The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, 2019

物体の重なり情報を用いた単眼深度推定

Monocular depth estimation using overlap information

高嶺 潮 *1 Michiru Takamine　遠藤聡志 *1 Satoshi Endo　Jakub Kolodziejczyk*1　西銘大喜 *1 Taiki Nishime

*1琉球大学工学部情報工学科
Faculty of Engineering Department of Information Engineering

Depth estimation is an important tool for machines to get spatial information. Human is realizing high-accuracy depth estimation by dividing problem area, but it is difficult for machine to calculate depth from a single RGB image. Our goal is to improve the accuracy of monocular depth estimation. From the past research, it has been found that obtaining the information stepwise in the global and local areas is effective for depth estimation. Therefore, In this research, we propose a method to utilize the anteroposterior relationship information of the object. Experimental results showed that the overlap information is useful for depth prediction.

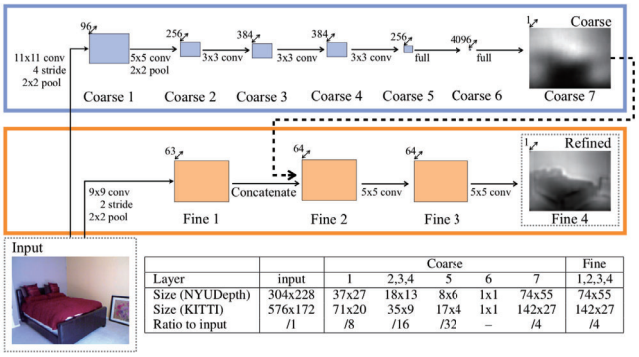
1. はじめに

カメラから被写体までの距離のことを深度と言い、画像から深度を推測する分野を深度推定と呼ぶ。深度推定のうちモノラル画像を入力として扱うものを単眼深度推定と呼ぶ。正確な単眼深度推定にはオブジェクトの実寸の情報が必要となるが、実寸情報はモノラル画像から直接求められない [1]。人間は深度推定に使用できる情報の種類を増やすことで実寸情報を補完しており、具体的には問題領域を分割することで精度の高い深度推定を実現している [2]。問題領域の分割は、取得できる情報に冗長性を持たせ、取得に失敗した真に重要な情報を補う役割を持つ。これを受け、深度以外の情報を RGB 画像から獲得することによって単眼深度推定を改善しようとする試みが幾つか存在する。Semantic ラベルを用いた実験では、ラベルの曖昧性によって学習に悪影響が生じることがわかり、人間の主観によって定義された情報の欠点を浮き彫りにした [3]。対して、深度のエッジ推定を扱った実験は、推定結果の外れ値の削減に大きく貢献している [4]。これらの結果は、数値的に定義可能なオブジェクト情報が、人間が深度推定を行う際に獲得する冗長性の再現に繋がることを示唆している。以上を踏まえて、本研究では、物体の前後関係に着目した情報（以下、これを重なり情報と呼ぶ）を深度推定に活用する手法を提案すると共に、対照実験により重なり情報の有効性を検証する。

2. 先行研究

単眼深度推定には研究者の事前知識と多くの仮定が必要とされ [5]、その弊害として入力画像に大きな制限が設けられた。Eigen ら [4] は CNN で人間の深度推定を模倣しようと試み、大域情報と局所情報を段階的に求めることで、入力画像に制限を設けない深度推定を可能とした。2 つの情報を分けて考える利点は問題領域の分割にあり、これは冗長性の獲得に対応する。Eigen らが使用した Multi-Scale Model:MSM（図 1）は 2 種類のネットワークを持ち、Global Coarse-Scale Network（図 1 青枠）が大域情報を、Local Fine-Scale Network（図 1 赤枠）が局所情報を導出し、2 つの情報を統合して最終的な深度を求める。本論文では MSM を雛形に深度推定モデルを構築していく。

連絡先: e155736@ie.u-ryukyu.ac.jp



Layer	input	Coarse 1	2,3,4	5	6	7	Fine 1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

図 1: MSM[4]

3. 提案手法

3.1 重なり情報の定義

重なり情報とは、物体もしくは空間の前後関係の順番のみに着目した情報と定義する。重なり情報は単一のチャンネルの画像として表現され、カメラに最も近い領域には 0、次に近い領域には 1、以下昇順で pixel ごとに整数値を持っている（以下、これを重なり画像と呼ぶ）。重なり画像 1 つに割り振られる整数値の種類の数を重なり数として定義する。図 2 は重なり画像の例で、近いほど青く、遠いほど赤く pixel が表現されている。重なり情報には複数の形態が考えられるが、本論文では以下の 3 種類の形態を取り扱う。

1. pixel を深度でクラスタリングしクラスタ毎に番号を割り振った重なり情報:Pixel Overlap（図 2(a)）
2. オブジェクト単位の領域毎に番号を割り振った重なり情報:Object Overlap
   (a) Object Overlap のうち、属している深度の区域に応じて番号を割り振った重なり情報:Object Clustering Overlap（図 2(b)）
   (b) Object Overlap のうち、遮っている画面上のオブジェクトの数に応じて番号を割り振った重なり情報:Object Visual Overlap（図 2(c)）

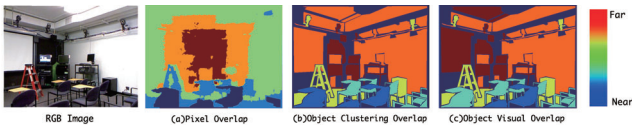


図 2: 重なり情報の可視化

Pixel Overlap は RGB-D 画像の pixel 値のみに依存し、Object Overlap は上記に加えてオブジェクトの位置関係と Semantic ラベル（以下、ラベルと呼ぶ）に依存している。

3.2 重なり情報の活用と利点

深度を求める前段階として RGB 画像から重なり画像を生成する。生成した重なり画像を深度推定モデルにおける大域情報として扱い、モデルの中間層に追加の入力として与えることで単眼深度推定の精度向上を図る。

重なり情報の推定問題を解くことは深度推定を分類問題に落とし込むことと同義である。よって、重なり情報を介して深度推定を行うことで、分類問題へのアプローチを深度推定に流用できるようになる。

4. 事前に重なり情報を与えた深度推定実験

4.1 実験設定

重なり情報が深度推定の精度向上に寄与するという仮説を検証するため、下記 5 種類の対照実験を行う。

1. 追加の入力を MSM に与えずに学習させる:Plane モデル
2. ラベルを追加の入力として MSM に与えて学習させる:Semantic モデル
3. 重なり画像を追加の入力として MSM に与え学習させる。重なり画像の形態は以下の 3 種類とする。
   (a) Pixel Overlap:Pixel モデル
   (b) Object Clustering Overlap:Cluster モデル
   (c) Object Visual Overlap:Visual モデル

各学習を 5 回ずつ行い、学習後のモデルの平均予測精度を 5 種類の評価関数を用いて比較する。学習の epoch は 20 を上限とした。また、ラベルに対する依存度をモデルの評価指標の一つとして用いる。

4.2 ラベルに対する依存度の比較方法

ラベルを用いてモデルを学習する場合、モデルがラベル情報に最適化され、RGB 画像に関係なくラベルの深度平均値を出力してしまう問題が考えられる。モデルの汎用性を確保したい場合この特性は好ましくないため、以下の方法でラベルに対する依存度を数値化し、モデルの評価指標として用いる。この指標をラベル依存度と呼ぶ。ラベル依存度が高ければ高いほど、入力画像に対するモデルの汎用性が低いことを表す。

1. RGB-D 画像内の各オブジェクトをラベル分類により領域分割した際、同一ラベルを持つ領域全体の平均深度とその標準偏差を計算する。この標準偏差をラベル標準偏差と定義する。
2. データセット内の全ての RGB-D 画像に対するラベル標準偏差を基準標準偏差とする。

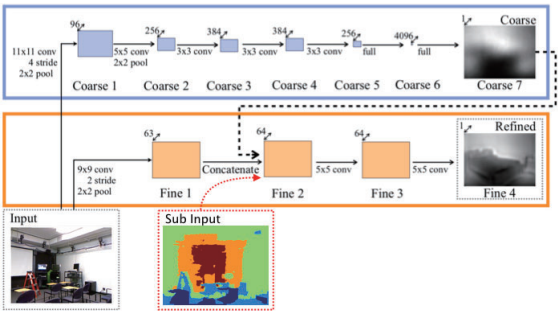


図 3: 重なり情報を追加入力とした MSM

Network	学習率	beta1	beta2	epsilon	decay
Global	0.001	0.9	0.999	10^{-7}	$1e-4$
Local	0.001	0.9	0.999	10^{-7}	$1e-4$

表 1: Adam に設定したハイパーパラメーター

任意のモデルを用い、データセット内の全ての RGB 画像から RGB-D 画像を求め、推定した RGB-D 画像からラベル標準偏差を求める。これをモデルラベル標準偏差とする。

3. モデルラベル標準偏差と基準標準偏差の差分を計算し、その絶対値の平均を求める。これをラベル依存度として定義する。

4.3 モデル構成

基本となるモデルに MSM を選択した。MSM の Fine2 ブロックと Fine3 ブロックの間にカーネルサイズ 1×1、フィルター数 64 の畳み込み層を加えたモデルを実験では使用する。追加の入力を与えない場合は上記のモデルをそのまま使用する。ラベル情報、重なり画像を入力として与える場合、追加の入力を 55×74 にリサイズし、各 pixel の重なり情報を one-hot 表現に直したものを上記のモデルの Fine2 ブロック（図 3 赤枠下部）に結合する。

損失関数には Scale-Invariant Error:sc-inv.[4] を使用し、定数 λ は $\lambda = 0.5$ の値を用いている。

評価関数には thresholded accuracy、mean squared error:MSE、root mean squred error:RMSE の linear と log、sci-inv. の 5 種類を使用した。

最適化手法には Adam を選択した。使用したハイパーパラメーターは表 1 の通りである。

活性化関数には ReLU を用いている。

4.4 データセット

4.4.1 NYU Depth Dataset v2

モデルの学習には、入力データとして RGB 画像と重なり画像、教師データとして RGB-D 画像が必要になる。よって、下記の 3 種類のデータの組を収録したデータセットとして NYU Depth Dataset V2[6]（以下 NYU Depth）を実験で用いる。ラベル情報は重なり画像の生成に使用する。

- RGB 画像
- RGB-D 画像
- ラベル情報

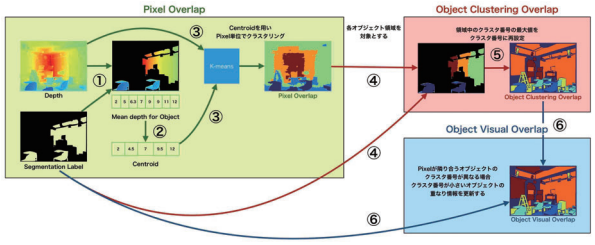


図 4: 重なり画像の作成手順

NYU Depth は屋内画像を中心とした画像データセットである。本実験では学習の度にデータ 1449 組のうち 1015 組を training set としてランダムに選出し、残りの 434 組を validation set として用いている。ラベル情報は参考文献 [7] を元に 41 種類に限定したものを用いる。

4.4.2 重なり画像

重なり画像は自明に求められる情報ではないため、以下の方法で各重なり画像を自動的に生成している。Pixel Overlap は RGB-D 画像とラベルから、Object Clustering Overlap は Pixel Overlap から、Object Visual Overlap は Object Clustering Overlap とラベルから順次求まる。重なり数は各形態共に 6 とし、うち整数 0 を Segmentation ラベル情報の存在しない pixel に割り振った。

- Pixel Overlap
  Segmentation ラベル情報から各オブジェクトの領域を特定し、各々の領域に対して RGB-D 画像の平均値を求める（図 4 緑枠 1）。平均値群の最小値から最大値の範囲を 4 分割し、最小値と各分割点と最大値の計 5 つの値を centroid として設定する（図 4 緑枠 2）。K-means 法によって RGB-D 画像を pixel 毎にクラスタリングする（図 4 緑枠 3）。属している pixel の値が小さいクラスから順番に、1 から始まる整数値をクラスタ番号として与える。クラスタ番号を重なり情報として扱う。

- Object Clustering Overlap
  Pixel Overlap の作成手順を適応する。次に、各オブジェクトの領域を再度対象とし（図 4 赤枠 4）、領域内におけるクラスタ番号の最大値をそのオブジェクト領域全体のクラスタ番号として再設定する（図 4 赤枠 5）。クラスタ番号を重なり情報として扱う。

- Object Visual Overlap（図 4 青枠）
  Object Clustering Overlap の作成手順を適応する。オブジェクトが持つ重なり情報を 1 で初期化する。次に、最も大きいクラスタ番号を持つオブジェクト A から順番に対象とし、画面上で隣り合うオブジェクト B と比較を行う。

  1. A と B のクラスタ番号が異なる場合、クラスタ番号が小さいオブジェクトを C とする。C の重なり数が比較対象先の重なり数以下の時、C の重なり情報をもう一方の重なり情報より 1 大きい値に設定する。
  2. A と B のクラスタ番号が同じ場合、A と B の重なり情報を比べてより大きい値を新しい重なり情報として両者に設定する。
  3. 上記の操作を全てのオブジェクトに対して行なった後、重なり情報の最大値より 1 大きい値を全ての重なり情報から引き、絶対値を取る。

	Plane	Semantic	Pixel	Cluster	Visual
$\delta < 1.25$	0.341	0.434	**0.465**	0.396	0.424
$\delta < 1.25^2$	0.601	0.737	**0.783**	0.682	0.724
$\delta < 1.25^3$	0.764	0.889	**0.927**	0.840	0.881
MSE	0.0258	0.0138	**0.0112**	0.0189	0.0145
RMSE(linear)	1.169	0.869	**0.779**	0.992	0.889
RMSE(log)	1.418	**0.407**	0.477	1.058	0.420
sc-inv.	0.287	0.157	**0.123**	0.209	0.163

表 2: 事前に重なり情報を与えた各モデルの深度予測精度

Plane	Semantic	Pixel	Cluster	Visual
0.0134	0.0126	0.0095	0.0095	0.0112

表 3: 事前に重なり情報を与えた各モデルのラベル依存度

4.4.3 データのかさ増し

任意の操作を加えることで、学習時に使用できる画像の量を 5 倍に増やしている。操作内容は以下の 4 種類であり、training set の各画像に各操作を一回ずつランダムに適応している。

- Scale:画像を拡大する。倍率は 1 から 1.5 倍。
- Rotation:画像を回転する。範囲は ±5 度。
- Shift:画像を上下または左右にシフトする。シフトの最大距離は画像の幅の 20%。
- Flips:画像を水平方向に反転する。

4.5 実験結果

表 2 の上 3 行の評価関数は値が大きいほど、下 3 行の評価関数は値が小さいほど優秀である。表 3 は各モデルのラベルへの依存度を表し、大きいほど依存度が高い。

表 2 より、RMS(log) を除いた全ての評価関数において Pixel モデルが他のモデルの精度を上回っており、残る提案モデル 2 つも Plane モデルを上回る精度を記録している。また、表 3 より、評価関数上では Cluster モデルと Visual モデルが Semantic モデルよりも低い精度を記録しているが、ラベルへの依存度は Semantic モデルの方が高く汎用性の面で劣っている。RMS(log) の評価関数において Semantic モデルの精度が高い理由は、オブジェクト単位で均一な深度を出力した結果、予測深度の外れ値が少なくなったためであると考えられる。

5. 推定した重なり情報を用いた深度推定実験

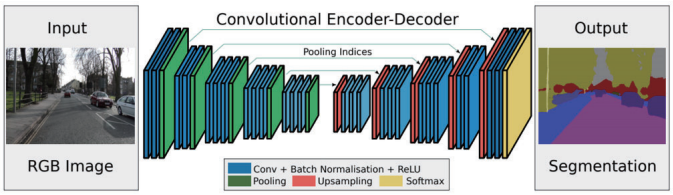
5.1 実験設定

重なり情報の推定可能性を検証するため、下記 3 種類の対照実験を行う。

1. 重なり画像を教師データとして SegNet[8] を学習させる。SegNet によって推定した重なり情報を追加の入力として MSM に与え深度を推定する。
   重なり画像の形態は以下の 3 種類とする。

   (a) Pixel Overlap:Get Pixel モデル

   (b) Object Clustering Overlap:Get Cluster モデル



	Layer	1.1	1.2	MaxPool	1.3	MaxPool	1.4	MaxPool	1.5
Encorder	Size	228 × 304	228 × 304	114 × 152	114 × 152	57 × 76	57 × 76	28 × 38	28 × 38
	chan	64	32	32	64	64	128	128	256
	ker. sz	1 × 1	3 × 3	2 × 2	3 × 3	2 × 2	3 × 3	2 × 2	3 × 3
	Layer	1.1	UpSamp	1.2	UpSamp	1.3	UpSamp	1.4	ZeroPad
Decorder	Size	28 × 38	56 × 76	56 × 76	112 × 152	112 × 152	224 × 304	224 × 304	228 × 304
	chan	256	256	128	128	64	64	32	32
	ker. sz	3 × 3	2 × 2	3 × 3	2 × 2	3 × 3	2 × 2	3 × 3	–
	Layer	1.5	MaxPool	1.6					
Decorder	Size	110 × 148	55 × 74	55 × 74					
	chan	64	64	6					
	ker. sz	9 × 9	2 × 2	1 × 1					
	stride	2 × 2	–	–					

図 5: SegNet[8]

Network	学習率	rho	epsilon	decay
SegNet	1.0	0.95	10^{-7}	0.0

表 4: AdaDelta に設定したハイパーパラメーター

(c) Object Visual Overlap:Get Visual モデル

評価方法とデータセットは 4.1 節と 4.4 節に譲る。

5.2 モデル構成

重なり情報の推定部に用いるモデルとして SegNet（図 5）を選択した。SegNet は Semantic Segmentation に用いられる pixel 単位の分類ネットワークである。

MSM の Coarse7 ブロック（図 3 青枠右）の出力を 4 倍に UpSampling し、ZeroPadding によって 228 × 304 にリサイズしたものを SegNet への追加の入力として input に結合する。重なり情報と RGB 画像から推定された重なり情報を 55 × 74 にリサイズし one-hot 表現に直したものを、4.3 節の MSM モデルの Fine2 ブロック（図 3 赤枠下部）に結合する。MSM モデルの学習は 4.3 節と同様に行う。SegNet の出力層の活性化関数には SoftMax を、SegNet の最適化手法には AdaDelta を使用している。AdaDelta のハイパーパラメータは表 4 の通りである。その他の設定は 4.3 節に譲る。

5.3 実験結果

表 5 より、全ての評価関数において Get Pixel モデルが他のモデルの精度を上回っており、 残る提案モデル 2 つも Plane モデルを上回る精度を記録している。対して、表 6 よりラベルへの依存度を見ると、Get Cluster モデルを除いて Plane モデルに汎用性の面で劣っている。最も汎用性の高かった Get Cluster モデルでも Semantic モデルの汎用性を下回る結果となった。

6. 考察とまとめ

本研究では単眼深度推定における重なり情報の有用性を証明することができた。しかしながら重なり情報の推定可能性については未だ課題が残り、現時点で最も推定精度の高い Object Visual Overlap でも accuracy は半分を下回っている（表 7）。3.1 節で述べた通り Pixel Overlap 以外の重なり情報はラベル情報に依存しているので、ラベル情報を SegNet に追加入力として与えることでこの問題は解決できる可能性が高い。

	Get Pixel	Get Cluster	Get Visual
$\delta < 1.25$	**0.430**	0.380	0.412
$\delta < 1.25^2$	**0.716**	0.656	0.702
$\delta < 1.25^3$	**0.875**	0.821	0.866
MSE	**0.0144**	0.0195	0.0153
RMSE(linear)	**0.888**	1.016	0.927
RMSE(log)	**0.398**	0.961	0.409
sc-inv.	**0.157**	0.217	0.169

表 5: 推定した重なり情報を用いた各モデルの深度予測精度

Get Pixel	Get Cluster	Get Visual
0.0135	0.0128	0.0138

表 6: 推定した重なり情報を用いた各モデルのラベル依存度

	Get Pixel	Get Cluster	Get Visual
accuracy	0.341	0.321	0.346

表 7: SegNet による各種重なり情報の予測精度

参考文献

[1] Li, Jun, Reinhard Klein, and Angela Yao. ”A two-streamed network for estimating fine-scaled depth maps from single rgb images.” Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy. 2017.

[2] J.M. Loomis. Looking down is looking up. Nature News and Views, 414:155
UTF2013156, 2001.

[3] Ladicky, Lubor, Jianbo Shi, and Marc Pollefeys. ”Pulling things out of perspective.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.

[4] Eigen, David, Christian Puhrsch, and Rob Fergus. ”Depth map prediction from a single image using a multi-scale deep network.” Advances in neural information processing systems. 2014.

[5] Saxena, Ashutosh, Min Sun, and Andrew Y. Ng. ”Learning 3-d scene structure from a single still image.” Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. IEEE, 2007.

[6] NYU Depth v2 `https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html`（参照 2019-02-14）

[7] Deng, Zhuo, Sinisa Todorovic, and Longin Jan Latecki. ”Semantic segmentation of rgbd images with mutex constraints.” Proceedings of the IEEE international conference on computer vision. 2015.

[8] Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. ”Segnet: A deep convolutional encoder-decoder architecture for image segmentation.” arXiv preprint arXiv:1511.00561 (2015).