AI 特許紹介(26)

AI 特許を学ぶ！究める！
～動的シーンにおける教師なし単眼深度学習～

2021 年 3 月 10 日

河野特許事務所
所長　弁理士　河野英仁

「AI 特許紹介」シリーズは、注目すべき AI 特許のポイントを紹介します。熾烈な競争となっている第４次産業革命下では AI 技術がキーとなり、この AI 技術・ソリューションを特許として適切に権利化しておくことが重要であることは言うまでもありません。

AI 技術は Google, Microsoft, Amazon を始めとした IT プラットフォーマ、研究機関及び大学から毎週のように新たな手法が提案されており、また AI 技術を活用した新たなソリューションも次々とリリースされています。

本稿では米国先進 IT 企業を中心に、これらの企業から出願された AI 特許に記載された AI テクノロジー・ソリューションのポイントをわかりやすく解説致します。

1.概要

特許権者　Google
出願日　2020 年 4 月 29 日
登録日　2020 年 10 月 20 日
登録番号 US10810752
発明の名称　画像深度とエゴモーション予測ニューラルネットワークの教師なし学習

752 特許は、単眼カメラにより、画像の深度予測、及び、カメラモーション推定を行う技術に関する。

2.特許内容の説明

図 1 は、ニューラルネットワークシステム 100 のアーキテクチャを示している。

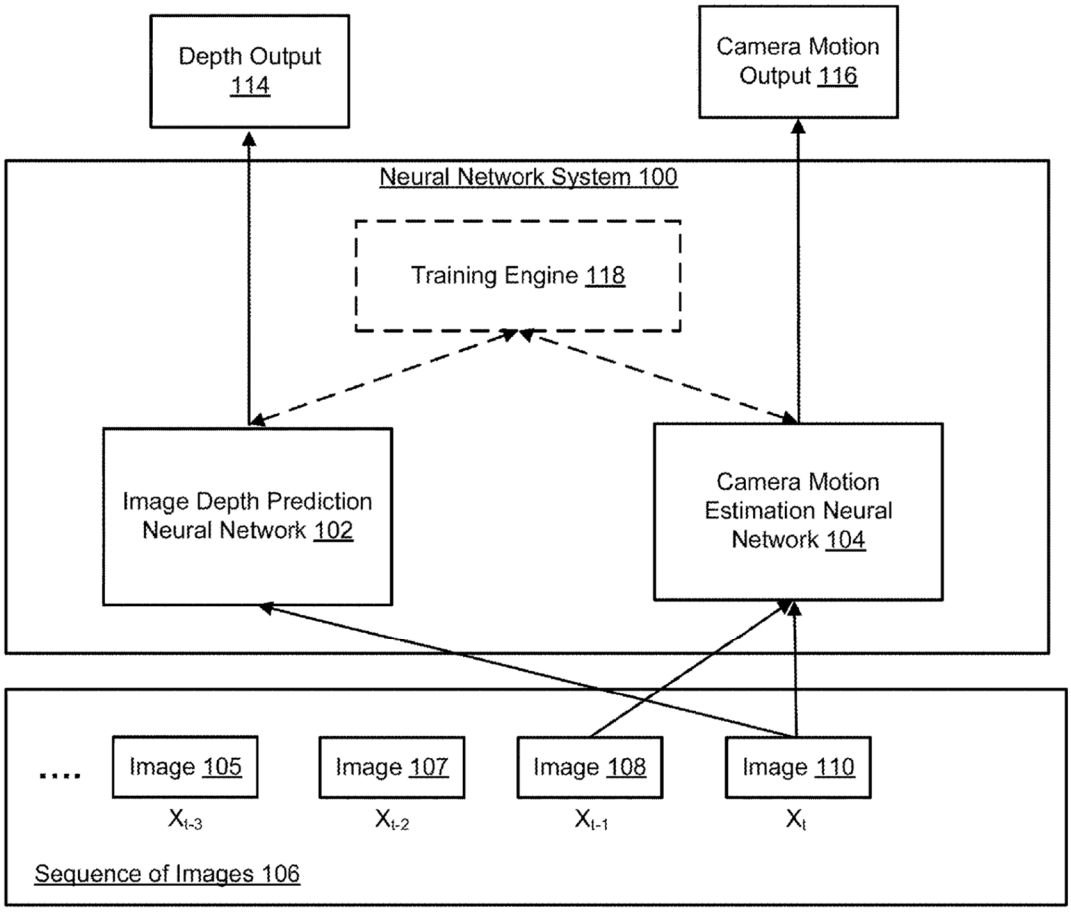


図 1

ニューラルネットワークシステム 100 は、画像 106 のシーケンスを受信し、シーケンス内の各画像を処理して画像の深度を特徴付ける深度出力を生成し、シーケンス内の画像のサブセットを処理して、サブセット内の画像間のカメラの動きを特徴付けるカメラモーション出力を生成する。

深度出力およびカメラモーション出力を生成するために、ニューラルネットワークシステム 100 は、画像深度予測ニューラルネットワーク 102 およびカメラモーション推定ニューラルネットワーク 104 を含む。深度ネットワーク 102 およびカメラモーションネットワーク 104 は、完全畳み込みニューラルネットワークである。

図 1 の例では、深度ネットワーク 102 は、画像 106 のシーケンスにおいて画像 110 を処理して、画像 100 の深度出力 114 を生成する。画像 110 は、Xt として表すことができる。ここで、t は、画像がカメラによってキャプチャされた時間である。深度出力 114 は、シーケンス 106 内の他の画像とは独立して、現在の画像 110 から生成される。

カメラモーションネットワーク 104 は、画像 106 のシーケンス内の画像のサブセッ

トを処理して、カメラモーションネットワーク 104 のモーションパラメータの現在の値に従って、サブセット内の画像間のカメラのモーションを特徴付けるカメラモーション出力を生成するように構成される。

図 1 に示すように、画像のサブセットは、２つの連続する画像 Xt-1 (108) 及び Xt (110)を含み、これらは、それぞれ、時間 t−1 及び t でカメラによって撮影される。

カメラモーションネットワーク 104 は、画像 Xt-1 および Xt を処理して、カメラの位置および向きを、画像 Xt-1 を撮影している視点から、画像 Xt を撮影している視点まで変換する変換行列であるカメラモーション出力 116（エゴモーション）を生成する。言い換えれば、カメラモーション出力 116 は、時間 t-1 から時間 t までのカメラの動き（位置および向き）を表す。

画像シーケンスの深度出力およびカメラモーション出力を効率的に生成するために、ニューラルネットワークシステム 100 は、教師なし学習技術を使用したトレーニングデータを用いて、深度ネットワーク 102 およびカメラモーションネットワーク 104 を共同でトレーニングするトレーニングエンジン 118 を含む。

各特定の画像について、トレーニングエンジン 118 は、深度ネットワーク 102 を使用して特定の画像を処理し、深度ネットワーク 102 の深度パラメータの現在値に従って、特定の画像の第 1 の深度を特徴付ける第 1 の深度推定値を生成する。

トレーニングエンジン 118 は、深度ネットワーク 102 を使用して、シーケンスの特定の画像に続く第 2 の画像を処理し、深度ネットワーク 102 の深度パラメータの現在値に従って、第 2 の画像の第 2 の深度を特徴付ける第 2 の深度推定値を生成する。

トレーニングエンジン 118 は、カメラモーションネットワーク 104 を使用して特定の画像および第２の画像を処理して、カメラの位置および向きを、特定の画像を撮影している際のその視点から、第 2 の画像を撮影している際のその視点まで変換する第 1 の変換行列を生成する。

次に、トレーニングエンジン 118 は、損失関数の勾配の推定値を逆伝播して、深度ネットワーク 102 およびカメラモーションネットワーク 104 のパラメータの現在値を共同で調整する。損失関数は、第 1 の深度推定値、第 2 の深度推定値、および第 1 の変換行列に基づいて計算される。

トレーニング中、トレーニングエンジン 118 は、上記の操作を繰り返し実行して、深さネットワーク 102 およびカメラモーションニューラルネットワーク 104 のパラメータ値を調整し、ミニバッチ確率的最適化または確率的勾配最適化方法を使用することによって損失関数を最小化する。

トレーニング後、ニューラルネットワークシステム 100 は、深度ネットワーク 102 を使用して、深度ネットワーク 102 の深度パラメータのトレーニングされた値に従って所与の入力画像の深度出力を生成し、カメラモーションネットワーク 104 を使用して、カメラモーションネットワーク 104 のモーションパラメータのトレーニングされた値に従った複数の入力画像のカメラモーション出力を生成する。

ニューラルネットワークシステム 100 は、トレーニングされた深度ネットワーク 102 およびトレーニングされたカメラモーションネットワーク 104、またはネットワークのパラメータのトレーニングされた値を外部システムに提供することができる。外部システムは、トレーニングされた深度ネットワーク 102 およびトレーニングされたカメラモーション 104 を使用して、上記の方法で一連の入力画像の深度出力およびカメラモーション出力を生成することができる。

例えば、トレーニング後、システム 100 または外部システムは、ロボットが環境と相互作用している間、深度出力およびカメラモーション出力を使用してロボットを制御することができる。深度出力とカメラモーション出力をロボットの制御ポリシーまたはプランナーへの入力として使用できるようにすることで、ロボットをより効果的に制御して、環境内の指定されたタスクを完了することができる。

3.クレーム

752 特許のクレーム 1 は以下の通りである。

1.画像深度予測ニューラルネットワークおよびカメラモーション推定ニューラルネットワークを含むニューラルネットワークを訓練するための方法であって、

シーケンス画像を含むトレーニングデータを取得し、

シーケンス画像内の特定の画像ごとに、

特定の画像の第 1 の深度を特徴付ける第 1 の深度推定値を生成すべく、画像深度予測ニューラルネットワークを使用して特定の画像を処理し、

第 2 の画像の第 2 の深度を特徴付ける第 2 の深度推定値を生成すべく、画像深度予測ニューラルネットワークを使用してシーケンス画像内の特定の画像に続く第 2 の画像を処理し、

カメラの位置と向きを特定の画像を撮影しているその視点から、第 2 の画像を撮影しているその視点に変換する第 1 の変換行列を生成すべく、カメラモーション推定ニューラルネットワークを使用して特定の画像と第 2 の画像を処理し、

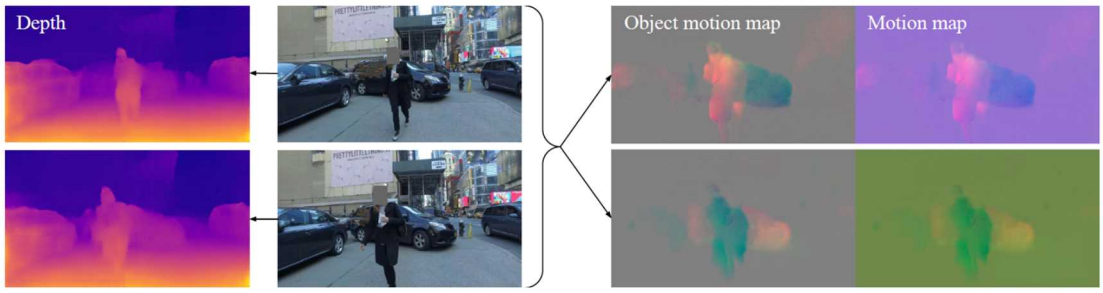
第 1 の深度推定、第 2 の深度推定、および第 1 の変換行列に基づいて、画像深度予測ニューラルネットワークおよびカメラモーション推定ニューラルネットワークのパラメータの現在の値を共同で調整するために、損失関数の勾配推定を逆伝播する。

4. ニューラルスタイル変換に関する論文

本特許に関連する論文[1]「動的シーンにおける教師なし単眼深度学習」が Anelia Angelova 氏らにより発表されている。

本論文では、単眼ビデオのみから 3D により、本特許の深度及びエゴモーション（カメラモーション）に加えて、高密度オブジェクトモーションマップをも用いて、共同で学習する方法を記載している。論文に記載された方法は、単眼ビデオ自体以外の補助信号を利用しておらず、セマンティック信号、ステレオ、またはあらゆる種類のグラウンドトゥルースは使用していない。

本論文では、深度、エゴモーション、及び、オブジェクトモーションが、単眼ビデオから同時に学習される。



上記図は、YouTube のトレーニングビデオに表示されている深度予測（フレーム毎）とモーションマップ予測（フレームのペア）を示す。全体の 3D モーションマップは、学習したカメラのモーションベクトルをオブジェクトのモーションマップに追加することによって取得される。

[1] Hanhan Li, Ariel Gordon, Hang Zhao, Vincent Casser, Anelia Angelova “Unsupervised Monocular Depth Learning in Dynamic Scenes” arXiv:1508.06576v2 [cs.CV] 2 Sep 2015

Method	Uses semantics?	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Struct2Depth [12]	Yes	0.145	1.737	7.28	0.205	0.813	0.942	0.978
Gordon [11]	Yes	0.127	1.33	**6.96**	0.195	0.830	**0.947**	**0.981**
Pilzer [43]	**No**	0.440	6.04	5.44	0.398	0.730	0.887	0.944
Ours	**No**	**0.119**	**1.29**	**6.98**	**0.190**	**0.846**	**0.952**	**0.982**

上記テーブルは、標準分割を使用する Cityscapes でトレーニングおよび評価されたモデルの、教師なし単眼深度学習アプローチのパフォーマンス比較を示す。

実験では、入力/出力に 416×128 の解像度画像を使用している。”uses semantics”列は、対応するメソッドが移動するオブジェクトの識別に役立つ事前トレーニング済みのマスクネットワークを必要とするかどうかを示す。本論文の手法ではセマンティクス情報を使用しない。

“AbsRel”, ”Sq Rel”, “RMSE”および”RMSE log”は、それぞれ平均絶対誤差、二乗誤差、二乗平均平方根誤差、および二乗平均平方根対数誤差を示す。δ /x は、グラウンドトゥルースと予測値の比率が x と 1 / x の間の分数を示す。赤の指標の場合、低いほど良く、緑の指標の場合、高いほど良いことを示している。本論文に示す手法の方が、他の手法と比較して精度が高いことが理解できる。

以上

著者紹介
河野英仁
河野特許事務所、所長弁理士。立命館大学情報システム学博士前期課程修了、米国フランクリンピアースローセンター知的財産権法修士修了、中国清華大学法学院知的財産夏季セミナー修了、MIT(マサチューセッツ工科大学)コンピュータ科学・AI 研究所 AI コース修了。
AI 特許コンサルティング、医療 AI 特許コンサルティングの他、米国・中国特許の権利化・侵害訴訟を専門としている。著書に「世界のソフトウエア特許(共著)」、「FinTech 特許入門」、「AI/IoT 特許入門 2.0」、「ブロックチェーン 3.0(共著)」がある。