M-002

Local Dynamic Map の利用を想定した車載での壁透視表示向け歩行者画像生成法
A Multi-view Pedestrian Image Generation Method for Practical Wall See-through Systems Using Local Dynamic Map

大濱 吉紘† 小島 真一† 後藤 邦博†
Yoshihiro Ohama Shin-ichi Kojima Kunihiro Goto

1. はじめに

近年、5G 無線通信システムの技術仕様の策定の進展とともに、ITS 技術による高度運転支援及び自動運転サービスを支える新たな情報処理技術として、車車間及び路車間での情報通信による Local Dynamic Map (LDM)システムの議論が活発に行われている[1][2]。LDM システムでは、交差点等に設置された監視カメラや車載のレーダ及びカメラといった多数のセンサ群の計測情報を比較的狭い範囲で集約し、構造化された時空間的データである LDM として管理する。より具体的には、交通環境中に存在する道路・建物のような静的な地理情報、信号機や交通規制のような準静的な制御情報、道路の混雑具合のような準動的な情報、可動物体の位置・速度や属性のような動的な情報の 4 階層で LDM を構成する仕様が策定されつつある[2][3]。この中で、最も情報の鮮度を求められる第 4 層(動的情報)においては、実世界に対する遅れが 1 秒未満というリアルタイム処理が求められている[4]。これまでに我々は、この LDM の第 4 層の技術要件を満たす技術構成を見出すことに焦点を当て、WiFi の複数のアクセスポイントを低遅延で切り替えながら実車上の車載センサを接続する、プロトタイプ・システムを実装した[5]。このプロトタイプ・システム上で、さらにロボティクスやサーベイランス・システムの技術分野における複数物体追跡技術を、複数台の車載カメラを用いてリアルタイム処理可能な形態で実装し、技術検討を開始している[6]。

高度運転支援における LDM の応用例の 1 つとして、壁などに隠ぺいされた歩行者や車両などを透過した映像を提供する、壁透視表示システムがあり得る[7][8][9]。従来からの壁透視表示システムは、映像のリアルタイム伝送が可能であることを想定している。しかし近い将来、5G 通信システムの実用化が予想されるものの、壁透視表示システムのためだけに、市街地の交差点で多数のノード間でリアルタイムに映像を伝送することは、無線通信帯域の有効な利用形態とは言い難い側面がある。そこで本稿では、高々数十バイトのデータ転送によっても、車載の情報提示装置(HMI; Human-Machine Interface)として効果のある画像を生成可能な手法の確立を目的とする。ここでは特に歩行者画像に焦点を当て、図 1 に示すように、あるカメラから隠ぺいされた歩行者の画像を、隠ぺいされていない別のカメラで捉えた歩行者画像から生成する、視点変換の問題設定を考える。既に我々は、交差道路が隠ぺいされた交差点を自動車に搭乗して直進通過する場面を想定した心理実験を行い、HMI に求められる要件の技術検討を報告している[10]。この中で、壁透視表示による搭乗者の衝突判定の反応時間には、表示画像に交差対象物の向き情報が含まれるか否かで有意差がみられることを確認している。

†株式会社豊田中央研究所, Toyota CRDL, Inc.

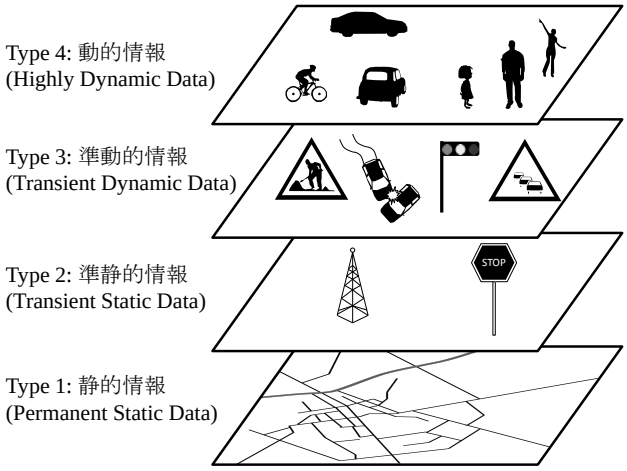


図 1　LDM の構成要素の概略図

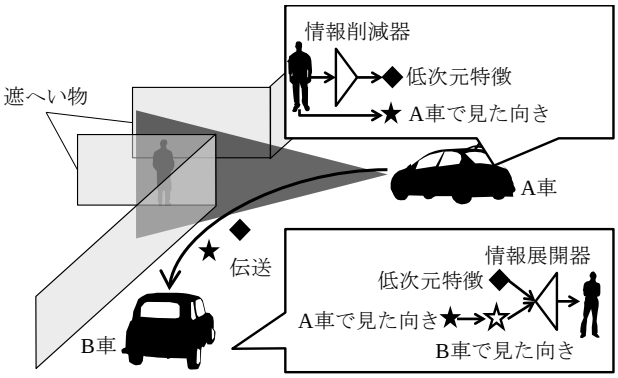


図 2　CVAE による壁透視表示の構成法の概略図

本稿では、深層学習器の 1 つである CVAE(Conditional Variational Autoencoder)[11]に、あらかじめ向きの判っている多数の歩行者画像を学習させて、画像を低次元化する情報削減器と、低次元化された画像を復元する情報展開器を構成して、各カメラには情報削減器を、各 HMI には情報展開器を備える、図 2 のようなシステム構成を提案する。この構成により、各カメラは画像中の歩行者検出枠の座標値に加えて、歩行者の向きと CVAE で低次元化された画像、いわゆる特徴量を送出することで、各 HMI の視点からの歩行者画像として復元できるようになる。この歩行者画像から、各車の搭乗者が向き情報を識別できれば、既報[10]より情報提示装置として効果のある HMI が構成できると期待される。そこで本稿では、既存手法[12]で学習された歩行者画像の向き識別器が、この生成画像の向き識別を適切に行えるかどうかにより、妥当性を検証する。

2. 壁透視表示のための歩行者画像の視点変換

ニューラルネットワークにおいて古くから学習データからの特徴量抽出の手法として用いられてきた Autoencoder(AE)を礎に、近年の深層学習の技術進展の中で

Variational Autoencoder(VAE)が提案され、学習済みの学習モデルのボトルネック層を操作することで、多様な画像を生成できることが知られている[13]。さらに、より選択的な情報生成が可能となるように VAE を拡張したものが CVAE である。本節では、AE、VAE 及び CVAE について概説し、壁透視表示のための歩行者の視点変換画像の生成例について述べる。

2.1 情報表現のための学習モデル

AE は階層型ニューラルネットワークにおいて、入力層と出力層の次元が同一で、ボトルネック層と呼ばれる入出力層よりも極端に少ない次元の中間層を少なくとも 1 つ備えた、図 3(a)に示すような形状の学習モデルである。さらに学習時には、入力層への入力データと出力層からの出力データが一致するような学習、いわゆる恒等写像学習を行う。この学習が適切に行われた後に、入力データに対応するボトルネック層の出力（情報表現）には、学習に用いたデータの本質的な特徴が表出することが知られている。このような学習モデルの構造と機能から、入力層からボトルネック層までの変換をエンコーダ、ボトルネック層から出力層への変換をデコーダと呼ぶことが多い。ここで、AE への入力を確率変数x、オートエンコーダの出力をy、エンコーダ部のニューラルネットの結合重みをθ、デコーダ部のニューラルネットの結合重みをφと記述する。また、ボトルネック層の情報表現をzとする。このとき恒等写像学習は、xに関する学習用データ列$X=\{x_1, x_2, \cdots, x_N\}$が与えられたときに、結合重み$\theta$及び$\varphi$を調整する最小二乗問題とすることができる。これは条件付き確率分布$p(X;\theta,\varphi)$に正規分布を仮定した、以下の尤度関数Lに関する、θ及びφの最尤推定問題と等価である。

$$L = \mathrm{E}[\log p(X;\theta,\varphi)] = -\frac{1}{2}\sum_{n=1}^{N}(x_n - y_n)^2 \quad (1)$$

$$\hat{\theta},\hat{\varphi} = \underset{\theta,\varphi}{\operatorname{argmax}} L(X;\theta,\varphi) \quad (2)$$

VAE では、AE に過学習を防ぐような制約を加えて、ボトルネック層での情報表現の質を高める仕組みを導入する。VAE の学習モデルは、図 3(b)のようにエンコーダ部の出力を正規分布のパラメータ（期待値μ、共分散Σ）とし、標準正規乱数をΣで重みづけたものをμに加える変数変換（Reparameterization Trick）を行う。これにより、情報表現zも、エンコーダ部のニューラルネットの結合重みθと、入力データxの条件付き確率分布$q(z|x;\theta)$に従う確率変数とみなすことができる。このとき、zが標準正規分布に従うとした尤度関数の変分下界L_{min}は、次式のようになる[13] 。

$$\begin{aligned} L_{min} &= -D_{KL}\big(q(z|x;\theta)\|p(z)\big) \\ &\quad + \mathrm{E}[\log p(x|z;\theta,\varphi)] \\ &= -\frac{1}{2}\sum_{n=1}^{N}[tr\Sigma(x_n;\theta) + \mu(x_n;\theta)^T\mu(x_n;\theta) \\ &\quad - d_z - \log\Sigma(x_n;\theta)] \\ &\quad + (x_n - y_n)^2 \end{aligned} \quad (3)$$

ここで、$D_{KL}(\cdot\|\cdot)$はカルバック・ライブラー・ダイバージェンス（KL 距離）、E[·]は期待値、d_zはzの次元数である。すなわち、式(3)の第 1 項の KL 距離はzに関する正則化項、第 2 項は式(1)と同様に恒等写像の精度を意味する。

様々な歩行者画像で構成された学習用データ列を AE や VAE で学習した場合、情報表現zには歩行者の向きだけでなく、全体の形状、衣服、所持品など多様な情報が畳み込まれている。そのため、学習後の AE や VAE のデコーダ部を取り出して、所望の向きの歩行者画像を生成することは容易ではない。しかし、あらかじめ学習用の歩行者画像に、向きをラベルとして付与することは可能である。CVAE では、各学習データに対応するラベルlが付与されている学習用データ列$X=\{(x_1,l_1),(x_2,l_2)\cdots,(x_N,l_N)\}$に対して、図 3(c)のような学習モデルを構成する。ラベルlはエンコーダ部とデコーダ部に入力され、情報表現zにはラベルlを除く特徴が表出することが期待される。また、デコーダ部を取り出してデータを生成する際には、所望のラベルlに対応する様々なデータを生成することができる。このとき、CVAE の尤度関数の変分下界L'_{min}は、次式のようになる。

$$\begin{aligned} L'_{min} &= -D_{KL}\big(Q(z|x,l;\theta)\|P(z)\big) \\ &\quad + \mathrm{E}[\log P(x|z,l;\theta,\varphi)] \\ &= -\frac{1}{2}\sum_{n=1}^{N}[tr\Sigma(x_n,l_n;\theta) \\ &\quad + \mu(x_n,l_n;\theta)^T\mu(x_n,l_n;\theta) \\ &\quad - d_z - \log\Sigma(x_n,l_n;\theta)] \\ &\quad + (x_n - y_n)^2 \end{aligned} \quad (4)$$

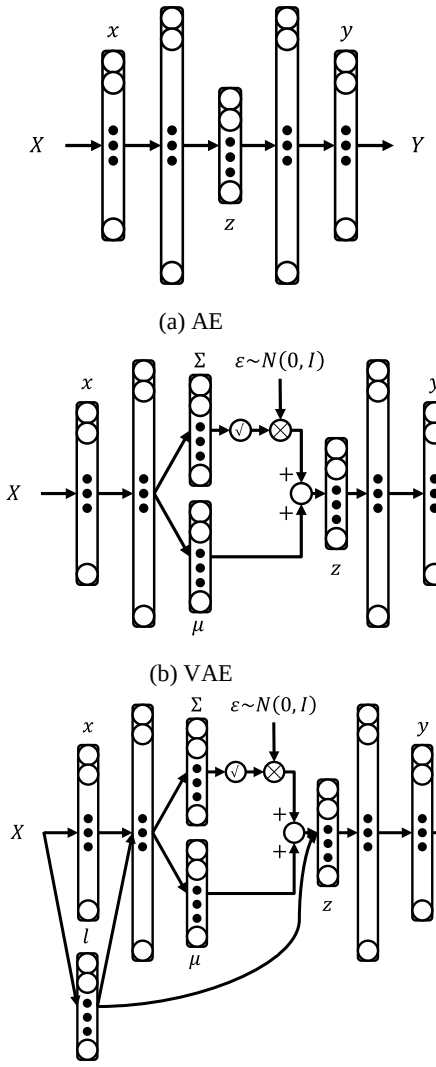


図 3　学習モデルの構成の比較

2.2 CVAE による歩行者の視点変換画像の生成

図 2 のように、歩行者が隠ぺいされていない視点から取得したカメラでの歩行者画像を、別の視点からの歩行者画像に変換して壁透視表示を実現するために、CVAE による視点変換画像の生成が可能かどうかを確かめる。CVAE を学習させるための歩行者画像データセットとして、Parse27k[14]を用いることにした。Parse27k は 27,081 枚の歩行者画像からなり、全画像に 4 方向、8 方向、16 方向の向きラベルも付与されている。このデータセットから 20,000 枚の画像をランダムに選び出し、32x32 ピクセルにリサイズして学習データとした。CVAE を構成する学習モデルは、図 3(c)の構成とし、画像の入出力層を 3,072 次元の線形関数、入力層の直後の層及び出力層の直前の層を 2,048 次元の ReLU 関数[15]とした。また、歩行者向きは、正面、右正面、真右、右背面、背面、左背面、真左、左正面の 8 方向の one-hot 表現とし、8 次元のラベル入力層を構成した。本節では CVAE の利用可能性の確認が目的であり、ボトルネック層は解釈を容易にするために 2 次元とした。

100 枚の歩行者画像を単位バッチとするミニバッチ学習を 100 エポック行った後に、ボトルネック層の 2 つの線形ニューロンを、値域[−2.0,2.0]を0.1刻みで格子分割して手動設定して得た歩行者画像を、図 4 に示す。図 4 では、ボトルネック層の格子分割と、生成画像の格子分割は対応付けて描かれている。この結果から、z_1は生成画像の明るさを、z_2は歩行者の衣服の輝度を表現していることがわかる。

次に、学習に用いなかった 7,081 枚の画像から白っぽい衣服の歩行者画像と黒っぽい衣服の歩行者画像を各 1 枚ずつ選んで CVAE に入力後、8 方向の歩行者向きラベルをボトルネック層に入力して得た、視点変換画像を図 5 に示す。いずれの生成画像も、入力された歩行者画像の衣服の特徴を保存したまま、指定した向きの画像を生成できていることがわかる。AE による生成画像はぼやけたものとなることは広く知られているが、歩行者画像においても図 5 からもわかるとおり、背景除去の効果はあるものの、頭部及び上肢のテクスチャも平均化されてしまっていることがわかる。しかし本稿の目的は、HMI としての効果が見込まれる画像の生成であり、搭乗者の衝突判定の反応時間に影響を与えた向き情報を、生成画像から推測可能であるかどうかが重要となる。

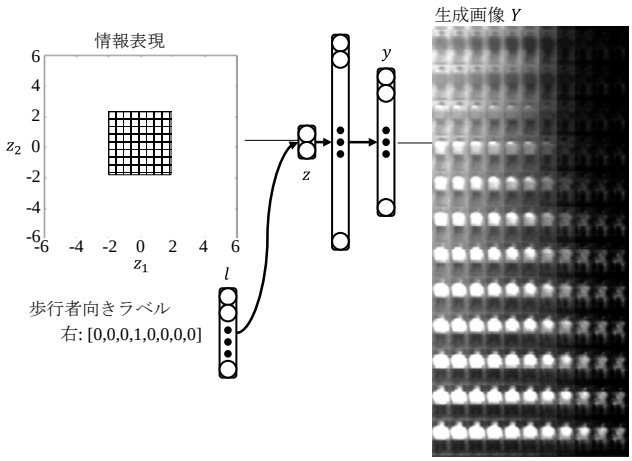
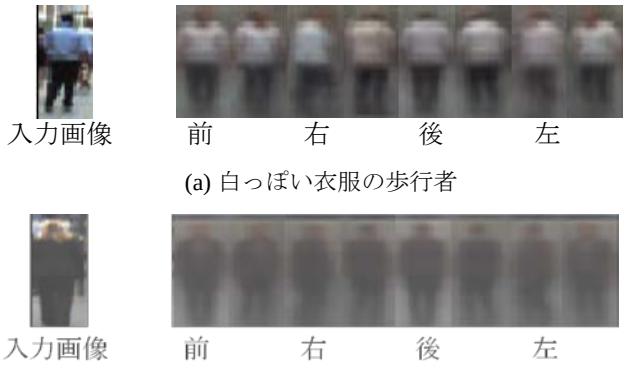


図 4　CVAE による右向き歩行者画像の生成



(a) 白っぽい衣服の歩行者

(b) 黒っぽい衣服の歩行者

図 5　向きを指定した歩行者画像の生成例

3. 歩行者の視点変換画像の数値実験による評価

CVAE による歩行者の視点変換画像が向き情報を含んでいるかどうかを、歩行者画像の向き識別器を用いた網羅評価によって定量化する。歩行者画像の向き識別には、いくつかの既存手法が存在する[16]。本稿の目的上、歩行者向き識別器の性能自体は重要ではなく、撮影した歩行者画像と生成した歩行者画像との間で、向き識別結果の差を定量化できれば良い。そこで、典型的な手法である歩行者画像から HoG 特徴量を生成して SVM によって識別する手法[12]を用いることにする。

3.1 評価方法

本節で利用する歩行者画像データセットは、2.2 節と同様に Parse27k を用いることとし、ランダム選択によって 20,000 枚の学習データと 7,081 枚の評価データとに分割した。各画像には 4 方向、8 方向、16 方向の向きタグが付与されているが、16 方向のタグを用いると、1 方向あたりのデータ数が極めて少なくなるため、ここでは 4 方向と 8 方向のタグのみを用いることにした。

評価に先立って、CVAE と歩行者向き識別器を、学習データによって十分に学習させる。ここで、CVAE のボトルネック層は 16 次元の線形関数とし、100 枚の歩行者画像を単位バッチとするミニバッチ学習を 100 エポック行った。また、HoG 特徴量は 6,804 次元とし、4 方向及び 8 方向の識別を行う SVM の最適なパラメータをグリッドサーチによって探索した。

評価にあたっては、評価データ（元画像）を直接 SVM で向き識別を行う一方で、評価データを CVAE に入力して各方向の歩行者画像（視点変換画像）を生成し SVM で向き識別を行った。このとき、向き識別の性能差が少ないほど、視点変換による向き識別のしやすさへの影響が少ないと考えることができる。

3.2 評価結果

識別精度(Accuracy)の評価結果を、歩行者向き識別が 4 方向及び 8 方向の場合について、図 6 に示す。この結果から、歩行者向き識別が 4 方向の場合の精度は、元画像については 0.70、視点変換画像については 0.66 であり、約 14.3%の精度低下であった。いくつかの歩行者向き識別器の性能評価を行った文献[16]においても、Accuracy は概ね

0.7 前後が報告されており、この歩行者画像の向き識別器を本稿の評価に用いることは、妥当と考えられる。一方で、向き識別が 8 方向の場合には、元画像については 0.51、視点変換画像については 0.43 であり、約 15.7%の精度低下がみられた。ただし 8 方向の場合には、元画像も視点変換画像も精度自体が低くなっており、HoG 特徴量を用いた SVM での多クラス識別自体が、歩行者画像については、そもそも困難であったという可能性もある。

さらに歩行者向き識別が 4 方向の場合について、適合率(Precision)と再現率(Recall)を算出した結果を、元画像について図 7(a)に、視点変換画像について図 7(b)に示す。まず元画像の結果である図 7(a)から、前後方向と左右方向では明らかに差が出ており、データセット中の歩行者向き画像に、偏りがあったことがわかる。次に視点変換画像の結果である図 7(b)から、Precision と Recall のバランスに偏りがみられ、左右方向に識別できたものは正しいものの、多くの左右向き画像が識別から漏れていることがわかった（Precision が高く、Recall が低い）。一方、後ろ方向に識別したものは半分程度、誤識別であった(Precision が低く、Recall が高い)。ただし、前方向の識別は概ね正しかった。この結果から、CVAE は見かけ上は多視点の画像を生成できているが、多くの画像を後ろ向きと識別しやすいような、偏ったデータセットとなっていると考えられる。

4. おわりに

本稿では LDM を活用し、車載の壁透視表示 HMI のための歩行者画像の視点変換法を提案した。そして、視点変換画像が生成でき得ることを、オープンデータセットを用いて確かめた。壁透視表示 HMI において、搭乗者が衝突判定に要する反応時間は、衝突対象物の向き情報に影響されるという既存知見がある。そこで、元画像と視点変換画像との間で、歩行者画像の向き識別器の精度比較を行った。その結果、視点変換画像では 15%前後の精度低下にとどまった。一方で、データセットの偏りから、後ろ向きの歩行者画像と識別される画像が多く生成されていることが示唆された。また、精度低下量が壁透視表示 HMI の効果に与える影響は未検証であり、今後の課題である。

今後は、学習用データセット及び学習モデルの見直しによる生成画像の品質向上、視点変換画像の向き識別の被験者実験による検証とタスク・パフォーマンス評価、実車への搭載について検討を進めていく予定である。

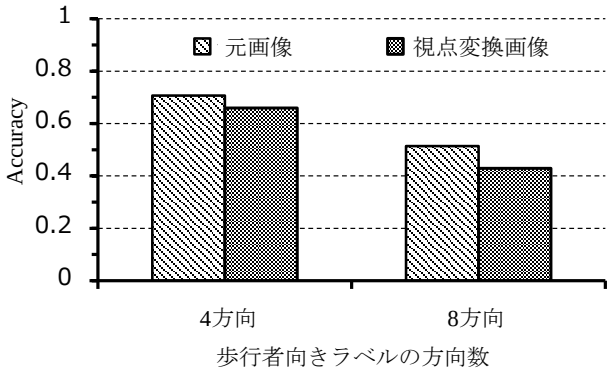


図 6　歩行者画像向き識別の精度の比較

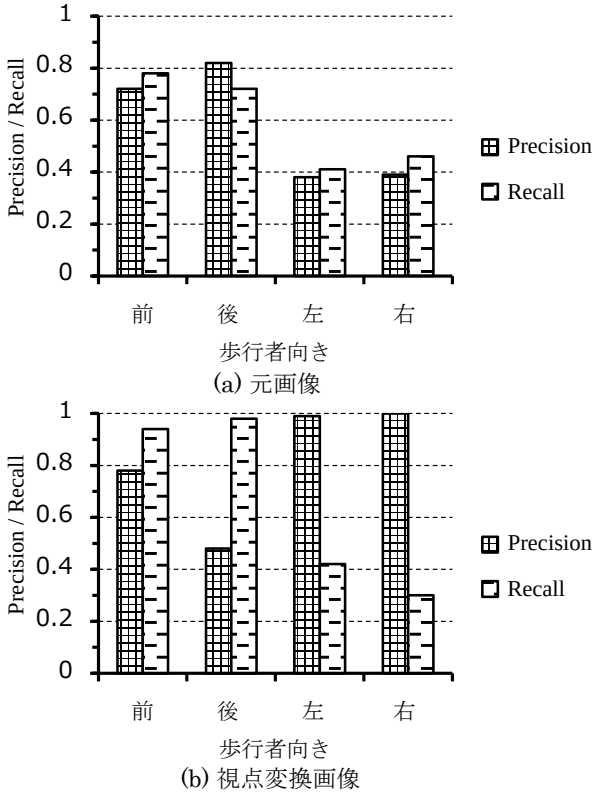


図 7　4 方向識別の場合の適合率と再現率の比較

参考文献

[1] A. Schalk, “The globally applicable concept of a Local Dynamic Map”, 6th ETSI ITS Workshop (2014).
[2] H. Shimada et. al., “Implementation and Evaluation of Local Dynamic Map in Safety Driving Systems”, Journal of Transportation Technologies, Vol.5 (2015).
[3] A. Shwarz, “TEAM Project Presentation”, 1st My Way Collaboration Workshop (2015).
[4] 葛巻清吾, “SIP-adus 進捗報告”, 戦略的イノベーション創造プログラム自動走行システム 第 5 回メディアミーティング(2015).
[5] K. Sasaki et. al., “Robust Low Latency Communication Using Commodity Wireless Networking and Edge Computing”, IEEE GLOBECOM (2018) submitted.
[6] 中村亮裕ら, “ローカルダイナミックマップ生成のための複数カメラを用いたリアルタイム多物体追跡システム” , FIT 講演論文集 (2018).
[7] P. Barnum et. al., “Dynamic seehroughs: Synthesizing hidden views of moving objects”, Proc. IEEE/ACM ISMAR (2009).
[8] C. Sandor et. al., “An augmented reality X-Ray system based on visual saliency”, Proc. IEEE/ACM ISMAR (2010).
[9] T. Tsuda et. al., “Visualization methods for outdoor see-through vision”, IEICE Trans. Information and Systems. Vol.89, No.6 (2006)
[10] 安田浩志ら, “拡張現実感による死角交差点での運転支援：リアルな壁透過表現は必要か?”, 日本バーチャルリアリティ学会論文誌, Vol.19, No.3 (2014)
[11] D. P. Kingma et. al., “Semi-supervised Learning with Deep Generative Models”, Proc. NIPS (2014).
[12] T. Gandhi et. al., “Pedestrian protection systems: issues, survey, and challenges”, IEEE Trans. ITS, Vol.8, No.3 (2007).
[13] D. P. Kingma et. al., “Auto-Encoding Variational Bayes”, Proc. ICLR (2014).
[14] RWTH Aachen, Parse27k, https://www.vision.rwth-aachen.de/page/parse27k (2016).
[15] V. Nair et. al. “Rectified Linear Units Improve Restricted Boltzmann Machines”, Proc. ICML (2010).
[16] M. Enzweiler et. al., “Integrated Pedestrian Classification and Orientation Estimation”, Proc. CVPR (2010).