Deep Learningを用いたデジタルカメラのシーン認識技術

Scene Recognition for Digital Cameras with Deep Learning Technology

疋田 聡*
Satoshi HIKIDA

要　旨

近年デジタルカメラにおいて，シーンに適した露出，フォーカス，ホワイトバランスなどの制御により高画質な画像を撮影する機能が望まれるようになってきている．しかし，従来の当社のデジタルカメラでは画像のシーン認識をする技術が限定的であった．そこで，最近有望な技術として非常に注目されているDeep Learning技術をデジタルカメラのシーン認識に利用することで，当社の従来のシーン認識よりも認識率を大幅に向上させ，認識対象も拡大することができる技術を開発した．また，1つの画像で複数のカテゴリが正解となる非排他的カテゴリ認識への対応として，Deep Learningの畳み込み処理のカテゴリ間共通化技術を考案し，単純に排他カテゴリ認識を並列に並べるよりも処理量とメモリ量を抑えることができた．

ABSTRACT

In recent years, controlling exposure, focus, and white balance so that they are suitable for scenes on digital cameras have been required, but the recognition categories of our scene recognition technology was determinative. Therefore, we have used deep learning technology for scene recognition, and it is possible to improve the recognition rate substantially by more than in the past and expand the recognition categories. We have developed a shared convolutional neural network method to save memory and increase throughput by more than when using a simple way.

* リコーICT研究所 システム研究センター
System Research & Development Center, Ricoh Institute of Information and Communication Technology

1. 背景と目的

近年デジタルカメラにおいて，シーンに適した露出，フォーカス，ホワイトバランスなどの制御により高画質な画像を撮影する機能が望まれるようになってきている．例えば，Fig. 1のように画像のシーンを認識できれば，デジタルカメラでシーンに適した画像処理や撮影制御に利用することで，様々なシーンにおいて高画質化を実現することができる．

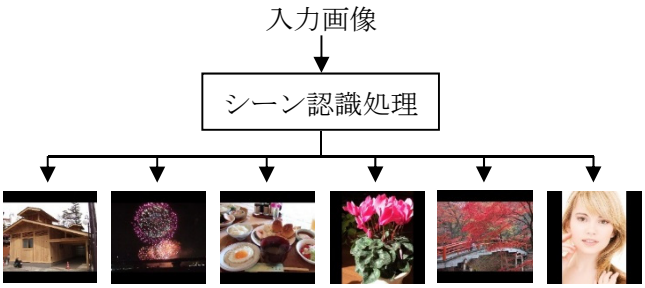

Fig. 1　Example of scene recognition.

しかしながら，当社の従来のデジタルカメラでは，「夜景」のシーンを認識する機能はあったが，その他のシーンには対応しておらず，また，認識率も十分とはいえなかった．

そこで，最近注目を集めているDeep Learning技術[1-3]を適用することで従来よりも認識率を向上させ，認識対象も拡大することを目的とした技術開発を行った．

Deep Learning技術は深い階層型ネットワーク構造を利用した学習方式の総称であり，最近（主に2006年以降），画像認識[2,4,5]，音声認識[6]，データマイニングなど様々な分野において，従来技術よりも圧倒的に良い結果が得られたため注目されている技術である．画像認識においては，特に2012年の一般画像認識コンペティション "IMAGENET Large Scale Visual Recognition Challenge 2012 (ILSVRC2012)" で2位以下に大差をつけて優勝した辺りから注目度が急激に上がっている[4]．

画像認識分野で利用されるDeep Learning技術としてはConvolutional Neural Networksがよく用いられるので，その例をFig. 2に示す．左側から入力された画像データが，Convolution処理とPooling処理を繰り返しながら，段々と低次特徴から高次特徴が用いられるようになっている．

このDeep Learning技術をシーン認識に利用し，例えば，シーン認識の結果を用いてシーンに適した露出，フォーカス，ホワイトバランスなどの制御をリアルタイムに行うことで，様々なシーンに適した高画質化を実現することができるようになる．

2. Deep Learningの畳み込み処理のカテゴリ間共通化

2-1　Deep Learningによる非排他的カテゴリ認識

画像認識コンペティションなどで用いられているDeep Learning技術を用いた画像認識では，1つの画像に1つのカテゴリが正解となる排他的なカテゴリ認識が一般的である[4]．排他的なカテゴリ認識では，ある画像のシーンが「料理」や「花」などの認識可能なカテゴリ中のどれか1つのカテゴリとして認識される．

しかし，デジタルカメラ等におけるシーン認識では，Fig. 5のように「料理」の画像の中に「花」が写っている場合も多くあり，学習データにそのような画像が入っていると，「花」の認識能力が下がり，例えば「花」の画像が「花」と認識されず，デジタルカメラで「花」に適した画像処理や撮影制御が行われなくなるという問題が発生する可能性がある．また，このような認識率の低下を防ぐために，複数のカテゴリが混在している画像を学習データから取り除いておくことも考えられるが，学習データは数万枚と大量であり，その中から複数のカテゴリが混在しているものを予め取り除くには多くの労力がかかってしまう．

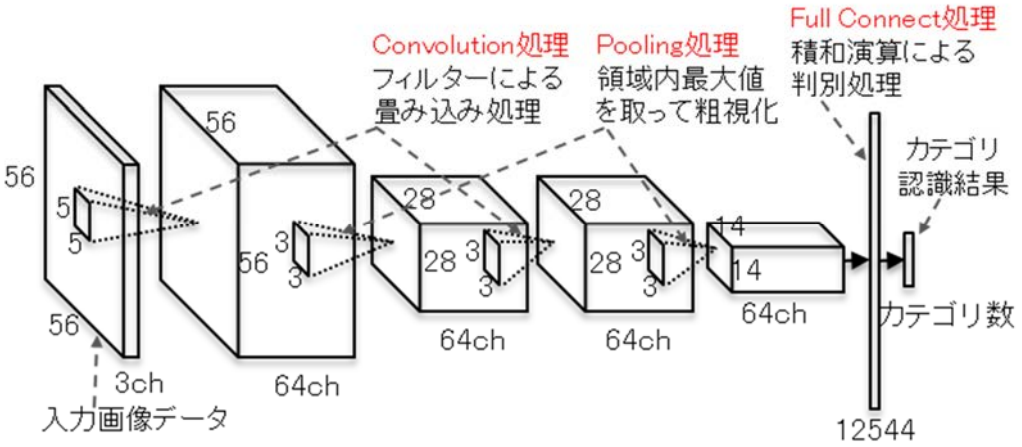

Fig. 2　Example of convolutional neural networks for scene recognition.

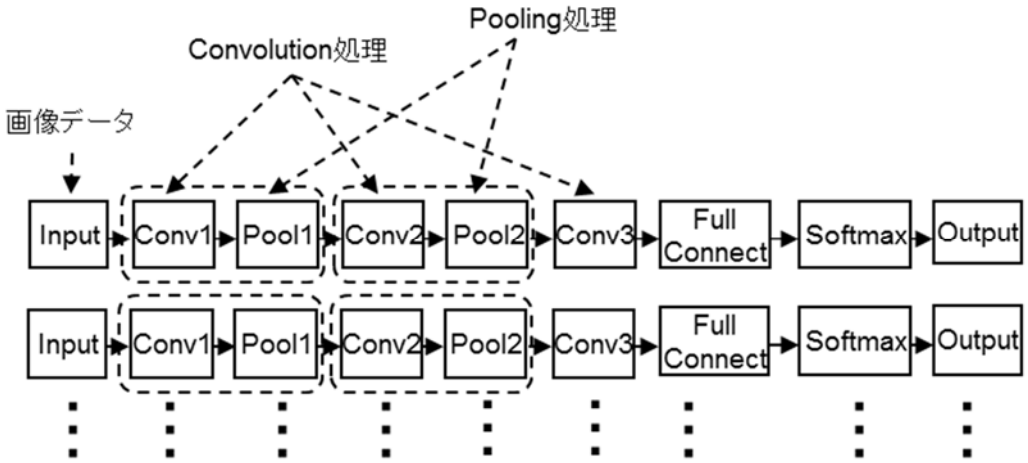

Fig. 3　Example of nonexclusive Deep Learning.

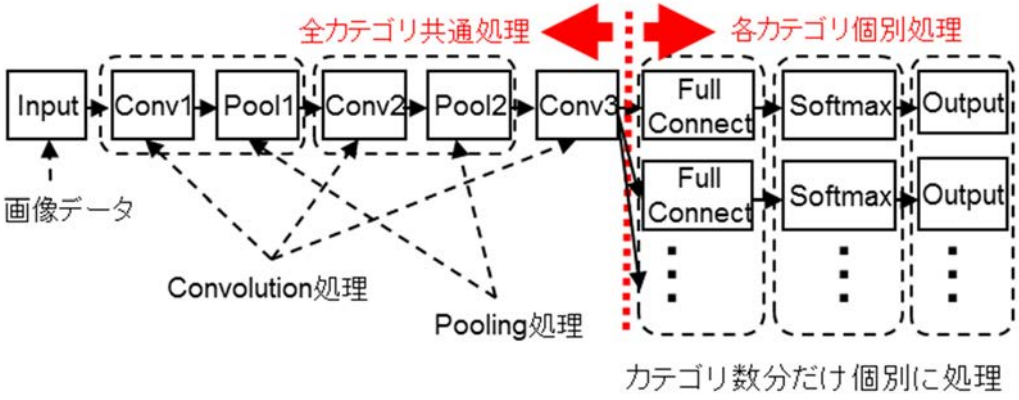

Fig. 4　Example of shared convolutional neural networks.

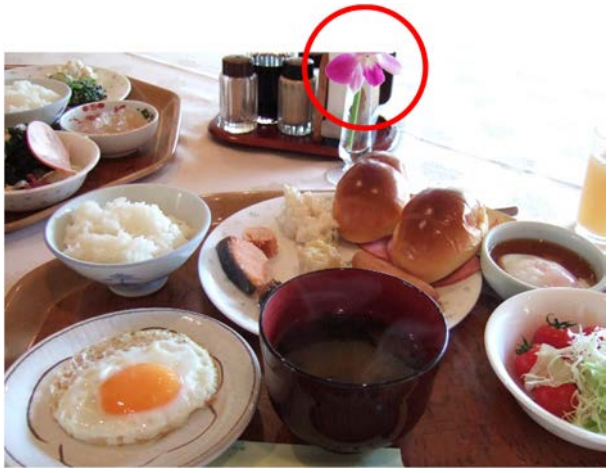

Fig. 5　Example of multi category image.

そこで，複数のカテゴリの混在への対応として，排他的カテゴリ認識ではなく非排他的カテゴリ認識を用いることにした．非排他的なカテゴリ認識は，1つの画像に対して複数のカテゴリが正解となるものであり，例えばFig. 5では「料理」と「花」の両方が正解であってもよいような認識方法である．このようなDeep Learningによる非排他的カテゴリ認識は，例えばFig. 3のように「料理」と「料理以外」の2クラスの排他カテゴリ認識と，「花」と「花以外」の2クラスの排他カテゴリ認識などとを並列に実装し，「料理」「花」などのそれぞれのカテゴリの出力値によって判断することで実現することができる．このような構成でのDeep Learningによる非排他的なカテゴリ認識では，「料理」と「料理以外」，「花」と「花以外」というように別々のDeep Learningのネットワークで学習するので，学習データの中にFig. 5のような画像が混ざっていても，認識率が下がるような副作用が発生しない．

ただし，Deep Learningによる非排他的カテゴリ認識を行うために単純にFig. 3の構成を用いると，Deep Learningのネットワーク分量が排他的カテゴリ認識のカテゴリ数倍になってしまい，必要となる処理量とメモリ量が大きくなるので，今回の開発では以下に述べるDeep Learningの畳み込み処理のカテゴリ間共通化技術を考案し，処理量とメモリ量を抑えることにした．

なお，Deep Learningによる非排他的カテゴリ認識の方法としては，Deep Learningの畳み込み処理のカテゴリ間共通化技術以外にも，画像中からのオブジェクト抽出の技術[7-11]もあるが，計算量が多くなり，ターゲットとしているデジタルカメラ上ではGPUも使用できなかったため，今回は用いなかった．

また，Deep Learningのメモリを削減する技術[12]も提案されているが，その技術ではFig. 4でいうところの各カテゴリ個別処理のFull Connect部分の削減効果は高いが，全カテゴリ共通処理のConvolution部分の削減効果は1/3程度とそれほどでもなく，今回の開発では全カテゴリ共通処理部分のメモリ量が各カテゴリ個別処理よりもはるかに大きかったので，Deep Learningの畳み込み処理のカテゴリ間共通化技術を用いた．

2-2　Deep Learningの畳み込み処理のカテゴリ間共通化

Deep Learningによる非排他的カテゴリ認識を行うために単純にFig. 3の構成を用いると，Deep Learningのネットワーク分量が排他的カテゴリ認識のカテゴリ数倍になってしまい，必要となる処理量とメモリ量が大きくなるという問題を解決するため，Fig. 4のようにDeep Learningによる排他的カテゴリ認識と非排他的カテゴリ認識のDeep Learningネットワークをカテゴリ共通処理とカテゴリ個別処理とに途中から分離し，畳み込み処理をカテゴリ間で共通化した．

この技術ではFig. 3とFig. 4を比較すると分かるように，全カテゴリ共通処理の部分がカテゴリ数に関わらず1個で済むため，全カテゴリ共通処理に必要となる処理量とメモリ量がFig. 4のカテゴリ数分の1となっている．Deep Learningによるシーン認識処理では，畳み込み処理を行っている全カテゴリ共通処理の方が各カテゴリ個別処理の部分より処理量が大きいので，この部分が節約できると効果が高くなっている．参考に今回の実装での全カテゴリ共通処理と各カテゴリ個別処理の処理時間の比率を示すと，

全カテゴリ共通処理部分が34倍の処理時間なのでこの部分をカテゴリ数分の1にできると効果が大きい．同様に，使用メモリ量の比率を示すと，全カテゴリ共通処理部分が16.4倍のメモリ量なのでこの部分をカテゴリ数分の1にできると効果が大きい．

また，各カテゴリ個別処理の部分では，シーン認識結果を非排他的カテゴリ認識と同様に「料理」と「料理以外」，「花」と「花以外」というように別々に出力することができるようになっており，複数のカテゴリの混在へも対応できている．

2-3　カテゴリ間共通化の学習時の処理

Deep Learningの畳み込み処理のカテゴリ間共通化によるシーン認識方式では，全カテゴリ共通処理とカテゴリ個別処理が混ざっているので単純には学習することができない．そこで，Fig. 6，Fig. 7に示すようにカテゴリごとに学習データをミニバッチ（学習時にニューラルネットワークのパラメータを更新するデータの塊．例えば画像128枚など）で与え，ミニバッチごとにカテゴリを切り替えていくことを繰り返して学習処理を行うという方式で学習を行った．（なお，Fig. 6, 7ではFig. 4を縦にして，Pooling処理の層は省略し，Full Connect層は "FC" と略している．また，学習時にはSoftmax層は使用しない．）

例えば，Fig. 6のカテゴリ1の学習フェーズでは，カテゴリ1の学習データをミニバッチで与え，全カテゴリ共通部分のDeep Leaningネットワークパラメータは通常の更新処理を行い，カテゴリ個別処理部分についてはカテゴリ1の処理だけを行い，その他のカテゴリの処理は休止させておく．次にFig. 7のカテゴリ2の学習フェーズになったときは，カテゴリ2の学習データをミニバッチで与え，全カテゴリ共通部分のDeep Leaningネットワークパラメータは通常の更新処理を行い，カテゴリ個別処理部分についてはカテゴリ2の処理だけを行い，その他のカテゴリの処理は休止させておく．このような処理をミニバッチごとに各カテゴリについて行い，最後のカテゴリまで処理したらカテゴリ1に戻ってさらに繰り返し学習処理を進めることで学習を行った．

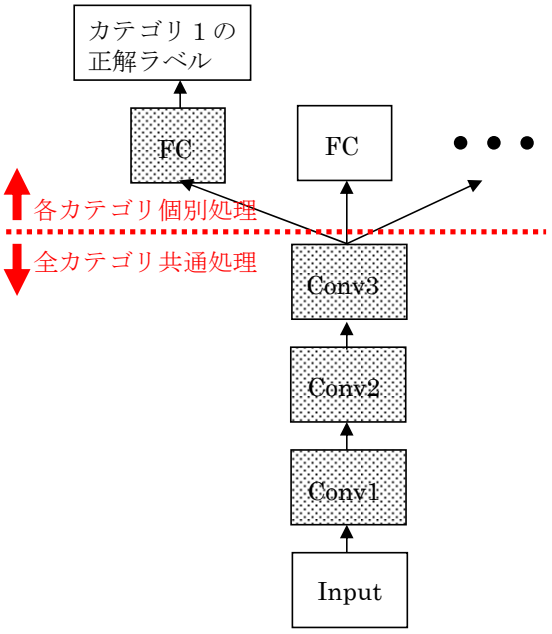

Fig. 6　Example of learning for category 1.

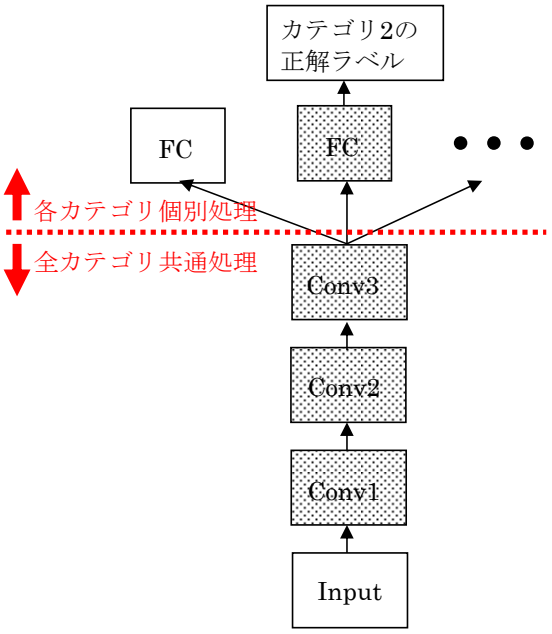

Fig. 7　Example of learning for category 2.

3. 実験結果

従来の当社のデジタルカメラによる画像のシーン認識よりも認識率が向上しているかを確認するため，従来方式と新方式の認識率の比較実験を行った．認識対象のカテゴリは従来方式で対応している「夜景」のカテゴリを用いた．その結果，Fig. 8に示すように従来方式の85.0%から97.0%と大幅な向上が確認できた．

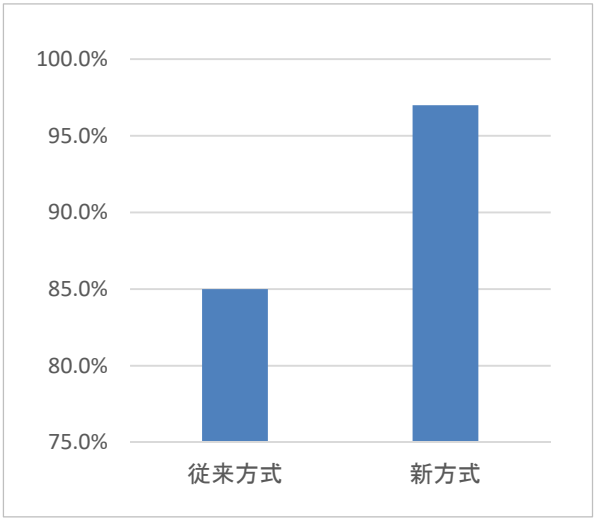

Fig. 8　Result of scene recognition.

また，認識対象の拡大の全てではないが一部の例として，「夜景」「料理」「花」「紅葉」「自動車」「夕焼け，朝焼け」の6カテゴリについて新方式での認識率をTable 1に示しておく．

Table 1　Detail of result of scene recognition.

カテゴリ	従来方式	新方式
夜景	85.0%	97.0%
料理	非対応	96.5%
花	非対応	96.0%
紅葉	非対応	94.5%
自動車	非対応	94.5%
夕焼け，朝焼け	非対応	94.5%

この結果では，拡大した認識対象に対しても，平均すると95%以上の認識率が得られている．

このように認識率が上がったことにより，デジタルカメラで様々なシーンに適した画像処理や撮影制御が行えるようになり，高画質化を実現できるようになった．

4. おわりに

以上のように，従来の当社のデジタルカメラでは画像のシーン認識をする技術が限定的であったが，Deep Learning技術を適用することで当社の従来のシーン認識よりも認識率を大幅に向上させ，認識対象も拡大することができる技術を開発した．

今後は，更なる効率化や認識率の向上，デジタルカメラ以外への応用なども検討していく予定である．

参考文献

1) Y. Bengio, A. Courville, P. Vincent: Representation Learning: A Review and New Perspectives, eprint arXiv:1206.5538 (2012).
2) 岡谷貴之: 連載解説「Deep Learning (深層学習)」第4回，人工知能学会誌, Vol. 28, No. 6, pp. 962-974 (2013).
3) 岡谷貴之: 深層学習，講談社 (2015).
4) A. Krizhevsky, I. Sutskever, G. E. Hinton: ImageNet Classification with Deep Convolutional Neural Networks, NIPS (2012).
5) K. He et al.: Deep Residual Learning for Image Recognition, eprint arXiv:1512.03385 (2015).
6) G. E. Hinton et al.: Deep Neural Networks for Acoustic Modeling in Speech Recognition, *IEEE SP Magazine* (2012).
7) R. Girshick et al.: Rich feature hierarchies for accurate object detection and semantic segmentation, eprint arXiv:1311.2524 (2013).

8) D. Erhan et al.: Scalable Object Detection using Deep Neural Networks, eprint arXiv:1312.2249 (2013).
9) C. Szegedy et al.: Scalable, High-Quality Object Detection, eprint arXiv:1412.1441 (2014).
10) R. Girshick: Fast R-CNN. arXiv:1504.08083 (2015).
11) S. Ren et al.: Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, eprint arXiv:1506.01497 (2015).
12) S. Han, H. Mao, W. J. Dally: Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding, eprint arXiv:1510.00149 (2015).