静止画像における奥行き予測のための空間情報抽出

Extracting image features for estimating depth information in static image

鈴木遵自（大阪大）　荻野正樹（関西大）　浅田稔（大阪大）

Jyunji Suzuki, Osaka University, jyunji.suzuki@er.ams.eng.osaka-u.ac.jp
Masaki Ogino, Kansai Univeristy, ogino@res.kutc.kansai-u.ac.jp
Minoru Asada, Osaka University, asada@ams.eng.osaka-u.ac.jp

Human is thought to learn to extract image features as important cues for depth estimation in the developmental process. In this paper, we make a hypothesis that pictorial depth cues are acquired so that disparities can be predicted well and make a model that extracts features appropriate for depth estimation from static images. The experiments with simulation and real environments show high correlation between estimated and real disparities.

Key Words: pictorial depth cues, visual learning, depth estimation, random forest

1. 緒言

ヒトが外界から受ける情報の多くは五感を通して受け取られ，中でも最も情報量が多いのは視覚である．我々は普段ほとんど意識することなく，様々な物体の立体的な形や空間内での位置を把握している．ヒトの視覚システムは，2次元平面として網膜上に投影された像から3次元世界を再構成する，という逆問題を，両眼視差，運動視差，眼球運動の情報などといった様々な奥行知覚の手がかりを用いて瞬間的に解くことができる．

さらに，ヒトは2次元平面上に描かれた絵画や写真などからも，その描かれた対象の奥行きを知覚し，立体感を得ることができる．このときに手がかりとなるのは，絵画的手がかりと総称される，物体の相対的大きさ，上下関係，遠近法などといった，単一の静止画像から得られる手がかりである．絵画的手がかりはそれぞれ意味づけされ，分けて考えられることが多い．しかし実際にはこれらの手がかりは，ヒトが視覚情報を統合する過程で，情報の信頼性に依存して重み付け平均化が行われることで，重要な情報の組み合わせとして経験的に得られるものであり[1][2]，その形成には豊富な視覚経験が不可欠である．また，絵画的手がかりのほとんどは文脈情報であり，いくつかの局所的な情報を組み合わせ，統合することで，奥行きを予測する手がかりとして機能する．

外界からの視覚入力は網膜，外側膝状体を通り，脳の後頭葉の最後方に位置する第一次視覚野(V1野)と呼ばれる神経組織に到達する．視覚神経系は並列階層的に構成された多くの領野から成り立っており，低次の領野から高次の領野に向かうに従い，ローカルな特徴からよりグローバルな視覚パターンの検出が行われると考えられている．最下層のV1野には，特定の方位を持った線分やエッジに選択的に反応する細胞があることが知られており[3]，V1野の上の階層にあたるV2野では，2本の線を組み合わせた十字や角などの刺激[4]，さらに高次の視覚野であるIT野と呼ばれる領野では，より複雑な図形に応答する細胞の存在が報告されている[5]．このような二次元形状の構造化と同様に，空間の認識に関しても，低次の領野で情報が抽出され，高次の領野で文脈情報として統合されることで，面などの高次の概念が構造化されていると考えられる．絵画的手がかりは，そのような高次の空間表現の基盤として抽出される情報である．

平面上に描画された平面画像からの3次元形状の復元という問題は，コンピュータビジョンの分野での主要なテーマの一つである．特に単一静止画像を対象とした近年の研究として，Bayesian NetworkやMRF(Markov Random Field)といった確率モデルを利用して奥行きを推定するモデルが提案されており[6][7][8]，確率モデルによって，静止画像を局所領域に分割し，局所領域間の条件付確率によって文脈情報を表現することで，局所領域毎の平面パラメータを推定している．これらのモデルは空間の構成要素が平面であるということを前提としており，3次元形状の復元という問題に焦点が当てられている．一方で，奥行知覚の手がかりの形成に焦点を当てた場合，これは奥行知覚の学習過程で得られるものであるため，空間の構成要素について前提を設けるのは適切ではない．奥行知覚の手がかりの形成という問題について，手がかりに対する事前知識を与えないボトムアップなアプローチを試みている研究例は少ない．

本研究では，奥行知覚の手がかりの形成メカニズムの理解を目指し，静止画像から主要な空間情報を抽出することを目的とする．静止画像からの奥行き予測をタスクとして，静止画像に含まれる画像特徴の中から，奥行き予測のための主要な情報を抽出し，それをもとに奥行き予測を行う学習モデルを提案する．モデルでの学習によって，奥行き知覚の絵画的手がかりとして説明可能な情報が抽出され，奥行きの予測に大きく寄与していることを示す．

2. 学習モデル

本研究で提案する学習モデルについて，Fig. 1にその概要図を示す．モデルでは，静止画像からの奥行き予測をタスクとして，静止画像の画像特徴の中から，主要な

情報を抽出し，それをもとに奥行き予測を行う予測器を構築する．入力画像から各点の画像特徴を求め，別途用意した，入力画像の奥行き情報を目的変数として，回帰分析を行う．学習の際に用いる説明変数を，

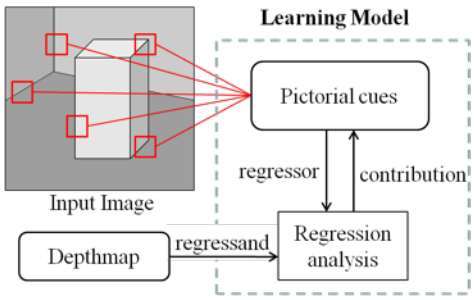


Fig. 1 Overview of the learning

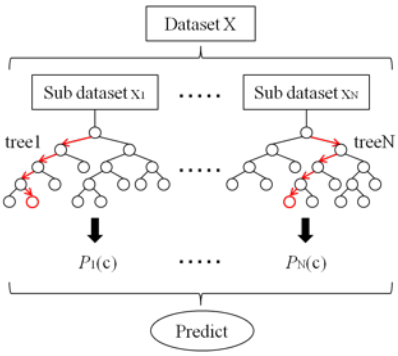


Fig. 2　Random forests

各点の画像特徴の中から回帰への寄与度を基準に選定する．抽出された奥行き予測のために主要な画像特徴の組み合わせとして，奥行知覚の手がかりが表現される．

予測器毎に，ランダムに100個のベクトル，u1～u100を設定する．奥行きを予測する2次元平面上の注目点を点xとして，点xにおける説明変数は，ヒトのV1野の細胞で検出される特徴量[3]を参考に，点xの座標(x,y)，及びx + u1～x+ u100の100点のエッジの強さと方向の2次元特徴量の，計202の変数を用いる．奥行き情報は，3次元空間における視差（運動視差または両眼視差）から得る．視差の大きさは対象までの距離に依存するため，奥行きに相当する情報として利用できる．視差の大きさとして得られる，点xにおける奥行きを点xの目的変数とする．以上，説明変数202と目的変数1の計203の変数を点xにおける入力データとし，1画像につきxを20点ランダムサンプリングする．大量の画像に対してこのデータセットを用意する．

学習にはFig. 2に示すRF(random forests)[9]を利用する．RFは多数の決定木を用いた集団学習である．RFは入力データを多数のサブサンプル集合に分け，各サブサンプル集合から決定木を生成し，その結果を統合することで予測結果を得る．決定木での分岐は分岐前後のデータ純度を基準とし，よりデータの純度が高くなるよう，分岐に用いる変数とその閾値が決定する．RFは予測精度が高く，様々な特徴を持つ．本研究では特に，説明変数の重要度を算出できる点に着目し，奥行き予測に大きく寄与する変数として奥行知覚の手がかりが表現されることを示す．

3. 実験

実験はシミュレーション空間と実環境で行った．共通の設定として，画像入力は400平方画素の濃淡画像であり，中央の300平方画素の範囲を奥行き予測の対象とする．エッジ情報は27画素平方の範囲から計算し，また視差が得られるのは特徴点についてのみであるため，対象範囲を6×6の格子状に区分し，点xの奥行きは，属する格子内に存在する全特徴点で得られた視差の大きさの平均値とする．学習データは3万点，テストデータは未学習のものから1万点をそれぞれランダムサンプリングする．

3.1 シミュレーション環境における実験

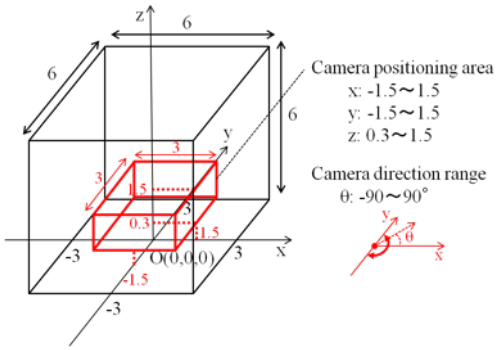


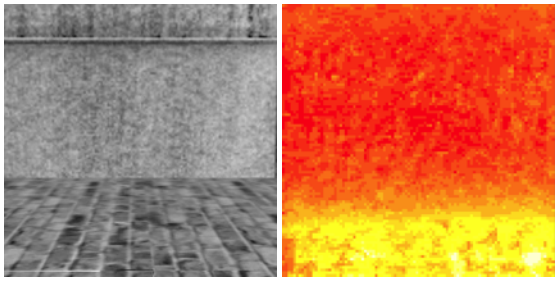
Fig. 3　Simulation environment

シミュレーション環境は，Fig.3に示すような各辺6の立方体空間内であり，床面及び内壁にはテクスチャが施されている．床面中心点の座標を(0,0,0)とし，水平方向にx軸及びy軸，鉛直方向にz軸をとる．カメラ視点の位置を，x及びy座標を-1.5～1.5，z座標を0.3～1.5の範囲，カメラ視線の水平方向を-90°～90°の範囲でランダムに決定する．カメラ視線の上下方向の角度は水平で固定，カメラの画角は標準レンズに相当する50°とする．視差は運動視差を用いる．ランダムに決定したカメラ位置から，カメラ視線方向に対して垂直平面上で移動距離1の並進運動を行い，視差を得る．

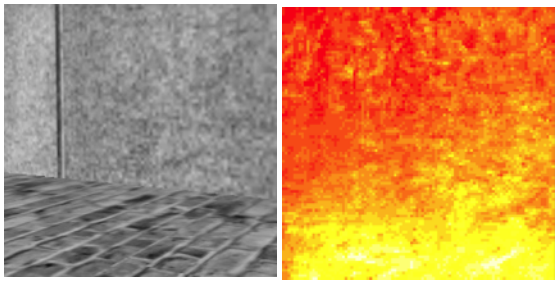
学習の結果，テストデータの奥行き情報とその予測値との相関係数が0.8115程度の高い相関を示す予測器が構築された．学習の結果を視覚的に捉えるため，得られた予測器に対し，未学習データにあたる1画像の全点を入力として与え，出力として得られた各点の奥行き予測値を2色画像に変換したものを，奥行き予測画像として，Figs. 4 (a),(b)に示す．左が元画像のグレイスケールイメージ，右が作成した奥行き予測画像であり，黄色が近距離，赤色が遠距離を表している．床面や，壁面の傾きの傾向が表れている．　また，それぞれの画像についての

予測精度を，次式で求められる平方平均二乗誤差率RMSPE(Root Mean Square Percentage Error)によって評価した結果を併記して示す．

$$RMSPE = \left(\sqrt{\frac{1}{n} \sum_{t=1}^{n} \left\{ (\hat{X}_t - X_t) / X_t \right\}^2} \right) \times 100$$



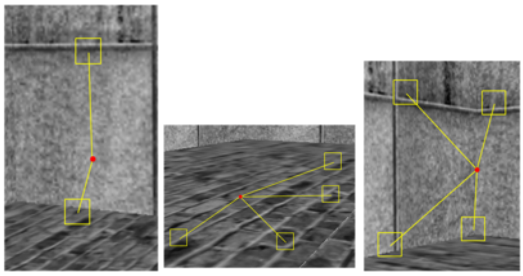
Original image　　Predicted image
(a) RMSPE: 11.96%



Original image　　Predicted image
(b) RMSPE: 12.21%

Fig. 4 Input images and estimated depth maps in simulation environment

学習データのいくつかについて，各説明変数の重要度を算出した．特徴量として，エッジの強さと方向，1点につき2変数を与えているため，2変数の重要度の相乗平均をとり，その点の重要度とした．各入力に対し，算出された100点の重要度のうち，上位5点（5%）以内のものを，重要度の高い特徴点とみなす．重要度の高い変数の組み合わせが，奥行知覚の絵画的手がかりとして説明できる例をFigs. 5(a)〜(c)に示す．赤色の点が奥行きを予測する注目点であり，注目点から黄色い線を伸ばした先が重要度の高い特徴点であり，それを囲う黄色い正方形で示した範囲が，その点で特徴量としてエッジを取得した範囲を示している．Fig. 5(a)は，壁面と床面の境界線上と，その境界線に平行な壁面テクスチャの直線上の点が，重要度の高い点の組み合わせとして抽出されている，線遠近法の手がかりとして説明できる例である．平行線の傾きとその間隔が，奥行きの手がかりとして機能していると考えられる．Fig. 5(b)は，床面テクスチャ上の点について，特にその勾配方向（画像上下方向）での複数の組み合わせがみられ，テクスチャ勾配の手がかりとして説明できる例である．エッジ情報，即ち輝度勾配から，テクスチャの模様だけでなく，テクスチャの粗密に相当する情報が得られるため，その勾配が奥行きの手がかりとなっていると考えられる．



(a)　　(b)　　(c)

Fig. 5　Examples of important image features

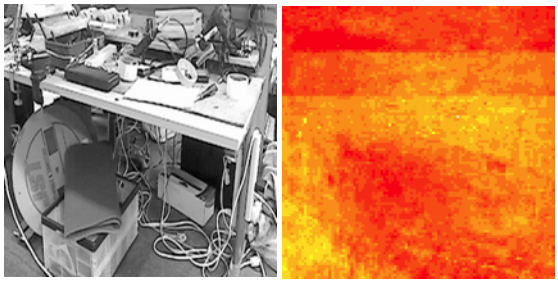
Fig. 5(c)には既知のものの大きさが奥行き手がかりとして機能していると見受けられる例を示した．隅の壁面テクスチャの，特にその角にあたる部分に，重要度の高い点が多く見られる．そのテクスチャの視えの大きさと奥行き情報とを関連付けて学習し，手がかりとして利用しているものと考えられる．

3.2 実環境実験

実環境は明るい室内環境とし，視差として両眼視差を利用する．左右平行に10cmの間隔で固定した2台のカメラを用いて画像を取得し，視差画像を得る．

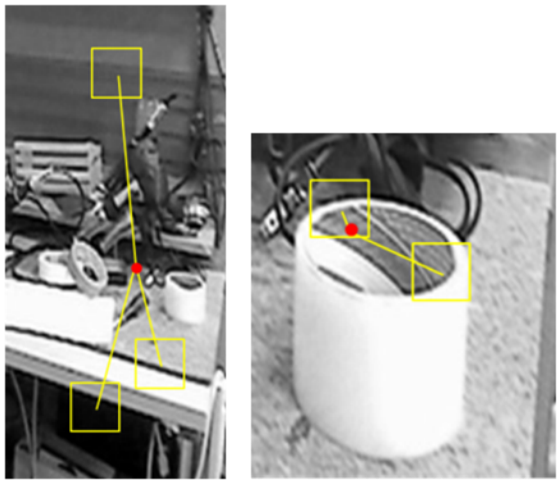
学習結果，テストデータの相関係数は0.7933となり，シミュレーションには及ばないものの，高い相関を示す予測器が構築された．シミュレーションと同様に，作成した予測画像とその入力元画像を，予測精度を平方平均二乗誤差率RMSPEによって評価した結果と共にFig. 6に示す．上方に視差が生じにくい遠方の景色が映ることが多かったため情報量が少なく，いずれも画像上方の予測が段階的になってしまっているが，デスクの端の傾きなどの大まかな奥行きの傾向は掴めている．

また，シミュレーション同様，説明変数の重要度を算出し，絵画的手がかりとして説明できるものをFigs. 7(a),(b)に示す．シミュレーション同様に，絵画的手がかりとして説明できる特徴点が，奥行き予測に重要な変数の組み合わせとして表れた．特に，Fig. 7(b)では，シミュレーション環境には存在しない，円形の物体について，その直径に相当する情報が得られる特徴点の組み合わせが表れた．円形物体の直径は，視点位置によらず距離のみに依存するため，その大きさを学習する機会が十分豊富であったものと予想される．



Original image　　Predicted image
RMSPE: 14.44%

Fig. 6 Input images and estimated depth maps in simulation environment



(a)　　(b)
Fig. 7　Examples of important image features

4.　結論と考察

静止画像からの奥行き予測をタスクとして，静止画像の画像特徴の中から，奥行き予測のための主要な情報を抽出し，それをもとに奥行き予測を行う学習モデルを提案した．提案モデルでの学習によって，静止画像の画像特徴から，線状透視，テクスチャ勾配，既知のものの大きさなどに相当する情報が抽出され，それらの情報が奥行きの予測に大きく寄与し，絵画的手がかりとして機能していることを示した．

実験結果として図示した絵画的手がかりの他に，上にあるのものほど遠く，下にあるものほど近く感じるという，上下関係の手がかりを挙げられる．実験では，説明変数として注目点の画像内の座標を与えている．各点の特徴量の重要度比較では，あくまで特徴量の中での重要度を比較したが，説明変数全体で見たとき，作成したどの予測器においても，最重要な変数は注目点のy座標であった．y座標に対して，x座標の重要性は他の変数と比べても目立ったものではないことから，単に画像内の座標が重要な要素なのではなく，その上下位置が特に重要であるということがわかる．これは，実験においてカメラ画像の上下方向が実空間の鉛直方向となるよう固定し，また床面を基準とした環境で学習を行ったため，「下方に地面がある」という一般知識を学習できた結果であるといえる．

今後の課題として，提案モデルでは，奥行き予測に重要な情報を抽出することができるが，その抽出された情報間の関係性を明確にできていない．その関係性をみることで，得られた情報の絵画的手がかりとしての普遍的構造を明らかにすることができると考えられる．予測結果についても，提案モデルでは各点の奥行き予測は独立したものである．今回は，絵画的手がかりの抽出を目的としたが，将来的には，今回のモデルを基盤とした，1画像中の各点の奥行きを他の点の予測にフィードバックできる予測器の構築によって，より精度の高い予測が実現できると予想される．その際に，本研究で得られた絵画的手がかりの抽出機構を入力の基盤とすることで，入力情報のより普遍的な表現が可能となり，予測に大きく貢献できると期待される

参考文献

[1] J. Fagot, I, Barbet, and C. Parron, "Amodal completion by baboons(Papio papio): contribution of background depth cues". Primates, 47, 145-150, 2006.

[2] M. S. Landy, L. T. Maloney, E. B. Johnston, and M. Young, "Measurement and modeling of depth cue combination: in defense of weak fusion". Vision Research, 35, 389-412, 1995.

[3] D. Hubel and T. Wiesel, "Receptive fields of single neurones in the cat's striate cor-tex". Journal of Physiology, 148, 574-591, 1959.

[4] J. Hegde and D. C. Van Essen, "Selectivity for complex shapes in primate visual area V2". J. Neurosci. 20:RC61-66, 2000.

[5] I. Fujita, K. Tanaka,M. Ito, K. Cheng, "Columns for visual features of objects in monkey inferotemporal cortex". Nature, 360: 343-346, 1992.

[6] E. Delage, H. Lee, and A. Y. Ng. "A dynamic Bayesian network model for autono-mous 3d reconstruction from a single indoor image". Computer Vision and Pattern Recognition (CVPR), 2006.

[7] A. Sexena, S. H. Chung, and A. Y. Ng. "Learning Depth from Single Monocular Images". Neural Information Processiong Systems (NIPS), 18, 2005.

[8] A. Saxena, M. Sun, and A. Y. Ng. "Make3D: Learning 3D Scene Structure from a Single Still Image". IEEE Transactions of Pattern Analysis and Machine Intelligence (PAMI), 2008.

[9] L. Breiman. "Random forests". Machine Learning, 45, 5-32, 2001.