Vol.2018-CVIM-213 No.37
2018/9/21

照度ロバストな画像特徴量を用いたセマンティックセグメンテーション

田口賢佑[†1]　菅原俊[†1]　渡邊正人[†1]　船津陽平[†1]

概要：近年、深層学習をベースとした一般交通シーンでの画像認識技術が研究されている。その中でも、セマンティックセグメンテーションはピクセル単位でクラスを推定する技術であり、自動運転等の高度なアプリケーションへの利用が期待されている。一般交通シーンは様々な被写体が存在し、さらに時間帯や天候、シーンによって照度変動が大きいため、大量の教師データが必要となり、データ収集コストが課題となっている。本稿では、このような課題に対応するため、セマンティックセグメンテーションネットワークへの入力データを局所領域輝度関係性を表現するピクセルベース画像特徴量とする手法を提案する。局所領域輝度関係性により照度不変特性を獲得することで、認識精度が教師データの照度バリエーションに依存しないことを狙う。検証として一般道路を撮影した様々な照度を含むデータを用い、学習データとテストデータの照度が異なる場合の認識精度において、提案手法は従来手法（RGB データ入力）と比較して改善されていることを確認した。

キーワード：Semantic Segmentation、Convolutional Neural Network、Robustness、画像認識、局所特徴量

1. はじめに

近年深層学習を用いた画像認識の分野で Convolutional Neural Network(CNN)[1]が非常に高い精度を達成している。CNN は物体認識だけではなく、セマンティックセグメンテーションにも応用されている。セマンティックセグメンテーションはピクセル単位でクラスを推定する技術であり、シーンの理解や走行可能領域の判断が必要な自動運転等の高度なアプリケーションへの利用が期待されている。そのため車載向けの大規模データセットも公開されており、車載シーンを対象としたセマンティックセグメンテーションの研究が行われている。

自動車は外界を走行するため、車載で用いられる画像認識は走行時の環境変化に対してロバスト性が必要である。走行時の環境変化は地域性から天候変化、時間帯による照度変化等の様々な要因が存在するが、日常的に起こりうる事例として昼/夜に代表される照度変化がある。昼シーンと比較すると夜シーンのような低照度環境ではカメラ露出が不足しやすい。従って夜シーンの画像データには、カメラ露出不足に起因するコントラスト低下、イメージセンサ高感度化によるランダムノイズ増加、シャッタースピード低下による被写体ボケ、発光体の出現などの変化が生じる。図 1 に t-SNE[9]を用いた RGB データの昼/夜の分布を示す。様々なノイズ要因により昼/夜のデータ間距離が生じている。よって、昼/夜の照度違いに対応したセグメンテーションを実現するためには、昼/夜で取り得るデータ空間を学習する必要がある。

RGB データは昼/夜のデータ間距離が大きく、そのデータ空間を網羅するためには昼/夜どちらかのシーンではなく両シーンのデータを学習する必要がある。そのため、学習データ数が増加しデータ収集及び教師データ作成に多大なコストが必要である。特に夜データは昼データと比較し、被写体の出現頻度が低いため収集コストが大きい。

一方、画像処理等で昼/夜データ間距離を近づけデータ空間上のバラつきを抑えることができれば、学習データの照度バリエーションが乏しい場合でも昼/夜のデータ空間を網羅することができる。

そこで、本研究では照度変化が原因で生じる入力データばらつきを抑える画像特徴抽出層を備えたセマンティックセグメンテーション手法を提案する。画像特徴抽出層出力を CNN への入力データとすることで、学習データの照度バリエーションが少ない場合でも照度変化への対応が期待できる。

提案手法の有効性を検証するために、昼シーンと夜シーンを含むセマンティックセグメンテーション用のデータセットを作成し評価実験に用いた。

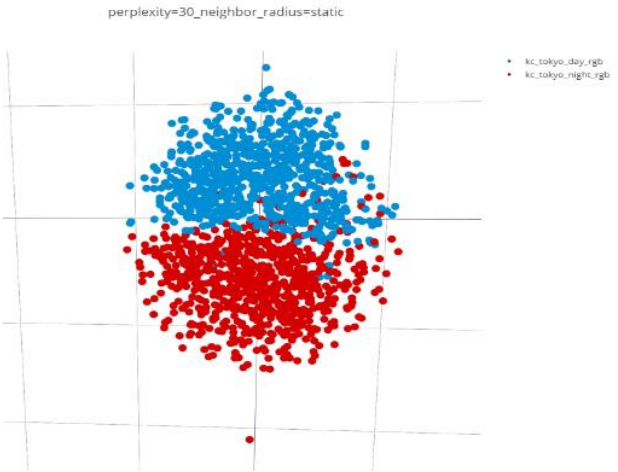

図 1　昼/夜データセットにおける昼/夜画像分布

†1 京セラ株式会社　研究開発本部　システム研究開発統括部
ソフトウェア研究所　画像処理研究部

ⓒ 2018 Information Processing Society of Japan

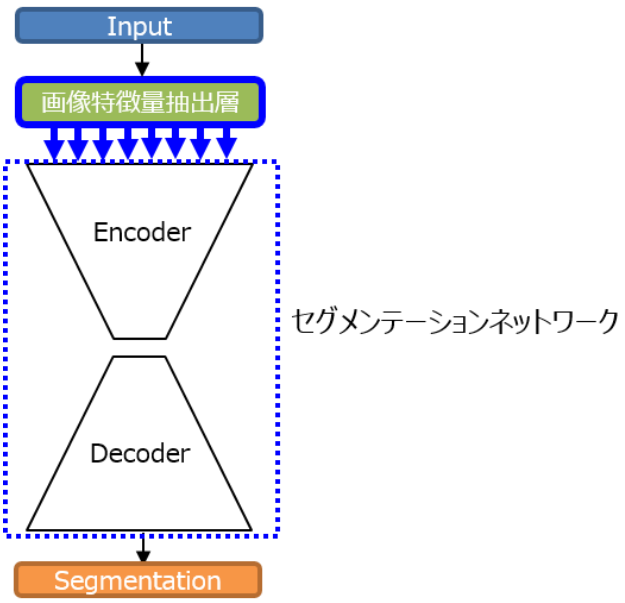

図 2　提案手法のネットワーク構成

2. 関連研究

CNN をベースとしたセマンティックセグメンテーションとして Fully Convolutional Network(FCN)[2]がある。FCN は CNN の全結合層を 1x1 畳み込み層に置き換え、出力層をピクセルワイズなソフトマックスレイヤーに置き換えたネットワークである。SegNet[3]は FCN で用いられる 1x1 畳み込み層を削除した Encoder-Decoder 構造を採用している。Encoder 部分は入力画像の特徴抽出を行い、Decoder は Encoder で抽出した特徴マップ画像を入力画像サイズまで戻す。

照度変化に頑健な画像特徴量として Local Binary Pattern(LBP)[4]がある。LBP は中心画素とその周辺画素の輝度差分を算出し、その符号によって二値を割り当てるため局所領域の輝度大小関係を記述可能である。照明変動等で画像全体の輝度が変化しても被写体が持つ局所領域の輝度関係性は不変なため、照度変化に頑健な出力となる。また、LBP は空間フィルタリング処理で算出されるためピクセルワイズな出力となる。Histogram of Oriented Gradients descriptor（HOG）[5]は局所領域内の輝度勾配の強度と方向をヒストグラム化した特徴量であり、大まかな物体形状を表現することができる。HOG は対象領域を小さなセルで分け、複数セルにより正規化しているため照明変動に対して頑健である。

車載環境の照度変化に対するセマンティックセグメンテーションの研究として、Haltakov ら[10]は昼/夜シーンの交通信号を対象としたセマンティックセグメンテーションを実現している。交通信号及び周囲の構造物は昼/夜で見えが大きく異なる。そのため、画像からの特徴抽出方法は昼/夜で共通ではあるが、分類器の学習は両シーンの学習データを用いている。

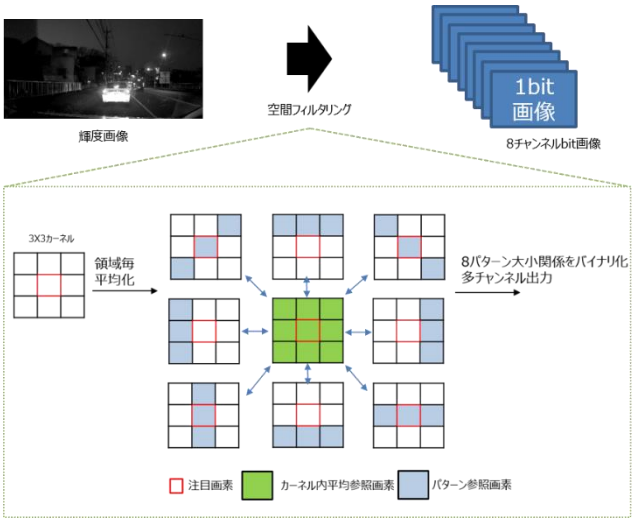

図 3　Local Area Pattern のアルゴリズム

照度変化等の環境変化を画像変換で実現する手法として Ming-Yu Liu ら[11]は、対象データと異なるドメインのデータセットを用い、異なるドメインへの画像変換を実現している。そのため、例えば教師データ付きの昼シーン画像と教師データ無しの夜シーン画像を用いて、教師データ付き夜シーン画像を生成することができるため、夜シーン画像のための教師データ生成コストを削減可能である。

[10]、[11]の手法は異なるドメインのデータを収集する必要があるため、データ収集コストに課題がある。よって、本研究では昼/夜の照度変化で生じるデータバラつきを学習データに依存しない手法で解決することを狙う。

3. 提案手法

本研究では、照度変化による入力データバラつきを抑える画像特徴抽出層を導入し、セグメンテーションネットワークへの入力とする構成を提案する。図 2 にネットワーク構成を示す。画像特徴抽出層を予め設計し固定パラメータとすることで、学習データに依存せず明示的にデータのバラつきを抑えることを狙う。以下に本提案手法の詳細を述べる。

3.1 画像特徴抽出層

画像特徴抽出層は入力データのバラつきを抑えつつ、CNN で扱うことができるデータを出力することが必要である。CNN がデータを扱うためには画像の幾何学的な連続性を維持することが必要である。CNN は画像データ入力を想定したアークテクチャーであり、近隣画素は強い関係性を持つ性質を利用した畳み込み構造を採用しているため、入力データの幾何学的な連続性は重要である。

本研究では画像特徴抽出層に用いるアルゴリズムとして 画像コントラスト変化とランダムノイズ量変化に対するロバスト性に着目し、LBP をベースにした画像特徴量 Local Area Pattern(LAP)を提案する。LAP は図 3 に示すように 3x3 カーネルで、カーネル内輝度平均値と 3 画素で構成

ⓒ 2018 Information Processing Society of Japan

される 8 パターン領域の輝度平均値の大小関係性を記述する。LAP は領域同士の大小関係を取るため、照明変動による画像コントラスト変化の影響を受けづらい。また、従来技術の LBP はピクセル間比較であり、比較対象ピクセルにノイズが乗ると被写体本来のピクセル間関係性が壊れる可能性がある。一方、LAP は比較対象を領域とすることで、

(1) ピクセル平均化によるノイズ低減効果。

(2) 複数画素の比較によるレベル方向へのダイナミックレンジ拡張。

が期待できる。LAP はピクセル単位ではなく領域単位で処理するため実質的な空間解像度では LBP に劣るものの、上記特徴を持つことでロバスト性向上を実現する。

また、LAP は LBP 同様にピクセルワイズな画像特徴量であり、入力画像の幾何学的な連続性は維持されているため、CNN で扱うことが可能である。

3.2 セグメンテーションネットワーク構成

セマンティックセグメンテーションを実施するネットワークは SqueezeNet[6] 構造をベースにした Encoder-Decoder 構成となっている。学習収束を早めるために各畳み込み計算後に Batch Normalization[7]を実施している。LAP を用いた画像特徴抽出層出力は 8 チャンネル画像であり、8 チャンネルを畳み込み層に入力しマージすることで適切なチャンネルフュージョンを実現することができる。

4. 評価実験

提案手法の有効性を確認するため、車載シーンを対象としたセマンティックセグメンテーション用データセットを用いて評価実験を実施する。提案手法と画像特徴抽出層無しの構成、画像特徴抽出層に別アルゴリズムを用いた構成を比較対象とした。画像特徴抽出層の別アルゴリズムはピクセルワイズで画像特徴抽出可能なラプラシアンフィルタ（Edge）と LBP とした。以下にデータセット、認識精度の評価方法、評価結果を述べる。

4.1 データセット

評価データには、Cityscapes[8]データセット及び独自で作成した昼/夜データセットを用いる。

Cityscapes データセットは車両に取り付けられたフロントカメラで撮影されており、ドイツ国内を中心としたヨーロッパ 50 都市で日中に撮影されている。アノテーションは詳細にアノテーションされた Fine annotation とラフにアノテーションされた Coarse annotation が用意されている。本実験では Fine annotation 画像を使用し、学習用として 2975 枚、評価用として 500 枚を用いる。Cityscapes データセットは 30 種類のクラス分けがされているが、本研究では車載シーンの画像認識として特に必要とされる路面、車両、人物、二輪車、その他の 5 クラスを対象とし評価を実施する。

図 4 昼/夜データセットの画像例

表 1 昼シーンの認識精度[%]

手法	Cityscapes	昼/夜データセット
画像特徴抽出層無し	71.91	58.02
LBP	71.98	55.78
Edge	72.91	58.75
LAP	71.39	56.25

昼/夜データセットは車両のフロントガラスに装着したドライブレコーダーで撮影されている。図 4 に昼/夜データセットの画像例を示す。撮影地域は日本国内の 2 地域であり、昼間画像は晴天時の 11～16 時、夜間画像は晴天時の 17 時～22 時に撮影している。画像サイズは 1920 x 1080 であり、本研究では昼シーン学習用に 1100 枚、夜シーン学習用に 1100 枚、昼シーン評価用に 200 枚、夜シーン評価用に 200 枚を用いる。アノテーションは詳細な Fine annotation が付けられており、対象クラスは先に示した 5 クラスである。

4.2 認識精度の評価方法

セマンティックセグメンテーションの評価指標として広く利用されている平均 IoU を算出する。各クラスの IoU は式（1）のように求める。TP は推論結果が正しい画素数、FN は別クラスとして誤判定した画素数、FP は正解が別クラスに属す領域に対して誤判定した画素数である。平均 IoU は各クラスの IoU を平均することで算出する。

$$\mathrm{IoU} = \frac{TP}{TP + FP + FN} \qquad (1)$$

4.3 昼シーンでの評価結果

提案手法は画像特徴抽出層で入力データ変換を行っているため、セグメンテーションに必要な情報が失われている可能性がある。そこで、画像特徴抽出層出力がセグメンテーションタスクに必要な弁別性能を有しているかを検証

ⓒ 2018 Information Processing Society of Japan

表 2　Cityscapes データセットでの認識精度比率[%]

手法	明るさ 1/2 画像/通常	明るさ 1/4 画像/通常	Sigma3 ノイズ/通常	Sigma5 ノイズ/通常	Blur(5x5)/通常	Blur(10x10)/通常
画像特徴抽出層無し	96.66	80.35	98.82	95.489	98.37	88.265
LBP	98.29	87.75	83.34	64.06	84.52	51.10
Edge	97.49	82.99	88.65	74.51	70.00	38.67
LAP	99.94	94.54	93.67	83.31	92.05	61.24

表 3　昼/夜データセットでの認識精度比率[%]

手法	明るさ 1/2 画像/通常	明るさ 1/4 画像/通常	Sigma3 ノイズ/通常	Sigma5 ノイズ/通常	Blur(5x5)/通常	Blur(10x10)/通常
画像特徴抽出層無し	91.85	67.20	98.50	97.31	97.60	89.08
LBP	98.94	95.49	93.64	85.14	94.47	78.51
Edge	93.24	74.04	98.38	95.21	90.31	66.75
LAP	99.45	97.05	96.49	92.28	96.06	85.71

するため、ノイズ要素の少ない昼データにて認識精度比較を行う

Cityscapes データセットと昼/夜データセットを用いた認識精度比較結果を表 1 に示す。昼/夜データセットは学習データと評価データを昼シーンとしている。どちらのデータセットにおいても提案手法は他手法と同等の認識精度を持つことが確認できる。よって、提案手法はセグメンテーションタスクに必要な弁別性を有していると言える。

昼/夜データセットにおいても各手法間で認識精度はほぼ同等であるが、一様に Cityscapes データセットにおける認識結果から悪化している。これは、学習データ数が少ないことやデータセット難易度が影響していると考えられる。

4.4 ノイズシミュレーションによる評価結果

次に提案手法が昼/夜の照度違いで生じるノイズに対するロバスト性を有するかを確認するため、昼シーン評価データに擬似夜間変換処理として明るさ低下処理、ランダムノイズ付与、ブラー効果を実施する。学習データは昼シーンを使用する。昼シーン評価データ認識精度に対する擬似夜間変換画像認識精度の比率を評価する。

Cityscapes データセットでの結果を表 2 に示す。元画像に対し明るさを 1/2 に低下させた場合はどの手法も認識精度を保っているが、1/4 に低下させると画像特徴抽出層無し構成の認識精度比率が 80.35%に対して提案手法は 94.54%であり最も高い。よって提案手法は明るさ不足に対してロバスト性が高いと言える。

ランダムノイズ増加は通常画像に対し平均値 1、分散値 3、5 のガウシアンノイズを付与させる。結果として画像特徴抽出層無し構成の認識精度比率が最も高い。画像特徴抽出層アルゴリズムの中では提案手法の認識精度比率が高く、特に類似アルゴリズムである LBP と比較して約 20%改善しているため、複数画素平均化によるノイズ低減効果が有効と言える。

ブラー効果は通常画像に対しカーネルサイズ 5x5、10x10 の平均フィルタで処理する。結果として画像特徴抽出層無し構成の認識精度比率が最も高い。画像特徴抽出層アルゴリズムは全般に認識精度比率が低下しており、ブラー効果を苦手としている。つまり画像特徴抽出層ではブラー効果に対するロバスト性を獲得しづらいため、学習データによる解決が必要である。例として、データオーギュメンテーションでブラー効果をシミュレートした画像を学習させることは有効と考えられる。

昼/夜データセットでの結果を表 3 に示す。Cityscapes デ

表 4　昼/夜データセットでの認識精度[%]

手法	夜シーン	昼シーン	夜シーン / 昼シーン
画像特徴抽出層無し	40.99	58.02	70.65
LBP	40.95	55.77	73.42
Edge	45.90	58.75	78.13
LAP	43.83	56.25	77.93
LBP（ブラー）	43.46	56.09	77.48
Edge（ブラー）	46.75	58.05	80.52
LAP（ブラー）	46.17	56.71	81.42
画像特徴抽出層無し（両シーン）（目標値）	58.81	59.03	99.63

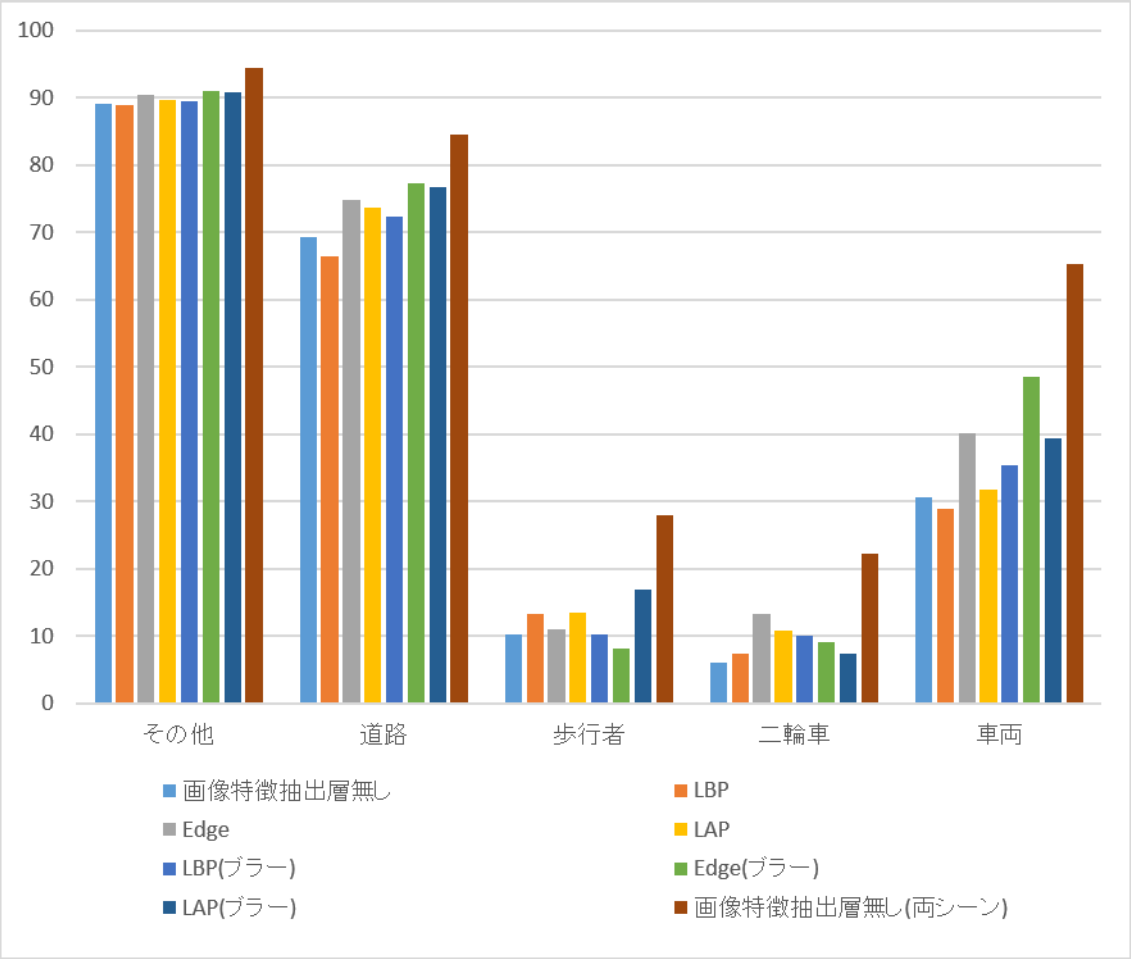

図 5　昼/夜データセットにおける各クラスの IoU 精度

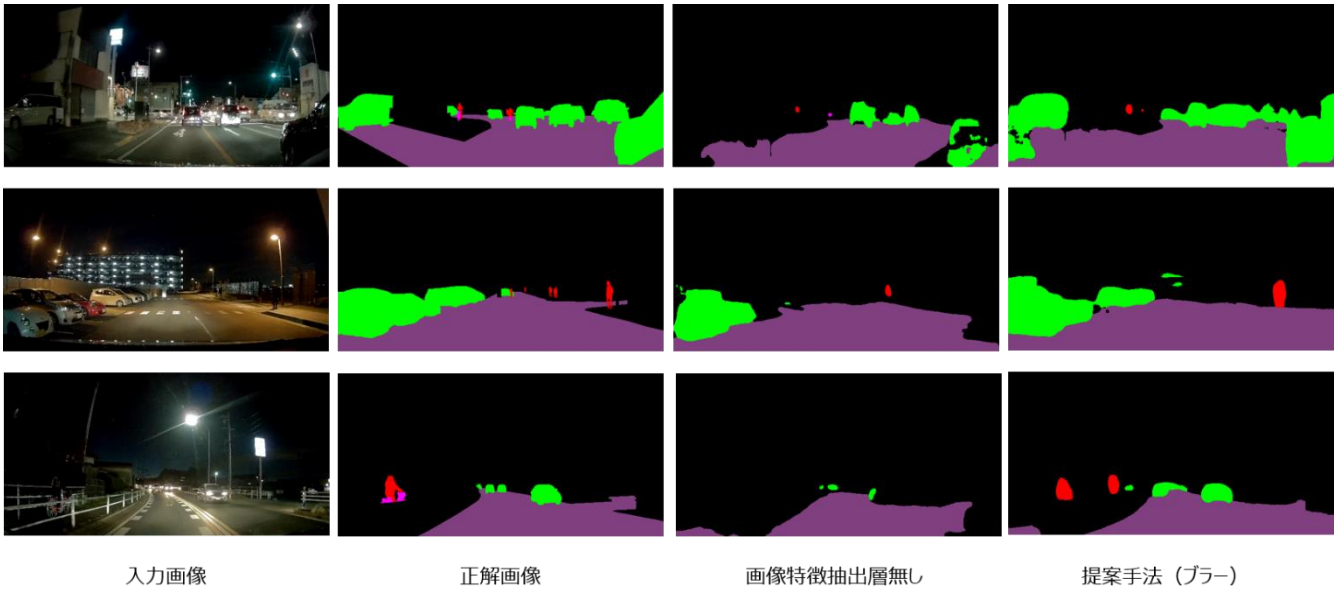

図 6　昼/夜データセットにおける結果例

ータセットでの結果と同様の傾向となっているが、ランダムノイズ増加とブラー効果に対するに認識精度比率が全体的に高い。これは昼/夜データセットは Cityscapes データセットと比較し、カメラ画質としてランダムノイズとブラーが多く含まれているため、加えられたノイズに対しロバスト性を持つと考えられる。

4.5 昼/夜シーンでの評価結果

前項で各ノイズに対するロバスト性を評価したが、実際の昼/夜シーンは各ノイズ要因が複合して発生する。よって昼/夜の照度違いデータに対し認識精度比較を行い、提案手法の照度変化ロバスト性を検証する。

昼/夜データセットを用いて昼シーン評価データ認識精度に対する夜シーン評価データ認識精度の比率を評価する。学習データは昼シーンを用いる。前項での結果から画像特徴抽出層アルゴリズムは明るさ低下、ランダムノイズ増加と比較するとブラー要因に対するロバスト性が低いため、学習データにランダムなブラー処理によるデータオーギュメンテーションを実施したモデルも比較対象に追加する。

表 4 に各画像特徴抽出層構成の認識精度及び認識精度比率を示す。目標値として学習データに昼と夜の両シーンを学習させたモデルの結果を記載している。提案手法の認識精度比率は 77.93%であり、画像特徴抽出層無し構成の 70.65%から大きく向上している。また、提案手法含む画像特徴抽出層構成はデータオーギュメンテーションを実施することで認識精度比率が平均 3%程度向上している。これは画像特徴抽出層によって特徴空間が絞られているため、不足していた特徴を効果的に学習できたと考えられる。データオーギュメンテーションを実施した提案手法の認識精度比率は 81.42%であり、目標値を除くと最も高い。このことから、実際の昼/夜の照度違いにおいて提案手法は照度変化で生じる変化にロバストな手法であると言える。

図 5 に夜シーン評価データのクラス別認識精度を示す。提案手法は歩行者の認識精度が高い。歩行者はヘッドライト範囲外の歩道に存在することが多く他クラスと比較し特に照度が低いため、明るさ不足に頑健な特性を持つ提案手法の認識精度が高いと考えられる。車両クラスは画像特徴抽出層にエッジを用いた場合の精度が高い。これは、車両のようになだらかな部分が支配的な被写体においては、連続値出力となるエッジが有効と言える。

図 6 に提案手法と画像特徴抽出層無し構成でのセグメンテーション画像を示す。全てのクラスにおいて提案手法の方が高精度にセグメンテーションできており、特に低照度領域が多い道路部分が高精度である。

一方、目標値の昼と夜の両シーンを学習させたモデルと比較すると提案手法含む画像特徴抽出層構成の認識精度比率が小さい。画像特徴抽出層では発光体の有無で生じるコンテキスト違いは吸収できないため、その部分が認識精度の差となっていると考えられる。

5. まとめ

本研究では、CNN に局所領域輝度関係性を表現するピクセルベース画像特徴量 LAP を用いた画像特徴抽出層を導入したセマンティックセグメンテーションネットワークを提案した。昼シーンを学習データとし、評価データが昼シーンと夜シーン時の認識精度の比率で照度変化ロバスト性を評価し、提案手法が他手法と比較し照度変化ロバスト性が優れていることを確認した。

一方、提案手法は昼/夜の両データを学習させたモデルの

ⓒ 2018 Information Processing Society of Japan

認識精度に劣る。今後は発光体のように昼/夜でコンテキストが異なる被写体への対応を検討する。また、提案手法は露出不足な暗部にロバスト性を持つため、カメラの特性や撮影の工夫も有効と考えられる。

参考文献

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, “Gradient-Based Learning Applied to Document Recognition," in Proceedings of the IEEE, vol. 86,no. 11, pp.2278-2324, 1998.

[2]J. Long, E. Shelhamer, T. Darrell, “Fully convolutional networks for semantic segmentation” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[3]V. Badrinarayanan, A. Kendall, R. Cipolla,”SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,"arXiv:1511.00561, 2016.

[4]T. Ojala, M. Pietikäinen, and D. Harwood , "Performance evaluation of texture measures with classification based on Kullback discrimination of distributions" 1994

[5]N. Dalal, B. Triggs: Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and Pattern Recognition, pp886- 893,2005

[6]F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J.Dally, and K. Keutzer. SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size.arXiv:1602.07360, 2016.

[7]S. Loffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", arXiv preprint arXiv:1502.03167, 2015.

[8]M. Cordts, M. Omran, S. Ramos, T. Rehfeld,M. Enzweiler, R.Benenson, B. Schiele, "The cityscapes dataset for semantic urban scene understanding", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016.

[9]Van der Maaten, Laurens, and Geoffrey Hinton. ”Visualizing data using t-SNE.” Journal of Machine Learning Research 9.2579-2605 (2008): 85.

[10]V. Haltakov, J. Mayr, and C. Unger, “Semantic Segmentation Based Traffic Light Detection at Day and at Night,” Pattern Recognition.Springer International Publishing,, vol. 9358, pp.446–457, 2015..

[11]M.-Y. Liu, T. Breuel, and J. Kautz.” Unsupervised image-to-image translation networks”. arXiv preprint arXiv:1703.00848, 2017

ⓒ 2018 Information Processing Society of Japan