JSiSE Research Report
vol.31,no.1(2016-5)

道路進行方向における新視点画像の生成システム

大政 孝充[*1], 鈴木 一史[*2]
放送大学大学院[*1], 放送大学[*2]

The System that Produces New View Synthesis along Roads

Takamitsu Omasa[*1], Motofumi Suzuki[*2]
Graduate School of the Open University of Japan[*1], The Open University of Japan[*2]

In the field of education, it is useful for students to learn outdoors rather than indoors. However, it is difficult to learn outdoors because of various restrictions. In order to solve these problems, we propose a new system by which students can obtain valuable experiences as if they were outdoors. In this system, it is important to produces image from a new viewpoint in this field, there are already several methods such as Sweep-Space algorithm or DeepStereo which uses CNN. While these methods produce realistic images, there are two main drawbacks. First, these methods require an enormous amount of time. Second, these methods require many images that are taken at closed viewpoints. We will limit these methods to the task of rendering digital images of road scenes, thereby finding solutions for these problems.

キーワード: 新視点画像合成，イメージベースドレンダリング，画像処理

1. はじめに

学校の生徒が戸外での体験を通して学ぶことは多い．もし仮に生徒が世界各地の道路上を自由に歩き回ることができれば、彼らが得られるものは計り知れないだろう．しかし実際には時間的制約や経済的制約のために戸外で歩き回ることは容易でない．一方で近年における情報技術の発展により、我々は様々な分野でヴァーチャルな体験を得ることができるようになった．しかしその現在においても、生徒が屋内にいながら戸外に出て自由に歩き回るような感覚を得ることは難しい．なぜなら戸外の完全な 3D モデルが存在しないため、外の景色と同じものをヴァーチャルに作り出すことが容易でないのだ．よって、なんらかの方法で 3D モデルを作り出すか、あるいは 3D モデルを代用する必要がある．

画像処理の研究分野において、この問題への対応策の 1 つとして image-based rendering という研究領域が存在する．これは複数の画像から 3D モデルを構築する領域である．この領域には幾つかの手法が提案されている．しかし本研究の目的を鑑みた場合、必ずしも完全な 3D モデルを構築する必要はない．なぜなら戸外を歩き回る状況においては、その主たる行動範囲は道路上に限定される．よって道路上からの視点による道路周辺を捉えた映像があれば足りるのである．

そこで我々は Google Street View に代表される、風景画像配信サービスに着目した．これらのサービスでは、道路上から見た風景画像を無料で大量に得られる．しかし一般的にこれらのサービスが配信するのは、道路に沿った数十メートル地点ごとの風景画像である．これだけでは道路を歩くような感覚のシステムは作れない．よって離れた 2 地点から撮影した風景画像から、その中間地点から見た風景画像を自動生生成する必要がある．

このように画像を生成する手法は new view synthesis と呼ばれる．同分野では従来からアルゴリズムで幾つかの手法が提案されている．いずれも挑戦的な試みではあるものの、実際の風景画像と見間違うような画像の生成には至っていない．しかし 2015 年において deep learning を用いた DeepStereo[5]と呼ばれる手法が提案され、その高い性能が確認された[6].

近年コンピュータビジョンの各分野では deep learning を用いることで従来のアルゴリズムによる結果を上回る、高い性能が得られている．このような状況を踏まえると、new view synthesis においても今後 deep learning が主流になると思われる．しかし DeepStereo を我々が目的とするタスクに使用することは、いくつかの問題のため困難である．このような背景から、本研究では DeepStereo の仕組みを真似た新手法を提案する．それは既存の手法を組み合わせ、それに独自のアルゴリズムを加えたものである．

2. 関連研究

2.1 アルゴリズムを用いた new view synthesis

従来からアルゴリズムで処理する new view synthesis の手法は幾つか提案されている．O. Woodford[8]らは CRF（conditional random fields）によるグラフカット法を改良したモデルを用いた．また G. Chaurasia[7]らのモデルは画像を superpixel に分割し、それを歪めて新視点画像を生み出した．

いずれの手法においても非現実的な画像が生成されるという問題点がある．これは特にオクルージョンの周辺で顕著である．

2.2 deep learning を用いた new view synthesis

DeepStereo は deep learning を使って新視点からの画像を創り出そうという意欲的な試みである．この手法では、まず複数視点からの画像から新視点における画像候補を plane-sweep 法により複数生成する．これら候補画像を deep learning のモデルへ入力する．deep learning のモデルは２つの Tower から構成される．１つ目は Color Tower であり、画像の色情報を学習する．２つ目は Select Tower であり、深さの確率を学習する．この２つを最上層にて統合し、出力画像を生成する．出力画像と実際に新視点から撮影された画像との誤差を求め、学習する．

この仕組みにより DeepStereo は高い性能を実現した．しかしこれを我々が目的とするタスクに使用する場合、２つの点で問題が生じる．１つ目の問題点は、学習済みのモデルで出力する場合においても、計算に膨大な時間を要することである．これにより、リアルタイムの処理が不可能となる．よってそのままでは我々の求めるシステムに対応できない．

２つ目の問題点は、入力画像として近接する５地点の画像が必要なことである．Google 自身はこのような画像を有しているが、一般にはその一部が公開されているのみである．代わりに離れた地点から撮られた画像を用いて同様の処理を行った場合、精度の低下が考えられる．よって DeepStereo の仕組みを用いるにはこれを変化させる必要がある．我々の手法は DeepStereo で用いられた plane-sweep 法を真似つつも、全体を学習に頼らない新たなものである．具体的には semantic segmentation で deep learning を用い、これとアルゴリズムを組み合わせている．

2.3 deep learning を用いた semantic segmentation

近年、deep learning を用いた semantic segmentation に関しては幾つかの手法が提案されている．J. Long ら[1]は CNN（convolutional neural network）で通常用いられる高層における全結合を畳み込み層に変えたモデルで従来のアルゴリズムの性能を大きく上回る精度を達成した．P. Pinheiro らの DeepMask[3]は、通常の CNN 構造の後に２つに分割された構造をつなげたモデルである．分割された１つ目の構造からは物体のセグメントを出力し、２つ目の構造からはその物体の可能性を出力する．S. Zheng ら[4]は条件付き確率場と RNN（recurrent neural network）を組み合わせたモデルを提示している．これらのモデルはいずれも学習済みモデルで出力する際にも多くの時間を要する．

V. Badrinarayanan らの SegNet[2]は CNN の encoder の後に同じ構造で順序が逆の decoder を繋げたモデルを用いた．SegNet は車載カメラからの映像のように路上からの視点に対応するため作られたものである．よってリアルタイムに対応する高い処理速度を達成している。本研究ではこの手法を利用する．一方で SegNet のセグメンテーションは他の手法に比べて幾分粗い．本研究ではこの点をアルゴリズムと組み合わせることで克服する．

2.4 アルゴリズムを用いた領域分割

アルゴリズムを用いた領域分割に関しては幾つかの手法が提示されている。ミーンシフト法[10]は標本

点から確率密度関数の極大点を探索する手法であり、対象画像の追跡にも用いられる．グラフカット法[11]は各画素を対象と背景に分割する問題を各画素のラベル付け問題として解く手法である．SLIC[12]は色情報と距離情報をもとに領域に分割する手法である．我々の手法では領域分割として SLIC を用いる．

2.5 plane-sweep 法

R. T. Collins[9]は複数の視点における画像から新たな視点における画像を生成する手法を提案した．まず求める新視点から複数の距離$d_i(i=1,2,\cdots,n)$に平面$Z_i(i=1,2,\cdots,n)$を想定する．この平面それぞれに対して複数の視点$m_j(j=1,2,\cdots,k)$から画像を投影する．これを全ての視点m_jに対して行う．ある注目物体 Yがd_iの距離に存在するとする．そうすると、d_iの距離にある平面Z_iにおいては、この注目物体 Yがいずれの視点m_jからも平面Z_i上の同一座標 X に投影されることになる．つまり座標 Xにおいては、それぞれの視点から投影された色は等しい．よって色が等しくなる座標は対象の物体が一致すると見なして、その対象物体の距離をd_iと考える．このようにしてそれぞれの物体の距離を求めると、新視点からの画像も作成できる．我々の手法では、plane-sweep の手法を用いて、２視点からの画像を別の新視点における画像上に投影する．

3. 提案手法

提案手法の概略は以下のようになる．まずある視点から撮影した画像に対して、何らかのアルゴリズムで superpixel の領域に分割する．一方で、同じ画像を deep learning の semantic segmentation 系モデルにかけ、その出力を得る．次にこの両者を照合し、superpixel 領域ごとにそのクラスが何であるかタグ付けする．この操作を別の視点から撮影した画像に対しても行う．

次に、plane-sweep 法により、新視点からの距離がそれぞれ違う n 個の平面を生成する．この平面を使って、２地点からその中間となる新視点に対してピクセルごとの投影を行う．これを全ての n に対して行う．最後に、どの n が最もらしいか判定する．この判定には、色情報およびクラス情報を用いる．

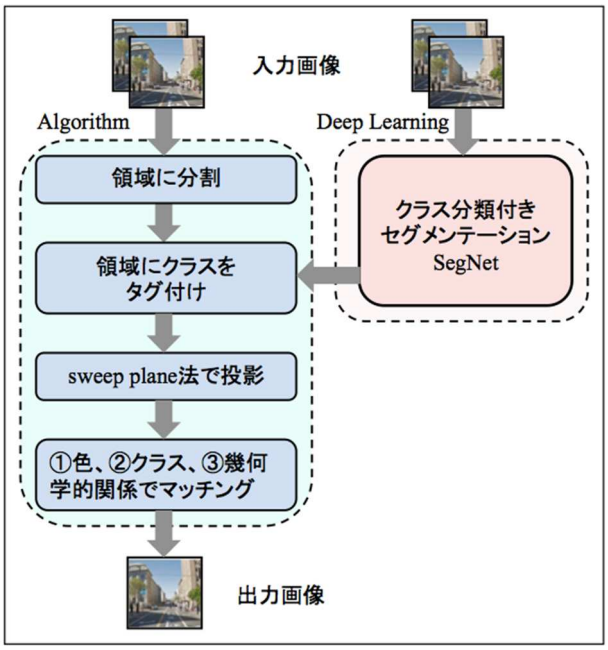

図 1 提案手法の全体図

以下、提案手法の詳細を述べる．

3.1 領域に分割

領域分割で用いる SLIC の概要は以下である．まず分割数 k を決定し、画像全体を k 個の格子状領域に均等に分ける．各領域に i の番号をラベリングする．初期の各領域における色情報及び座標を

$$C_i^0 = \left(l_i^0, a_i^0, b_i^0, x_i^0, y_i^0\right) \cdots (3.1.1)$$

とする．分割領域のサイズを$S \times S$とする．注目画素に対し、$2S \times 2S$の範囲で(3.2)～(3.4)式を用いて探索を行う．

$$D = \sqrt{{d_c}^2 + m^2 {d_s}^2} \cdots (3.1.2)$$

$${d_c}^2 = (l_j - l_i)^2 + (a_j - a_i)^2 + (b_j - b_i)^2 \cdots (3.1.3)$$

$${d_s}^2 = (x_j - x)^2 + (y_j - y_i)^2 \cdots (3.1.4)$$

ここでd_Cは注目画素と中心画素との色の差、d_Sは注目画素と中心画素とのユークリッド距離、m は色に対する距離の重みである．全ての画素をいずれかの領域にクラスタリングすれば、0 回目の操作は終了である．以上の操作を各領域の中心座標が動かなくなるまで繰り返す．

図 2 に SLIC による領域分割の例を示す．

図 2 SLIC による領域分割例　元画像（左）、SLIC による処理結果（右）

3.2　deep learning による semantic segmentation

semantic segmentation で用いる SegNet は encoder 部の後に decoder 部をつなげた構成である．encoder 部では畳み込み層とプーリング層の組み合わせを 5 回繰り返す．decoder 部では upsampling 層と畳み込み層の組み合わせを 5 回繰り返し、最後に softmax 層で出力する． SegNet を用いた場合の 2 地点入力画像を図 3 に示す．

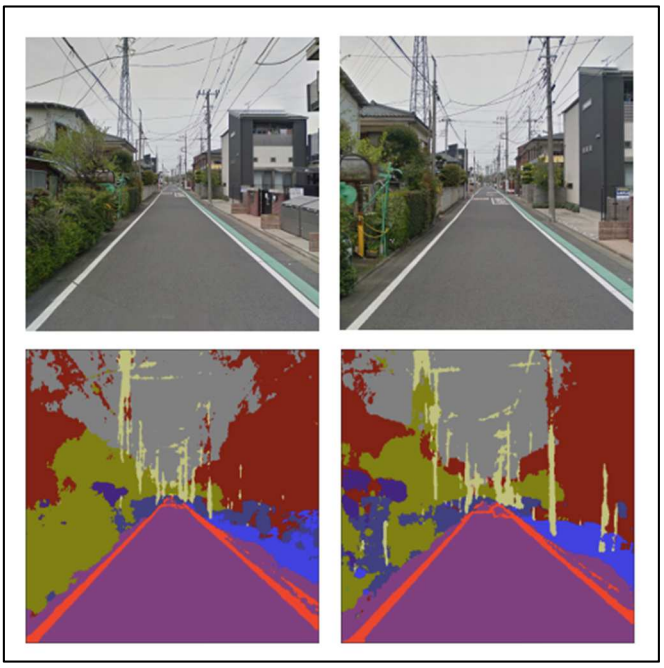

図 3 SegNet による出力結果　上側が入力画像、下側がそれぞれの出力結果．

3.3　領域にクラスをタグ付け

3.1 で得られたそれぞれの領域に対してクラス付けを行う．i番目の領域S_iに含まれる全てのピクセルの個数をN_{S_i}とする．i番目の領域S_iに含まれるピクセルのうち、3.2 で求めた j 番目のクラスに属する個数を$N_{S_i}^j$とする．

$$\frac{N_{S_i}^j}{N_{S_i}} \geq Th \cdots (3.3.1)$$

ならば、領域S_iはj番目のクラスとする．Th は閾値である．この結果を図 4 に示す．

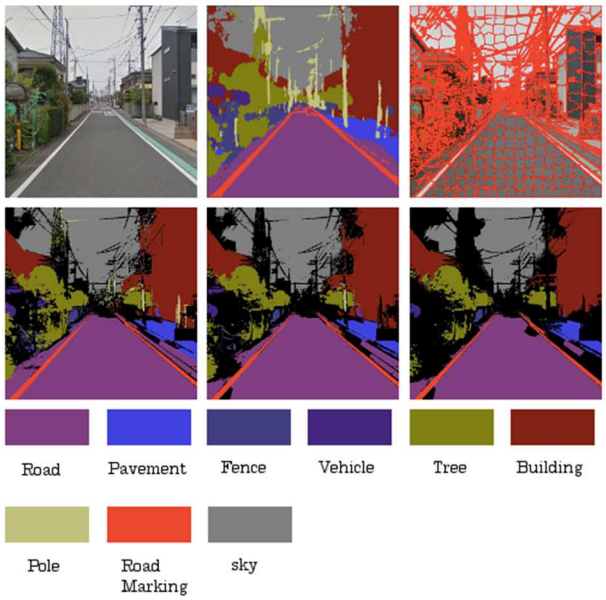

図 4 領域ごとにタグ付けされたクラス　元画像（左上）SegNet による出力（中央上）、SLIC による領域分割（右上）. 下段はクラスをタグ付けされた領域で左から閾値 0.6、0.7、0.8. 黒色はクラスをタグ付けされていない領域

3.4　sweep-plane で投影

図 5 のように幾何学的関係を単純化する．

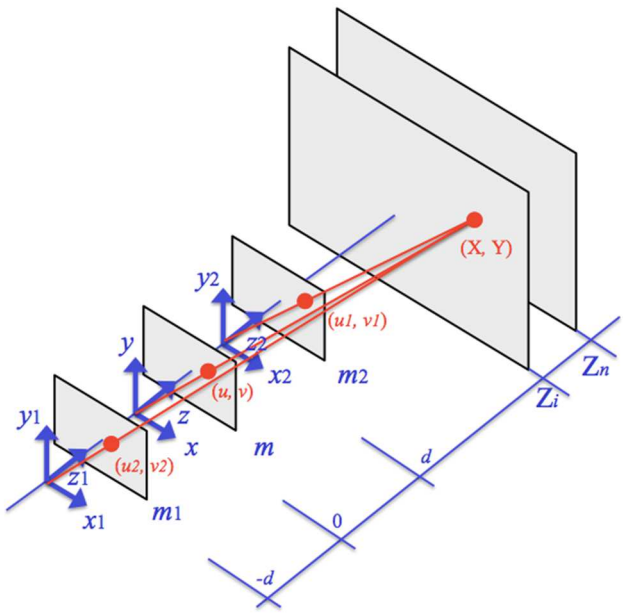

図 5 撮影地点の幾何学的関係

m_1地点とm_2地点から撮られた 2 枚の画像から新視点mにおける画像を生成する．m 視点のカメラ座標系をワールド座標系と一致させる．m_1のカメラ座標系およびm_2のカメラ座標系はワールド座標系をz軸方向にそれぞれ$-d$、d平行移動させたものとする．カメラ座標系はそれぞれの画像の中心を通るようにする．sweep-plane を n 枚想定する．i番目のsweep-planeのワールド座標系におけるz座標の値をZ_iとする．ワールド座標系における$X(X,Y,Z_i)$の値をm、m_1、m_2からの画像に投影した場合、その画像上の座標をそれぞれ$m(u,v)$、$m_1(u_1,v_1)$、$m_2(u_2,v_2)$とする．X およびこれらの点における同時座標をそれぞれ$\tilde{X}(X,Y,Z_i,1)$、$\tilde{m}(u,v,1)$、$\tilde{m_1}(u_1,v_1,1)$、$\tilde{m_2}(u_2,v_2,1)$とする．幾何学的関係から

$$\tilde{m}\sim A(R|t)\tilde{X}\cdots(3.4.1)$$

$$\tilde{m_1}\sim A(R|t_1)\tilde{X}\cdots(3.4.2)$$

$$\tilde{m_2}\sim A(R|t_2)\tilde{X}\cdots(3.4.3)$$

となる．ここで A はカメラの内部パラメータを表す3×3の行列である．R はワールド座標系からそれぞれのカメラ座標へ変換する3×3の回転行列であるが、今回は単位行列である．t はワールド座標系からそれぞれのカメラ座標への平行移動を表す3×1の行列である．〜は両辺が定数倍の違いを残して等しいことを表す．図 5 の幾何学的関係から、(3.4.1)式は

$$\begin{bmatrix}u\\v\\1\end{bmatrix}\sim\begin{bmatrix}\frac{f}{\delta}&0&0\\0&\frac{f}{\delta}&0\\0&0&1\end{bmatrix}\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\end{bmatrix}\begin{bmatrix}X\\Y\\Z_i\\1\end{bmatrix}\cdots(3.4.4)$$

となる．ここで画素の物理的な間隔は x 軸方向、y 軸方向いずれにおいても δ としている．f は焦点距離である．これを計算して

$$\begin{cases}u=\frac{f}{\delta}\frac{X}{Z_i}\\v=\frac{f}{\delta}\frac{Y}{Z_i}\end{cases}\cdots(3.4.5)$$

が求まる．同様にして(3.4.2)式、(3.4.3)式から

$$\begin{cases}u_1=\frac{f}{\delta}\frac{X}{Z_i+d}\\v_1=\frac{f}{\delta}\frac{Y}{Z_i+d}\end{cases}\cdots(3.4.6)$$

$$\begin{cases}u_2=\frac{f}{\delta}\frac{X}{Z_i-d}\\v_2=\frac{f}{\delta}\frac{Y}{Z_i-d}\end{cases}\cdots(3.4.7)$$

がそれぞれ求まる．(3.4.5)、(3.4.6)、(3.4.7)から(u,v)と(u_1,v_1)、(u_2,v_2)との関係が以下のように求まる．

$$\begin{cases}u=\frac{Z_i+d}{Z_i}u_1\\v=\frac{Z_i+d}{Z_i}v_1\end{cases}\cdots(3.4.8)$$

$$\begin{cases}u=\frac{Z_i-d}{Z_i}u_2\\v=\frac{Z_i-d}{Z_i}v_2\end{cases}\cdots(3.4.9)$$

(3.4.8)式はm_1地点の画像上(u_1,v_1)におけるピクセルが表す物体がZ_iの距離にあると想定した場合、m地点における画像上でその物体の座標が(u,v)となることを表している．これを全てのピクセルに関して行う．これにより m 地点において 2 枚の画像M_i^1、M_i^2が作成される．

さらにこの操作を全ての$Z_i(i=1,2,\cdots,n)$に対して行う．これにより $2n$ 枚の画像$M_i^j(i=1,2,\cdots,n;j=1,2)$が作成される．画像$M_i^j$の各ピクセル$m_i^j(u,v)$は色情報のほか、3.3 で求めたクラス情報を保持している．

3.5 マッチングによる出力

マッチングは 2 段階で行う．第 1 段階において、sweep-plane 法に馴染まない道路、路面表示、空を独自に求める．第 2 段階において、道路、路面表示、空以外のマッチングを行う．具体的には 3.4 で求めたM_i^jに対し、$i=1$から順に、M_i^1とM_i^2とのマッチングをピクセルごとに判定していく．判定には色情報とクラス情報を用いる．$i=2$以降はマッチングしたと判定されていないピクセルに対してのみ行う．

4. 今後の課題

特に 2 つの点で大いに改良の余地があると思われる．1 つ目は最後のマッチングの段階において領域の特性をいかすことである．同じ領域であれば同程度の距離にあると思われる．これらを同一に扱うような改良が考えられる． 2 つ目は同じくマッチングにおいて、SegNet で得られた意味的情報を十分に生かすことである．例えば、pole と判定された電信柱の位置はおお

よそ道路の端である．このような典型的な位置関係をいかすことは精度の向上に貢献すると思われる．あるいは街路樹はどれも似ているので、マッチングの際に別の木とマッチングされないような工夫が必要である．

5. おわりに

本研究では 2 地点から撮られた風景画像から新たな視点における風景画像を生成する新たな手法を提案した．これはアルゴリズムとディープラーニングを組み合わせたものである．これによりこれまでにない新たな仕組みを提示できた．しかし先に述べたように改良の余地は多い．またこの手法の優位性を示すには、精度、処理速度両面における定量的な評価が必要となる．

参 考 文 献

(1) J. Long, E. Shelhamer, and T. Darrell: “Fully convolutional networks for semantic segmentation”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3431-3440 (2015)

(2) V. Badrinarayanan, A. Kendall, R. Cipolla: “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling”, arXiv:1505.07293 (2015)

(3) P. O. Pinheiro, R. Collovert, P. Dollar: “Learning to Segment Object Candidates”, arXiv:1506.06204 (2015)

(4) S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P. H. S. Torr: “Conditional Random Fields as Recurrent Neural Networks”, arXiv:1502.03240 (2015)

(5) J. Flynn, I. Neulander, J. Philbin, N. Snavely: “DeepStereo: Learning to Predict New Views from the World’s Imagery”, arXiv:1506.06825 (2015)

(6) DeepStereo: Learning to Predict New Views from the World’s Imagery – YouTube, https://www.youtube.com/watch?v=cizgVZ8rjKA

(7) G. Chaurasia, S. Duchene, O. Sorkine-Hornung, G. Drettakis: “Depth synthesis and local warps for plausible image-based navigation”, ACM Transactions on Graphics, 32. (2013)

(8) O. Woodford, I. Reid, P. Torr, A. Fitzgibbon: “On new view synthesis using multiview stereo” (2007)

(9) R. T. Collins:”A Space-Sweep Approach to True Multi-Image Matching”, Technical Report 95-101, Computer Science Department, Univ. of Mass (1995)

(10) D. Comaniciu, P. Meer:”Mean Shift: A Robust Approach Toward Feature Space Analysis”, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no.5, pp. 603-619 (2002)

(11) Y. Boykov M-P. Jolly: “Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images”, Proc. of IEEE International Conference on Computer Vision, I, pp. 105-112 (2001)

(12) R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Susstrunk:”SLIC Superpixels Compared to State-of-the-art Superpixel Methods”, PAMI (2012)