である Shift-variant convolution を使用した SVC2D を用いた場合は，他の単一画像からの行動認識の結果よりも高い認識精度を示した．通常畳み込みでは，符号化露光の固有な時空間パターンは画像の一部から決定されるため，シーンによって誤対応が発生する．我々は符号化露光パターン上の位置に応じて畳み込みカーネルを変更するというシンプルなアプローチでこの問題を解決した．

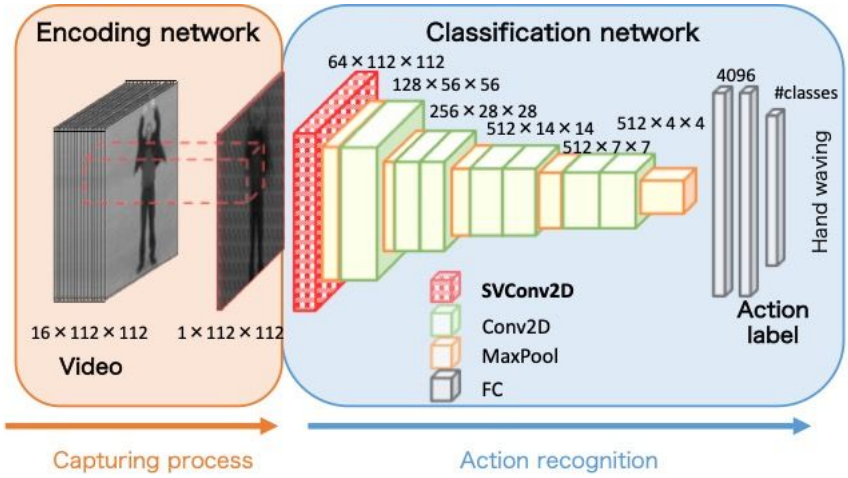

図 5: 行動認識のための DCP ネットワーク

表 1: Something Something データセットでの検証

Input	Pattern	Model	Top-1	Top-3	Top-5
(a) Coded	optimize	SVC2D	29.37	47.39	56.33
	optimize	C2D	26.77	45.66	54.37
	random		12.75	25.63	32.69
(b) Long		C2D	10.82	22.83	30.20
(c) Short		C2D	10.32	21.85	28.56
(d) Video		C3D	**39.31**	**61.97**	**70.05**

[1] 長原一, “I コンピュテーショナルフォトグラフィ-符号化撮像、ライトフィールド、圧縮センシングなどの新しい画像センシングとその応用-”, 画像センシングシンポジウム 2017 チュートリアル講演会(招待講演, 参加者 3000 人規模).

[2] T. Sonoda, H. Nagahara, K. Endo, Y. Sugiyama, R. Taniguchi, "High-speed imaging using CMOS image sensor with quasi pixel-wise exposure", International Conference on Computational Photography (ICCP), pp.1-11,2016.

[3] M.Courbariaux, I.Hubara, D.Soudry, R.El-Yaniv, Y.Bengio, "Binarized neural networks: Training neural networks with weights and activations constrained to +1 or − 1." arXiv preprint arXiv:1602.02830 2016.

[4] Goyal, R., Kahou, S. E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M. et al.: The ” Something Something ” Video Database for Learning and Evaluating Visual Common Sense., Proceedings of International Conference on Computer Vision (ICCV), Vol. 1, No. 2, p. 3, 2017.

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 Yoshida Michitaka、Sonoda Toshiki、Nagahara Hajime、Endo Kenta、Sugiyama Yukinobu、Taniguchi Rin-ichiro	４．巻 6
２．論文標題 High-Speed Imaging Using CMOS Image Sensor With Quasi Pixel-Wise Exposure	５．発行年 2020年
３．雑誌名 IEEE Transactions on Computational Imaging	６．最初と最後の頁 463～476
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/TCI.2019.2956885	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9件（うち招待講演　3件／うち国際学会　8件）

１．発表者名 Michitaka Yoshida, Akihiko Torii, Masatoshi Okutomi, Kenta Endo, Yukinobu Sugiyama, Rin-ichiro Taniguchi, Hajime Nagahara
２．発表標題 Joint optimization for compressive video sensing and reconstruction under hardware constraints
３．学会等名 International Workshop on Image Sensors and Imaging Systems（国際学会）
４．発表年 2018年

１．発表者名 Michitaka Yoshida, Akihiko Torii, Masatoshi Okutomi, Kenta Endo, Yukinobu Sugiyama, Rin-ichiro Taniguchi, Hajime Nagahara
２．発表標題 Joint optimization for compressive video sensing and reconstruction under hardware constraints
３．学会等名 European Conference on Computer Vision（国際学会）
４．発表年 2018年

１．発表者名 大河原 忠，吉田 道隆，長原 一，八木康史
２．発表標題 符号化露光画像を用いた人物の行動認識
３．学会等名 情報処理学会CVIM研究会
４．発表年 2019年

１．発表者名
Hajime Nagahara

２．発表標題
Computational photography using programmable sensors

３．学会等名
International Workshop on Image Sensors and Imaging Systems（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Hajime Nagahara

２．発表標題
Coded Computational Photography

３．学会等名
Korea-Japan workshop on Digital Holography and Information Photonics（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Hajime Nagahara

２．発表標題
Coded Computational Photography

３．学会等名
International Workshop on Advanced Image Technology（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Michitaka Yoshida, Akihiko Torii, Masatoshi Okutomi, Kenta Endo, Yukinobu Sugiyama, Rin-ichiro Taniguchi, and Hajime Nagahara

２．発表標題
Joint optimization for compressive video sensing and reconstruction under hardware constraints

３．学会等名
International Conference on Computational Photography（国際学会）

４．発表年
2019年

１．発表者名 Kenta Endo, Yukinobu Sugiyama, Michitaka Yoshida, Hajime Nagahara, Kento Kaneta, Keisuke Uchida, Yasuhito Yoneta, and Masaharu Muramatsu
２．発表標題 Functional CMOS Image Sensor with flexible integration time setting among adjacent pixels
３．学会等名 International Conference on Computational Photography（国際学会）
４．発表年 2019年

１．発表者名 Tadashi Okawara, Michitaka Yoshida, Hajime Nagahara, Yasushi Yagi
２．発表標題 Action Recognition from a Single Coded Image
３．学会等名 International Conference on Computational Photography（国際学会）
４．発表年 2020年

〔図書〕　計0件

〔出願〕　計1件

産業財産権の名称 動画像処理方法及び動画像処理装置	発明者 長原一、大河原忠、吉田道隆	権利者 大阪大学
産業財産権の種類、番号 特許、特願2019-001491	出願年 2019年	国内・外国の別 国内

〔取得〕　計0件

〔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