深層学習モデルを用いたステレオカメラ画像による高精度な3D計測技術

High-Precision Depth Map Estimation Technology for Stereo Images Applying Deep Neural Networks

関　晃仁 SEKI Akihito　　マーク　ポリフェイ Marc POLLEFEYS

自動車の自動運転，ロボットやドローンなどの自律移動に際しては，周囲の3次元（3D）環境を把握することが必要不可欠になる。ステレオカメラを用いた3D計測では，2台のカメラで同一物体が写っている場所を基に距離を計測するため，正確な左右画像間の対応付けが重要である。

東芝は，ステレオカメラで撮影された左右の画像を用いて各画素の視差を正確に求めるため，画像全体にわたって画像間の対応付けを正確かつ緻密に推定できる，Semi-Global Matching（SGM）手法をベースにした深層学習モデルSGM-Netを開発した。ノイズや，模様のない領域，隠れている部分などの影響を取り除いて画像内の対応付けを全体最適化する際に，従来は人手によるパラメーターの設計・調整が必要であったが，SGM-Netでは自動で推定できる。これにより，対応付けの正確性が向上し，より高精度な3D計測が可能になった。

Technologies capable of precisely understanding three-dimensional (3D) environments are essential for self-driving cars and autonomous mobile robots and drones. In 3D measurement using a stereo camera system equipped with two cameras to generate binocular disparity, accurate correspondence of the identical points in the left- and right-hand images using a disparity map is important to estimate the distance to objects, providing a so-called depth map.

Toshiba Corporation has developed a deep neural network called SGM-Net that can achieve accurate dense correspondence of stereo images based on the semi-global matching (SGM) method, which is widely used as a regularization method. SGM-Net can automatically estimate the optimal parameters by more effectively removing noise, textureless areas , and occluded parts compared with conventional manual tuning methods. From the results of quantitative evaluation using an open dataset, we have confirmed that SGM-Net offers higher performance for depth map estimation than other reported methods.

1. まえがき

自動車の自動運転，ロボットやドローンの自律移動などの研究が世界中で過熱している。このような機能を実現するには，コンピューターが自分の周囲の環境を正確に認識する必要がある。周囲の3D環境を計測するには，カメラや，レーザーレーダー，ミリ波レーダーなどを使う方法がよく用いられる。レーザーレーダーやミリ波レーダーは，光や電波を送信し，その反射波を受信することで距離を計測するアクティブセンサーであり，一般的に距離の計測精度は高いが，空間分解能が低いという特徴がある。単眼カメラやステレオカメラは，画像に映った物体同士の“画像間での対応位置”を求め，**図1**に示すように，三角測量の原理で距離を計測するパッシブセンサーであり，空間解像度が高く，小さい物体までの距離も計測できることから，重要な役割を担っている。3D計測技術として，正確な距離を求めるには“画像間での対応位置”，つまり視差を正確に求められることが重要である。

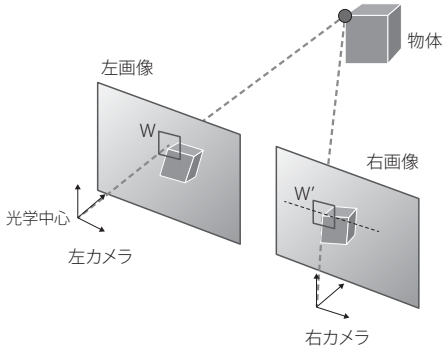


図1．ステレオカメラを用いた三角測量の原理に基づく距離計測
2枚の画像に写った画像間の対応位置から，3D位置を計測する。
Distance measurement based on triangulation using images captured by stereo camera

東芝は，画像間の対応付けを画像全体にわたって正確かつ緻密に推定する深層学習モデルを開発し，SGM-Netと名付けた[1]。これにより，従来よりも正確に距離計測を行うことができる。ここでは，その原理や有効性について述べる。

2. 画像間の対応付け

2章では，ステレオカメラで撮影された左右画像から各画素の視差を推定する，処理フロー（**図2**）の全体について説明する。

⑴ 前処理　　左右のカメラで撮影された2枚のステレオ画像に対し，レンズの歪み（ひずみ）補正を行うとともに，2枚の画像間の高さ方向に対して同一の物体が映り込むように平行化を行う。

⑵ ローカルな対応付け　　平行化されたステレオ画像に対して，注目する画素周辺の小領域を左カメラで撮影された画像（左画像）から切り出し，その小領域（図1のW）と最も一致する場所（図1のW'）を右画像から探索する。これは，画像内の各位置に対して独立に行うため，一般にローカルな対応付けと呼ばれる。左画像の各位置について同様の処理を施すことで，画素ごとに視差画像を生成できる。

⑶ グローバルな対応付け　　前述の⑵で求められた視差画像は，各画素で独立に視差を求めるため，画像に重畳されたノイズや，模様のない領域，左右画像間での隠れなどの影響で，視差画像に相当量の対応付けエラーが入ってしまう。当社は，このようなエラーを特定する手法も開発した[2]。ここでは，周辺画素で推定された視差情報も利用して視差を新たに推定し，その際に深層学習モデルとして独自に開発したSGM-Netを用いて全体最適化処理を行う。次の3章で詳しく説明する。

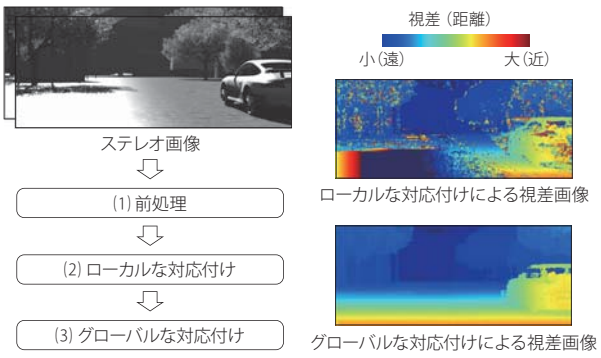


図2. ステレオ画像からの視差推定処理フロー
ローカル及びグローバルな対応付けを行うことで，視差を推定する。グローバルな対応付けにより，正確な視差画像が生成される。
Flow of processes for disparity estimation from stereo images

3. 深層学習モデルを用いたグローバルな対応付け

当社が開発した，深層学習モデルを用いた視差の推定方法について述べる。画素単位で大きく奥行きが変化するシーンは，現実の世界では皆無であることから，隣接画素間では視差に急激な変化が起きにくいように，拘束条件を付けて視差画像を推定する。この拘束条件を考慮したエネルギー関数Eは，式⑴で表される。

$$E(\hat{D})=\min_{D}\sum_{x}\left(C(\mathrm{x},d^{x})+\sum_{y}P|\delta^{x,y}|\right) \quad (1)$$

ここで，Dは，存在する全ての視差画像を表し，推定したい視差画像$\hat{D}$では，画素位置xにおける視差d^{x}が持つマッチングコスト$C(x,d^{x})$，及び$\delta^{x,y}=d^{x}-d^{y}$（xの視差とx以外にある画素位置yの視差との差）に対して正値であるペナルティーPが加算される。つまり，二つの画素位置で視差が異なる場合にはエネルギーが大きくなり，視差が同じになる場合にはエネルギーが小さくなる。もし，Pがゼロなら，周辺の画素位置の情報は考慮されず，マッチングコスト$C(x,d^{x})$が最も小さくなる視差が各画素位置で選ばれるので，ローカルな対応付けと完全に一致する。Pがゼロでない場合には，式⑴を実時間で解くことは難しいため，式⑵のように変形したSGM[3]を用いる。

$$E(\hat{D})=\min_{D}\sum_{x}\left(C(x,d^{x})+\sum_{y\in N_x}P_1T\left[|\delta^{x,y}|=1\right]+\sum_{y\in N_x}P_2T\left[|\delta^{x,y}|>1\right]\right) \quad (2)$$

ここで，Tは，因数が“真”のときに1で，それ以外では0の演算子，N_xは，画素位置xの近隣画素を表している。式⑴との違いは，画素位置xの視差と周辺にある画素位置yの視差との差$\delta^{x,y}$が1画素の場合のペナルティーをP_1とし，2画素以上の差があるときはP_2とすることである。更に，画素位置yは，画素位置xと隣接する画素だけにする。これにより，P_1は，徐々に奥行きが変化する場合（例えば道路の路面），P_2は急激に奥行きが変化する場合（例えば物体の境界）に対するペナルティーとなる。また，画素位置yに隣接する八つの画素に対しては，同時にエネルギー関数を最小化するのではなく，縦・横・斜めなどの走査方向を一つ決め，その方向で式⑵によるエネルギーの累積値を独立に求

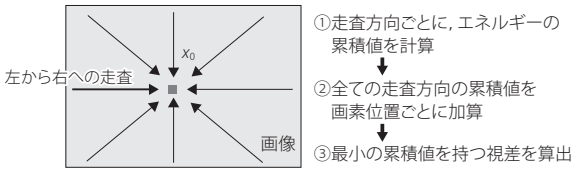


図3. グローバルな対応付けのSGM-Netによる処理手順

画像の走査方向ごとにエネルギーの累積値を求め，累積値の和が最小となる視差を各画素で求める。

Global matching procedures using SGM method

める。加えて，ほかの走査方向も同様に求めて最後にそれらを加算し，各画素位置で累積値が最小値となる視差を求める（**図3**）。これによって，画素ごとのグローバルな対応付けがなされる。SGMでは，パソコンなどの高性能なCPUだけでなく，組み込みCPUなどの比較的低速な環境でもリアルタイムに動作させることができる。

式(2)のP_1とP_2の二つの係数は，負でなければ任意に決めることができ，ペナルティーが大きいほど，そのペナルティーに相当する視差の変化が生じにくい作用を生む。従来，これらのペナルティーは，人手で設計されていた。例えば，物体の境界では，物体と背景の間に輝度差が生じることが多いということが経験的に分かっているため，エッジがある場合にP_2を減じるようにしていた。ところが，エッジは物体境界以外にも存在するため，これが必ずしも適切な仮定ではないという問題があった。

当社は，ペナルティーの設計を人手によらず，深層学習モデルを用いて自動的に推定できるSGM-Netを開発した。ペナルティーの設定方法だけを変更するため，SGMが高速に計算できる特性は，これまでどおり維持される。

4. 深層学習モデルによるペナルティーの推定

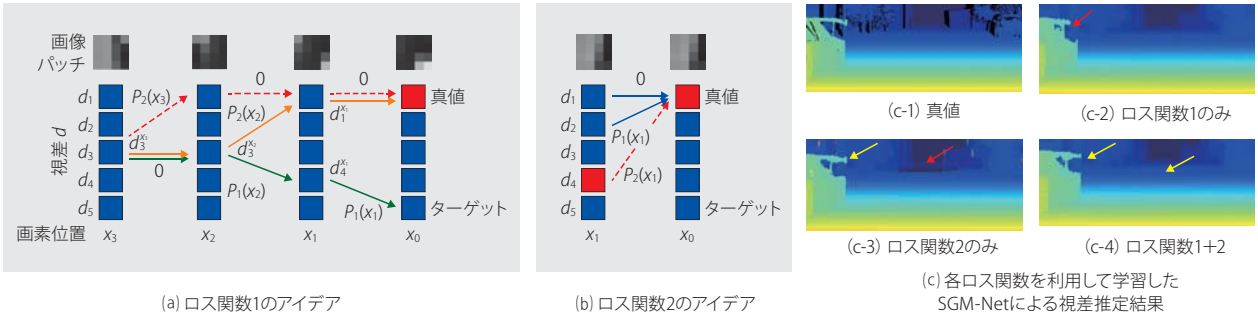
正確な視差画像を生成できるペナルティーは未知であるため，従来提案されている回帰モデルと呼ばれるような，あらかじめ定められた値になるようにペナルティーを直接的に深層学習で求めることはできない。そこで，その代わりに，視差画像が正確に生成されるように深層学習モデルで学習する。それには，新たなロス関数が必要である。ここでは，そのアイデアについて説明する。

まず，式(2)を基にしたエネルギーの累積値Lに着目すると，左から右へ走査する場合の累積値$L_{\rightarrow}$は式(3)となる。

$$L_{\rightarrow}(x_0, d)=C(x_0, d)+\min\{L_{\rightarrow}(x_1, d),\ L_{\rightarrow}(x_1, d\pm1)+P_1,\ \min_{i\neq d, d\pm1} L_{\rightarrow}(x_1, i)+P_2\} \quad (3)$$

$L_{\rightarrow}$は，現在の画素位置x_0における視差dでのマッチングコスト$C(x_0, d)$と，走査方向に対して一つ前の画素位置x_1の$L_{\rightarrow}$と，その視差との差に応じたペナルティーのうち最小のものとの和となり，それを画素位置に対して再帰的に求めていく。**図4**(a)は，式(3)の計算過程を図示したものである。例えば，x_0において，正しい視差（真値）とターゲットとする視差があったとする。真値に達する視差では，x_2でP_2が発生（以下，$P_2(x_2)$と略記）している（同図のオレンジ線の経路）。一方，ターゲットとする視差では，$P_1(x_2)$と$P_1(x_1)$が生じている（同図の緑線の経路）。

グローバルな対応付けを行う各画素の視差は，累積値Lが最小のものを求めるため，累積値Lが最小の視差（＝正しい視差）となるように，ペナルティーを深層学習モデルで学



(a) ロス関数1のアイデア　(b) ロス関数2のアイデア　(c) 各ロス関数を利用して学習したSGM-Netによる視差推定結果

図4. 深層学習モデルを学習するためのロス関数のアイデアと視差画像

ロス関数1は，累積値が最小となるように学習し，ロス関数2は，隣接する一つ前の画素位置だけを考える。

Ideas for loss functions of SGM-Net and estimated disparity maps trained by several combinations of loss functions

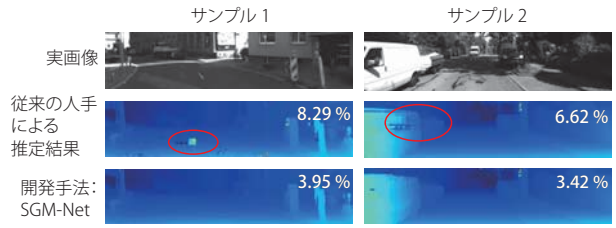


図5. 実画像における視差推定結果の比較

赤丸の領域では，推定誤りが特に大きく，右上に推定エラー率を示している。従来の方法に比べ，開発した方法は推定誤りが低減できている。

Comparison of disparity maps using real image datasets estimated by conventional manual tuning method and SGM-Net

表1. 公開データセットを利用した定量評価結果

Results of quantitative evaluation using open dataset

番号	推定方法	推定エラー率
1	SGM-Net	2.29 %
2	PBCP[2]	2.36 %
3	Displets v2[6]	2.37 %
4	MC-CNN-acrt[7]	2.43 %

PBCP：Patch Based Confidence Prediction
MC-CNN-acrt：Matching Cost - Convolutional Neural Network - Accurate architecture

習すればよく，これを一つ目のロス関数（ロス関数1）とする。真値となる視差でのLは，ターゲットとする視差でのLよりも小さくなる必要がある。よって，$P_2(x_2)$は小さくなり，$P_1(x_2)$と$P_1(x_1)$は大きくなるようにすればよい。その際，各画素位置には画像の輝度があるため，その画像パターンとなるときのペナルティーとして学習を行う。つまり，SGM-Netでは，各画素位置で得られる輝度などの情報を入力として，画素ごと及び方向ごとにペナルティーを推定する。

全ての画素位置において正しい視差が推定されるには，x_0での真の視差に達するLは，各画素位置において正しい視差を通る必要がある。図4(a)の赤破線の経路が，全ての画素位置において正しい視差を通った場合のLの計算過程であり，前述したオレンジ線の経路と異なる視差を通っていることが分かる。つまり，オレンジ線の経路に沿ったLだけでは，必ずしも正しくペナルティーが学習されるわけではないことが分かる。図4（c-2）は，このロス関数1を用いてSGM-Netを学習し，得られたペナルティーを用いて視差画像を生成したものである。全体としては正しく推定できているが，赤矢印で示すように，細かい形状が潰れてしまっている。そこで，次に説明する二つ目のロス関数を導入する。

二つ目のロス関数（ロス関数2）では，ロス関数1が現在の画素位置に到達するために通った全ての画素位置を考慮したのとは異なり，隣接する一つ前の画素位置だけを考える。図4(b)では，x_1での正しい視差d_4と，x_0での正しい視差d_1を通る経路が，そのほかの経路よりも小さくなければならない。そこで，$P_2(x_1)$は小さく，$P_1(x_1)$は大きくなるようにすればよい。図4（c-3）はこのロス関数2を用いて推定された視差画像である。ロス関数1に比べて細かい形状が推定されている（黄矢印の部分）とともに，部分的に誤った視差（赤矢印の部分）も推定されている。これは，ロス関数2では隣接する2画素しか考慮していないことと，ロス関数2を適用できる画素位置x_1は正しく視差が推定されていること，の二つの制限によっている。

そこで，ロス関数1とロス関数2を同時に利用してSGM-Netの学習を行う。その結果，図4（c-4）のように正確な視差画像が得られる。

5. 実験結果

SGM-Netの学習には，合成画像120枚[4]と実画像約200枚[5]を用いた。画像には，真の視差が教示されている。SGM-Netは，各画素位置の周辺5画素×5画素の画像パッチと，その位置を入力としたペナルティーP_1及びP_2を走査方向数だけ出力する。今回の実験では，走査方向は，縦と横の4方向とした。SGM-Netの計算時間は，汎用GPU（Graphics Processing Unit）を用いて0.02秒であった。実画像に対して適用した結果を，**図5**に示す。従来の人手による結果と比べ，特に赤丸で示す領域内の推定エラーが少ないことが確認できる。公開データセット[5]を利用したベンチマーク結果が**表1**である。表の1行目は，開発したSGM-Net，2行目は，以前に開発した対応付けの信頼度を用いる方法[2]，3行目は，車両などの形状モデルを用いる方法[6]，4行目は，人手によりパラメーターを調整した手法[7]である。この表から，SGM-Netはほかの手法に比べて精度良く推定できていることが分かる。

6. あとがき

当社は，深層学習モデルを用いることで，ステレオ画像から視差を正確に求める技術を開発した。画像間のグローバルな対応付けにおいて，深層学習モデルで推定したパラメーターを用いて，視差を正確かつ緻密に推定した。

自動車の自動運転，ロボットやドローンなどの自律移動のための周辺監視機器は，晴天の昼間に限らず，夜間や荒天などの様々な環境下でも安定に動作することが求められる。

今後は，耐環境性などについても検討を進めていく。

文　献

(1) Seki, A.; Pollefeys, M. “SGM-Nets: Semi-Global Matching with Neural Networks”. Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, HI, USA, 2017-07, IEEE. 2017, p.6640–6649.

(2) Seki, A.; Pollefeys, M. “Patch based confidence prediction for dense disparity map”. Proc. British Machine Vision Conference 2016 (BMVC). York, UK, 2016-09, BMVA. 2016, p.98.

(3) Hirschmuller, H. Stereo Processing by Semi-Global Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2008, **30**, 2, p.328–341.

(4) Mayer, N. et al. "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation". Proc. of 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016). Las Vegas, NV, USA, 2016-07, IEEE. 2016, p.4040–4048.

(5) Geiger, A. et al. "Are we ready for autonomous driving? the KITTI vision benchmark suite". Proc. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012). Providence, RI, USA, 2012-07, IEEE. 2012, p.3354–3361.

(6) Guney, F.;Geiger, A. "Displets. Resolving stereo ambiguities using object knowledge". Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015). Boston, MA, USA, 2015-06, IEEE. 2015, p.4165–4175.

(7) Zbontar, J.; Y.LeCun, Y. Stereo matching by training a convolutional neural network to compare image patches. Journal of Machine Learning Research, 2016, **17**, 1, p.2287–2318.

関　　晃仁　SEKI Akihito, D.Eng.
研究開発本部 研究開発センター
メディアAIラボラトリー
博士（工学）　情報処理学会会員
Media AI Lab.

マーク ポリフェイ　Marc POLLEFEYS, D.Eng.
スイス連邦工科大学 チューリヒ校
コンピューターサイエンス学部 教授
博士（工学）　IEEE会員
Institute of Visual Computing Department of Computer Science, ETH Zurich