コグニティブセンシング：マルチピンホールカメラ画像からのディープラーニングによる奥行推定

Cognitive Sensing: Depth Estimation with Multi-Pinhole Camera using Deep Learning

佐藤 智 Satoshi Sato
ポンサク ラサン Pongsak Lasang
シェン メイシェン Shen Shengmei, Jane
若井 信彦 Nobuhiko Wakai
登 一生 Kunio Nobori
吾妻 健夫 Takeo Azuma

要　旨

認識システムの入力デバイスとしてカメラが広く利用されている．しかし，撮像された画像には冗長な情報が多く含まれ，また，入力データ量は認識システムのハードウェアサイズに直結するため，従来のカメラは最適な入力でなかった．筆者らは，認識に有効な情報のみを効率的に取得し，軽量のハードウェアで高性能な認識を実現するコグニティブセンシングを提案する．本稿ではその一例として，マルチピンホールカメラ画像からの奥行推定について述べる．マルチピンホール画像は，視点の異なる画像が重畳されており，人間が見るには適さないが，認識処理に有効な被写体のテクスチャ情報と奥行情報が圧縮されて含まれるため，認識システムには有効である．ディープラーニングを用いたシミュレーションにより従来のカメラを利用した場合と比較して，入力データ量を増加させずに，奥行推定精度が向上することを確認した．

Abstract

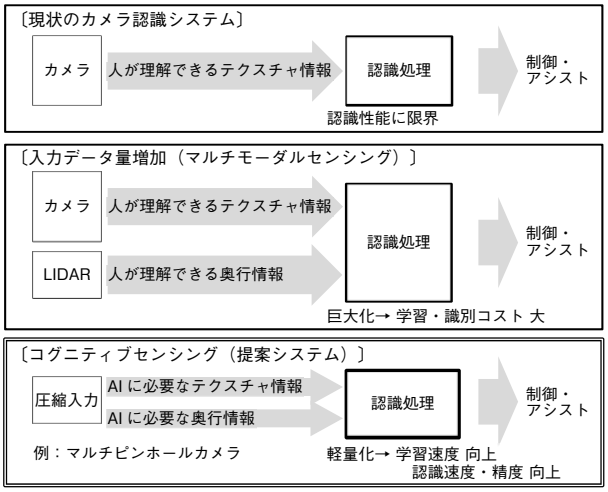
We propose a new recognition system named cognitive sensing, which acquires effective input only compared to a conventional camera including pinhole cameras. Although the conventional camera is one of the most popular input devices, it is not the most effective device owing to its redundant information, which incurs heavy computational cost. To overcome this problem, we use a multi-pinhole camera as an example of an effective input device. These cameras are effective for recognition systems because the captured images include both texture and depth information. Simulation results show that deep convolutional neural networks using multi-pinhole images estimate higher accuracy depth maps than those using pinhole images.

1. はじめに

認識システムは，処理アルゴリズムと入力デバイスの両面で大きな変化を迎えている．処理アルゴリズムに関しては，ディープラーニングをはじめとする機械学習技術の急速な発展により，さまざまな認識タスクの性能が飛躍的に向上している．入力デバイスに関しては，カメラの高解像度化･高フレームレート化･超広角化に加え，Time of Flight（TOF） センサをはじめとするレーザスキャナやステレオカメラなどの奥行センサ，赤外線カメラやマルチスペクトルセンサを組み合わせたマルチモーダルセンシングの一般化により，取得できるデータ量がますます増加している．このような入力デバイスのデータ量の増加は，認識性能の向上につながる[1]が，一方で，ネットワークの巨大化，処理時間・学習負荷・ハードウェアコストの増加をまねき，実用化の大きな障壁となっていた．

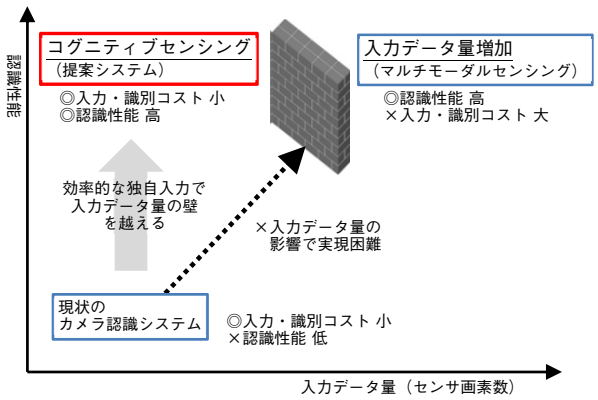
これに対し，筆者らは，入力デバイスから認識処理までのシステム全体を認識に必要なデータ量の観点から最適化した，コグニティブセンシングを提案する．ディープラーニングの認識処理は，Autoencoderとしても知られているように，次元圧縮を行っている．すなわち，入力データ量が豊富であっても認識に有効な情報量はより小さく，さらにその情報は人間が見ているカメラ画像とも異なっていると想像される．コグニティブセンシングではこの知見に基づき，入力データ量を効率的に圧縮して取得する圧縮入力デバイスを利用し，さらに，この圧縮入力データを直接，認識する処理アルゴリズムを組み合わせる．これにより，認識性能を落とさずに入力デバイスの負荷を減少させ，ハードウェアの軽量化，学習速度の向上，さらには認識速度･精度の向上が期待できる（**第1図**）．言い換えると，コグニティブセンシングにより，現状のカメラ認識システムと同程度の入力データ量でありながら，マルチモーダルなシステムと同等の認識性能を実現できる（**第2図**）．これは，対象となる信号をできるだけ少ない観測によって復元する圧縮センシング技術[2][3]とも深い関係がある．

コグニティブセンシングの圧縮入力デバイスの一例として，筆者らは，マルチピンホールカメラに着目する．マルチピンホールカメラは，重畳画像を撮像する．この画像は，人間が見るには適していない．しかし，重畳された各画像は視点位置が異なるため，入力データ量を増



LIDAR：Light Detection and Ranging

第1図　コグニティブセンシングとマルチモーダルセンシング
Fig. 1　Cognitive sensing and multimodal sensing



第2図　コグニティブセンシングの概念
Fig. 2　Concept of Cognitive Sensing

加させることなく，テクスチャ情報に加えて奥行情報を圧縮して取得しており，認識処理に適している．

本稿では，コグニティブセンシングの圧縮入力デバイスにマルチピンホールカメラを利用し，Deep Convolutional Neural Network（DCNN）により奥行マップを推定する．

2. マルチピンホール画像からの奥行推定

本章では，コグニティブセンシングの圧縮入力デバイスとしてマルチピンホールカメラを利用した，奥行推定方法について説明する．まず，マルチピンホール画像はテクスチャ情報と奥行情報が圧縮された圧縮入力であり，認識システムの入力デバイスに有効であること，さらにマルチピンホールカメラの将来性を説明し，その後，奥行推定に利用したDCNNの構成を説明する．

2.1　ピンホールカメラとマルチピンホールカメラ

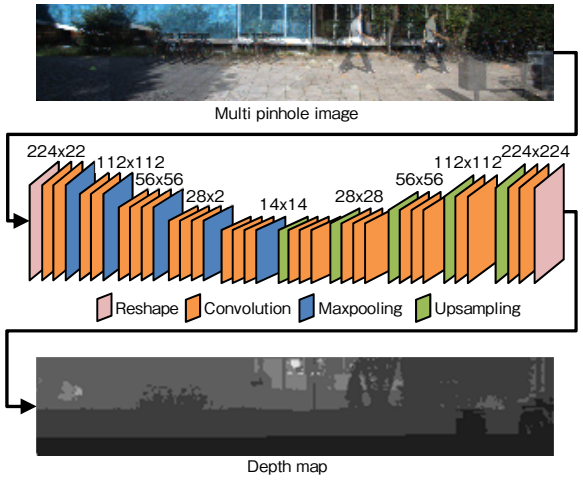
通常のカメラは，広義のピンホールカメラである．ピンホールカメラは，**第3図**に示すように，空間上に1点の焦点（ピンホール）を有する．このピンホールを通過した光線のみをイメージセンサで受光し，合焦した画像をテクスチャ情報として取得する．一般的に集光機能を有する主レンズを利用するが，これは，ピンホールを通過しない光線は捕捉されないために光量が低下し，SN比（Signal to Noise Ratio）が低くなるためである．また，光線上のいずれの距離にある被写体もイメージセンサ上の同じ位置に撮像されるため，被写体の奥行情報は取得できない．

一方，マルチピンホールカメラは，Coded Aperture Cameraの1種であり，**第3図**に示すように複数のピンホールを有し，各ピンホールを通過した光線を1つのイメージセンサで受光する．複数のピンホールを有することで，光量すなわちSN比が向上するため，主レンズを有しない超薄型・低コストのレンズレスカメラの実現も可能である[4]．また，撮像されたマルチピンホール画像は異なる視点位置のピンホール画像が重畳された画像，すなわちステレオ画像の多重像となるため，人間が見る画像には適さないが，認識処理に有効な被写体のテクスチャ情報と奥行情報が圧縮されて含まれる．このことから，マルチピンホールカメラは認識システムに有効なマルチモーダル情報の圧縮入力デバイスであると期待される．

第3図　ピンホールカメラとマルチピンホールカメラ
Fig. 3　Pinhole camera and multi-pinhole camera

2.2　ディープラーニングによる奥行推定

マルチピンホール画像から奥行を推定するDCNNの構成を**第4図**，各パラメータを**第1表**に示す．このネットワ



第4図　奥行推定のためのDCNN

Fig. 4　Deep convolutional neural networks for depth estimation

ークは，パッチに切り出したカラー画像を入力とし，畳み込み層とプーリング層を繰り返すことで，各画素について量子化された奥行値，すなわち奥行マップを推定する．**第4図**の各層上部に，今回利用した特徴マップのサイズを示す. このDCNNは, Deep Depth From Focus（DDFF）ネットワーク[5]の入力層のみを変更して作成した．DDFFネットワークはVGG-16[6]をベースにしており，焦点位置をカメラの近傍から遠方まで変化させて撮像した複数枚の画像群であるFocal Stack画像から奥行マップを推定する．

各画像は224×224画素ごとのパッチに切り出してネットワークに入力した．ストライドは56である．すべての畳み込み層のカーネルサイズは3×3とし，その直後にBatch Normalization層を挿入した[7]．活性化関数はReLUを，ネットワークの初期化にはHeの手法を利用した[8]．学習は，Softmax Cross Entropy損失をMomentum Stochastic Gradient Descent （Momentum SGD） により最適化した．MaxPooling層，Upsampling層のDropout率は0.5である．

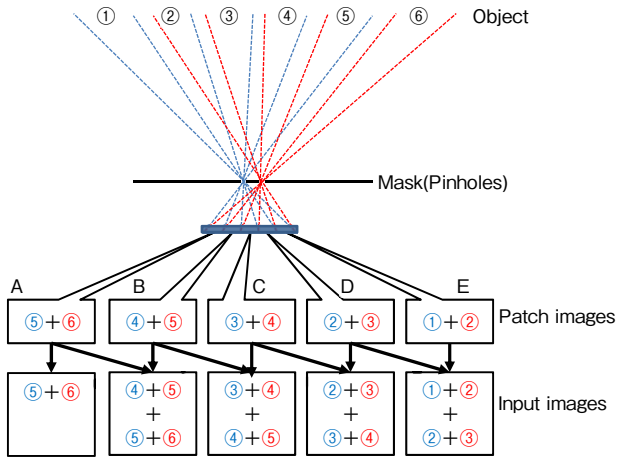
マルチピンホール画像は，各ピンホール位置に対応する視差を有する画像が重畳されており，この視差情報が奥行情報と関係している．しかし，画像をパッチに切り出すため，**第5図**に示すように，各パッチには被写体の1つのピンホール画像しか含まれなくなる可能性がある．特にピンホール間隔が広い場合, この影響が大きくなる．この問題に対応するため，筆者らは重畳画像であるマルチピンホール画像と，各ピンホールの位置に合わせて平行移動させた画像を重畳し，これを切り出したパッチをネットワークへ入力した．これにより，各ピンホール画像は各パッチ内で重畳されるため，切り出しによる奥行推定精度の低下を抑えることができる．**第5図**において，

第1表　DCNNパラメータ

Table　DCNN parameters

	in channels	Out channels	kernel size	stride	padding
Convolution 1_1	3	64	3×3	1	1
Convolution 1_2	64	64	3×3	1	1
MaxPooling	64	64	2×2	2	0
Convolution 2_1	64	128	3×3	1	1
Convolution 2_2	128	128	3×3	1	1
MaxPooling	128	128	2×2	2	0
Convolution 3_1	128	256	3×3	1	1
Convolution 3_2	256	256	3×3	1	1
Convolution 3_3	256	256	3×3	1	1
MaxPooling	256	256	2×2	2	0
Convolution 4_1	256	512	3×3	1	1
Convolution 4_2	512	512	3×3	1	1
Convolution 4_3	512	512	3×3	1	1
MaxPooling	512	512	2×2	2	0
Convolution 5_1	512	512	3×3	1	1
Convolution 5_2	512	512	3×3	1	1
Convolution 5_3	512	512	3×3	1	1
MaxPooling	512	512	2×2	2	0
Upsampling	512	512	1×1	1	0
Convolution 6_1	512	512	3×3	1	1
Convolution 6_2	512	512	3×3	1	1
Convolution 6_3	512	512	3×3	1	1
Upsampling	512	512	1×1	1	0
Convolution 7_1	512	512	3×3	1	1
Convolution 7_2	512	512	3×3	1	1
Convolution 7_3	512	256	3×3	1	1
Upsampling	256	256	1×1	1	0
Convolution 8_1	256	256	3×3	1	1
Convolution 8_2	256	256	3×3	1	1
Convolution 8_3	256	128	3×3	1	1
Upsampling	128	128	1×1	1	0
Convolution 9_1	128	128	3×3	1	1
Convolution 9_2	128	64	3×3	1	1
Upsampling	64	64	1×1	1	0
Convolution 10_1	64	64	3×3	1	1
Convolution 10_2	64	8	3×3	1	1

パッチEには，左のピンホールを通した被写体①と，右のピンホールを通した被写体②が重畳撮像されるが，被写体が異なるため，視差情報は含まれない．そこで，画像処理によりパッチDとEを重畳し，ネットワークへ入力した. この入力画像は, 左のピンホールを通した①と②, 右のピンホールを通した②と③の被写体の重畳画像となり，被写体②の各ピンホール画像が含まれるため，ネットワークは視差情報を取り出すことができる．また，この処理は，入力パッチサイズを拡大する場合と比較し，ネットワークの入力データが小さく，さらに入力画像の線形変換で実現できるため，演算量が小さくなり，実用上でも有効である．



第5図　パッチ画像とネットワーク入力画像
Fig. 5　Patch images and input images

3. シミュレーション結果

提案法の有効性を確認するため，シミュレーション実験を行った．画素数の等しいピンホール画像とマルチピンホール画像を入力画像としたDCNNを個別に学習した後，奥行マップの推定性能を比較した．

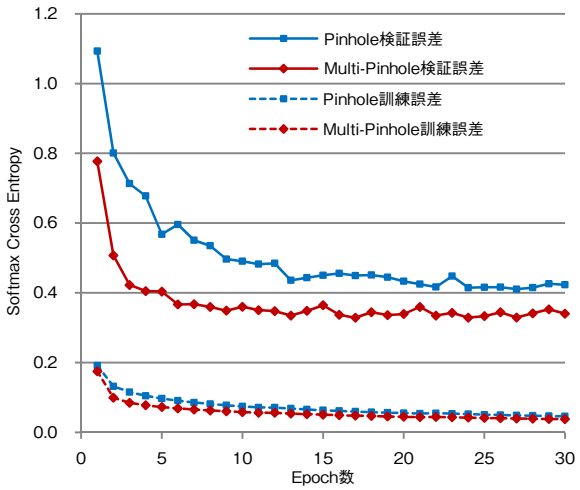
シミュレーション実験には，KITTI Object Detection Evaluation 2012[9] を利用した．このデータセットに含まれるステレオ画像，レーザスキャナによって取得したDepthデータ，カメラキャリブレーションデータを利用して，ピンホール数2のマルチピンホール画像を合成した．学習・推定対象である奥行マップは，Depthデータを入力画像と同じ画像数に補間拡大し，8階調（0 m～5 m, 5 m～10 m, …, 30 m～35 m, 35 m以上）に量子化して作成した．しかし，Depthデータはレーザスキャナによって取得されているため，一部の領域で欠損が生じている．そこで，補間拡大したDepthデータがパッチ内全画素の2割以上欠落したパッチは，学習・評価対象から削除した．利用したパッチ数は，学習データとしては1950 枚の画像から切り出した75832パッチ，評価データとしては学習データと異なる50 枚の画像から切り出した1963パッチである．

第6図に各手法での学習曲線と評価結果を示す．この図において，横軸は学習のEPOCH数，縦軸は8階調の奥行マップのSoftmax Cross Entropyを示している．また，破線は訓練誤差，実線は検証誤差であり，青線は入力画像にピンホール画像を用いたネットワークの評価値，赤線は提案法であるマルチピンホール画像を用いた評価値を示している．

検証誤差に着目すると，提案法であるマルチピンホール画像を利用したネットワークはピンホール画像のものと比較し，より高速に，より小さな評価値に収束している．この結果から，マルチピンホールカメラを圧縮入力デバイスに利用するコグニティブセンシングは，ピンホール画像を利用する従来の識別システムと比較し，入力データ量を増加させずに，より短い学習時間で，より高精度の奥行推定を実現できることが確認できた．

第7図に各入力画像に対する奥行マップ推定結果の一例を示す．**第7図**（a）はピンホール画像，**第7図**（b）は正解奥行マップ，**第7図**（c），（d）はピンホール画像およびマルチピンホール画像からの奥行マップ推定結果，**第7図**（e），（f）は推定した奥行マップの精度を評価するために，各推定結果と正解奥行マップで差分がある領域を黒，差分がない領域を白で示した図である．**第7図**（e），（f）より，画面の右上の領域で，提案法であるマルチピンホール画像の効果が確認できる．この領域は路側の植物領域であり，画像は距離によって大きな違いが生じない．そのため，テクスチャ情報だけでの奥行推定は困難である．一方，マルチピンホール画像ではテクスチャ情報に加え，視差情報であるステレオ画像の多重像を利用することで，より高精度の奥行マップ推定が可能である．

第2表に，ピンホール画像からの奥行推定画像と正解画像，マルチピンホール画像からの奥行推定画像と正解画像との平均絶対誤差を示す．提案手法であるマルチピンホール画像からの奥行マップ推定は，ピンホール画像からの推定と比較し，推定誤差を17 %低減した．



第6図　シミュレーション結果
Fig. 6　Simulation results

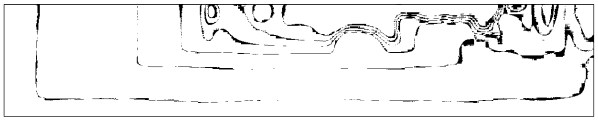

(a) ピンホール画像


(b) 正解奥行マップ


(c) ピンホール画像からの奥行マップ推定結果


(d) マルチピンホール画像からの奥行マップ推定結果


(e) ピンホール画像の推定誤差（黒領域が誤差発生画素）


(f) マルチピンホール画像の推定誤差（黒領域が誤差発生画素）

第7図　奥行マップ推定結果
Fig.7　Depth estimation results

第2表　奥行マップ推定の平均絶対値誤差（階調）
Table 2　Mean absolute errors of depth estimation

	訓練誤差	検証誤差
Pinhole Image	0.065	0.163
Multi-Pinhole Image	0.053	0.135

4. まとめ

入力データ量を効率的に圧縮して取得するコグニティブセンシングの取り組みの一例として，マルチピンホール画像からの奥行マップ推定性能を評価した．入力データ量の等しいピンホール画像と比較し，より高精度の奥行情報が取得できることを確認した．今後，識別システムとしてのコグニティブセンシングの有効性を確認するため，マルチピンホール画像を利用した識別処理を実装し，演算量の少なさを確認するとともに，圧縮入力を行わない通常のマルチモーダルセンシングを行うシステムと比較する予定である．

また，ピンホールの数と配置に関しても，検討をする予定である．本稿では横方向にピンホールを2個設置した場合のみを説明したが，縦方向にもピンホールを増やすことで，縦方向の視差情報も利用できるため，より識別性能を向上できると考えられる．

参考文献

[1] A. Eitel et al., “Multimodal Deep Learning for Robust RGB-D Object Recognition,” IROS, Hamburg, Sept.-Oct., 2015.

[2] D. L. Donoho, “Compressive Sensing,” IEEE Transactions on Information Theory, vol. 52, no. 4, pp. 1289-1306, 2006.

[3] S. Sato et al., “Compressive Color Sensing Using Random Complementary Color Filter Array,” IAPR Conf. on Machine Vision and Application (MVA), Nagoya, pp. 30-33, May, 2017.

[4] M. S. Asif et al., “FlatCam: Replacing Lenses with Masks and Computation,” ICCV, Santiago, pp.12-15, Dec. 2015.

[5] C. Hazirbas et al., “Deep Depth From Focus,” In ArXiv preprint arXiv, 1704.01085, Nov. 2017.

[6] K. Simonyan et al., “Very Deep Convolutional Networks for Large-Scale Image Recognition,” ICLR, San Diego, May 2015.

[7] S. Ioffe et al., “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” ICML, Lille, July 2015.

[8] K. He et al., “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification,” ICCV, Santiago, pp.1026-1034, Dec. 2015.

[9] A. Geiger et al., “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite,” Rhode Island, CVPR, Jun. 2012.

執筆者紹介

佐藤　智　　Satoshi Sato
ビジネスイノベーション本部
AIソリューションセンター
AI Solutions Center, Business Innovation Div.

ポンサク ラサン　Pongsak Lasang
パナソニックR&Dセンター シンガポール
Panasonic R&D Center Singapore
Ph. D

シェン メイシェン　Shen Shengmei, Jane
パナソニックR&Dセンター シンガポール
Panasonic R&D Center Singapore

若井　信彦　　Nobuhiko Wakai
ビジネスイノベーション本部
AIソリューションセンター
AI Solutions Center, Business Innovation Div.
博士（科学）

登　一生　　Kunio Nobori
イノベーション戦略室
Innovation Strategy Office

吾妻　健夫　　Takeo Azuma
ビジネスイノベーション本部
AIソリューションセンター
AI Solutions Center, Business Innovation Div.
博士（工学）