様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １７ 日現在

機関番号：５３４０１
研究種目：若手研究
研究期間：2020～2022
課題番号：２０Ｋ１４１０９

研究課題名（和文）ワクワクを創出するポーズ入力型プログラミング教材の開発

研究課題名（英文）Development of pause-input programming materials that create excitement

研究代表者
小松　貴大（Komatsu, Takahiro）

福井工業高等専門学校・電子情報工学科・准教授

研究者番号：６０６３８７６６

交付決定額（研究期間全体）：（直接経費）　　3,000,000円

研究成果の概要（和文）：　カメラ画像から人の骨格（関節位置）を推定する機械学習モデルを開発し、そのモデルを用いてキーボード等を必要としないポーズ入力型のプログラミング教材を開発した。機械学習モデルはスマートフォンなどのモバイルデバイスで動作することを想定し、軽量かつ高速に処理ができるようにOctave Convolutionを用いた。開発したモデルはモデルの読み込みから推定結果の表示までに約4秒で、推定時間のみでは約0.135秒となり、プログラミング教材の入力デバイスとして使用する分には問題ない程度の遅延時間であると考えられる。

研究成果の学術的意義や社会的意義
小学校では2020年度から、中学校では2021年度からプログラミング教育が必修化されたが、実際にプログラミング言語を学ぶことではなく論理的思考を身につけることが目的である。一方で論理的思考力は幼児期から身につけることができるが、プログラミングする際のキーボード・マウス操作に慣れるための身体的な訓練が必要となり、従来のプログラミング教材ではその点が問題である。本研究では、スマートフォンなどのモバイルデバイスのカメラ機能を用いてプログラミングすることができ、ポーズにアサインされた命令を組み合わせることで論理的思考力を身につけることが可能な教材となっている。

研究成果の概要（英文）：A machine learning model was developed to estimate the human skeleton (joint positions) from camera images, and the model was used to develop programming materials that do not require a keyboard or other tools. The machine learning model was designed to run on mobile devices such as smartphones, and Octave Convolution was used for lightweight and high-speed processing. The developed model takes approximately 4 seconds from loading the model to displaying the estimation results, and the estimation time alone is approximately 0.135 seconds, which is considered to be a sufficient delay time for use as an input device for programming materials.

研究分野：知能情報学

キーワード：Skeleton Estimation　Machine Learning Model　Programming Materials

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

（1）本研究を申請した２０１９年度は小中学校でプログラミング教育が必修化（２０２０年度から小学校、２０２１年度から中学校でプログラミング教育が必修化）されることもあり、様々なプログラミング教材が小中学校で試されていた。また、本研究者が所属する機関にプログラミング教育のための出前授業等を依頼されることもあり、プログラミング教育必修化の前に様々な問題を発見することができた。その中でも、パソコン等に慣れていない小学生では特にキーボードやマウスなどの操作に戸惑うことが多々見受けられた。小学生でも低学年になる程それは顕著に見られ、見本となるプログラミング言語を入力することに時間を取られてしまい、プログラミング教育の重要となる「論理的思考力を身につける」ということが達成できないのではないかという懸念が生まれた。

（2）論理的思考力は幼児期から身につけることができるスキルであり、幼児期から多くの言葉に触れて、語彙力を増やしていくことで、「何で？」「どうして？」という問いかけに対して答えられるようになっていく。しかしながら、論理的思考力はそのようなバーバルコミュニケーションの中でのみ発達していくのではなく、顔の表情やジェスチャーなどを使ったノンバーバルコミュニケーションの中でも発達する。相手に対してどこか適当な方向を手で指し示すと、人はそちらの方向を見てしまうように、「こういう動きをすると、人はこう動く」といった関連付けなどによっても論理的思考力が発達していく。

（3）そのため、直感的にプログラミングができる教材、ツールとしてビジュアルプログラミングに注目が集まった。代表的なものとして Scratch やロボットの動きとして結果を確認するレゴ・プログラミングキットなどがある。プログラミング言語取得には繋がりにくいが、比較的容易なキーボード、マウス操作によってプログラミングができる。本研究者は幼児期を対象としたプログラミング教材の開発を目指し、２０１７年度からジェスチェー（ポーズ）入力型のプログラミング教材を開発してきた。人の体の動きを精度良く検出するためには赤外線センサなどのデバイスを外部に設置したり、身につけたりしないといけない。本研究を始める以前までは Kinect というデバイスを人の動きを検出するために用いていた。しかしながら、Kinect からのデータを処理するパソコンやプログラミングした結果を表示するためのロボットとの接続などが必要となり、幼児期の子供が誰でも取り扱える手軽なプログラミング教材とは言えなかった。

２．研究の目的

（1）本研究では幼児期からでも論理的思考力が身に付く、キーボードやマウス操作を一切必要としない、ポーズ入力型のプログラミング教材の開発する。具体的にはプログラミング教材における、ポーズの認識部分（スマートフォンなどのモバイル端末のカメラで撮影した画像から、人の関節位置を推定する機械学習モデル）を開発することが目的である。これにより、プログラミングしたい人はカメラに向かってポーズをすることで、ポーズに割り当てられた命令がロボットに送信されてロボットが動作し、プログラミング結果を確認することができる。

３．研究の方法

（1）機械学習モデルによって、カメラ画像から人のポーズを認識する方法は大きく分けて２通り考えられる。１つ目はカメラ画像が事前に準備された複数のポーズのどれに値するか判別する方法である。事前に準備されたどのポーズに該当するのか０から１の数値（確率）によって表す方法であり、機械学習における分類問題を解くモデルを用いて実現することが可能である。しかしながら、新しいポーズを登録するたびにモデルの再学習が必要となる。２つ目は人の関節位置を推定し、それぞれの関節の位置関係からポーズを判別する方法（関節位置/骨格推定モデルを開発する方法）である。人の関節位置を推定する方法は、一度モデルを構築さえすれば例え新しいポーズの登録が必要となってもモデルの再学習は必要ない（関節位置の位置関係を判別する分岐処理を追加するだけで良い）。また、本研究で開発するプログラミング教材は今後、複数人でポーズをとってプログラミングするようなペアプログラミングの観点も考慮しているため、複数人の関節位置の推定においても、開発した骨格推定モデルを応用できる点で汎用性が高い。このため、関節位置を推定する骨格推定モデルを開発することにした。

① 関節位置を推定するモデルを作成することにしたが、分類問題を解くモデルによって何のポーズかを判定するモデルにも利点がある。分類問題を解くモデルは非常に軽量であり、高価な GPU などを搭載したデスクトップパソコンでなくても、スマートフォンなどのモバイルデバイスでも動作させることができるという点である。そこで、まずは分類問題を解くモデルを用いて、関節位置を推定することができないか検証した。具体的には、分類問題に代表されるモデルの中でも最も軽量である Mobile Net V3 というモデルを転移学習させた。具体的には、出力層を人体の 13 個のキーポイントの座標位置を出力するように変更し、活性化関数はソフトマックスではなく、座標位置を表現できるように Sigmoid 関数、恒等関数、ReLU 関数に変更する。

② ①で示した分類問題を解く機械学習モデルは Convolutional Neural Network（畳み込みニューラルネットワーク）に加えて、Depthwise Separable Convolution が使用されており、空間方向とチャネル方向の畳み込み層を分解して計算することで計算コストを下げてモデルの軽量化を図っている。さらに強力な畳み込み層の計算方法として Octave Convolution を用いることにし、転移学習などさせずに一から関節位置を推定するモデルを構築することにした（図１参照）。Octave Convolution モデルは画像を低周波成分と高周波成分に分解し、それぞれの画像に対して畳み込みを行うという考え方である。これにより、Mobile Net V3 よりもさらに軽量かつ関節位置を推定するのに特化したモデルを開発する。

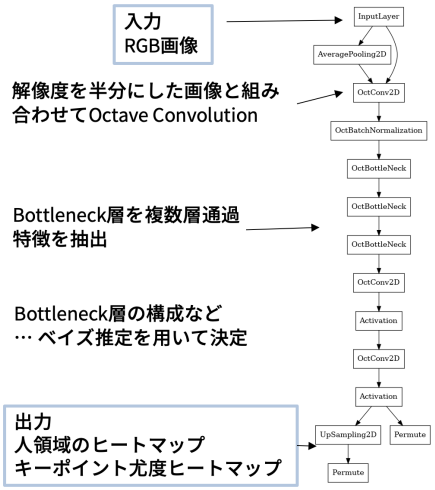

図１　開発したモデル概要

真値

部位	x	y
鼻	0.325	0.4015625
左肩	0.35625	0.421875
右肩	0.34375	0.4203125
左肘	0.35833332	0.4578125
右肘	0	0
左手首	0.35833332	0.4875
右手首	0	0
左臀部	0.35208333	0.484375
右臀部	0.34166667	0.48125
左膝	0.35833332	0.515625
右膝	0.32916668	0.5140625
左足首	0.38541666	0.5515625
右足首	0.33958334	0.5609375

図２　画像 A のキーポイントの座標位置

４．研究成果

（１）研究の方法①で示した方法で、モデルの学習を行い、ある画像 A（図２参照）に対する人体の 13 個のキーポイントの座標位置 xy（真値）に対してどのような結果が出力されたかを表１に示す。画像 A を含めたテスト画像全体として各活性化関数における真値との相対誤差の平均及び標準偏差は Sigmoid 関数では 558.04±452.38%、恒等関数では 72.96±23.37%、ReLU 関数では 73.45±52.51%となり、いずれも真値から大きく異なる xy 座標を出力していることがわかった。そもそも分類問題に代表されるモデルは、画像全体に対してその画像がどのような物体であるかを判別するモデルであり、画像から人体の複数のキーポイントの位置を推定することが困難であることがわかった。

表１　活性化関数を変更したモデルで推定した画像 A のキーポイントの座標位置

Sigmoid関数

部位	x	y
鼻	-0.2554913	-2.7011592
左肩	7.026405	2.8945267
右肩	-2.776651	0.5879204
左肘	-4.4607835	-1.6488379
右肘	1.099322	3.9825082
左手首	4.7063117	-0.14676449
右手首	2.8244193	-0.9059477
左臀部	1.8190684	4.573053
右臀部	1.6953034	2.886945
左膝	3.4676237	-0.02896352
右膝	-0.35325533	2.7600138
左足首	1.3148052	2.0522847
右足首	-0.4895978	-0.78038883

恒等関数

部位	x	y
鼻	0.1552352	0.03128973
左肩	0.45691702	0.13593632
右肩	0.15700054	0.08919221
左肘	0.33331162	0.14700222
右肘	0.08431667	0.0685454
左手首	0.12872091	0.11106607
右手首	0.03540698	0.05039927
左臀部	0.1380758	0.14059767
右臀部	0.16380334	0.11734888
左膝	0.01779315	0.03033808
右膝	0.01253778	0.04375857
左足首	0.00954884	0.04355258
右足首	0.01820496	0.02835528

ReLU関数

部位	x	y
鼻	0.48540854	0
左肩	0.6246797	0.5292142
右肩	0	0
左肘	0.35161	1.1906457
右肘	0.5921772	0.7602953
左手首	0.48825085	0.87940675
右手首	0.7524627	0.8259384
左臀部	0.32277867	0.55931824
右臀部	0.5248027	1.1355332
左膝	0.04931466	1.1527766
右膝	0.3170771	0.33121926
左足首	1.1960341	0.12702425
右足首	0.678536	0.34932896

（２）研究方法①で示したように分類問題に代表されるような機械学習モデルを転移学習させたモデルは人体の複数のキーポイントの座標位置を推定するのには適していない可能性がある。そこで、研究方法②で説明した Octave Convolution を用いて一からモデルを構築し、計算量が軽くなった分人体のキーポイントも 17 点に増やして関節位置の推定を行った。その結果を図３に示す。頭部や手首などキーポイントの特徴が他の関節位置などと比較して捉えやすい部

分は高い確率で関節位置を推定できていることがわかった。しかしながら、図３で示すような画像Ｂでは、色が一色で特徴が捉えにくくなっている足の各関節位置などは精度良く推定できていない。しかしながら、17 個のキーポイントを推定するのに要する時間は 113ms(約 7fps)となっており、かなり高速に推定できていることがわかった。この速度であればポーズ入力型プログラミング教材の入力速度としては申し分ない速度であり、モバイル端末で動画を撮影しながらプログラミングすることもできることが予想される。今回はモデルを軽量にすることに趣をおいてモデルの開発を行ったが、今後は速度―精度のトレードオフ関係を調べ、精度を上げる必要がある。精度を上げるために開発したモデルの中間層の改良、上半身/下半身に特化したモデルの開発、複数人の関節位置推定に対応するためにモデルの並列化などを検討している。

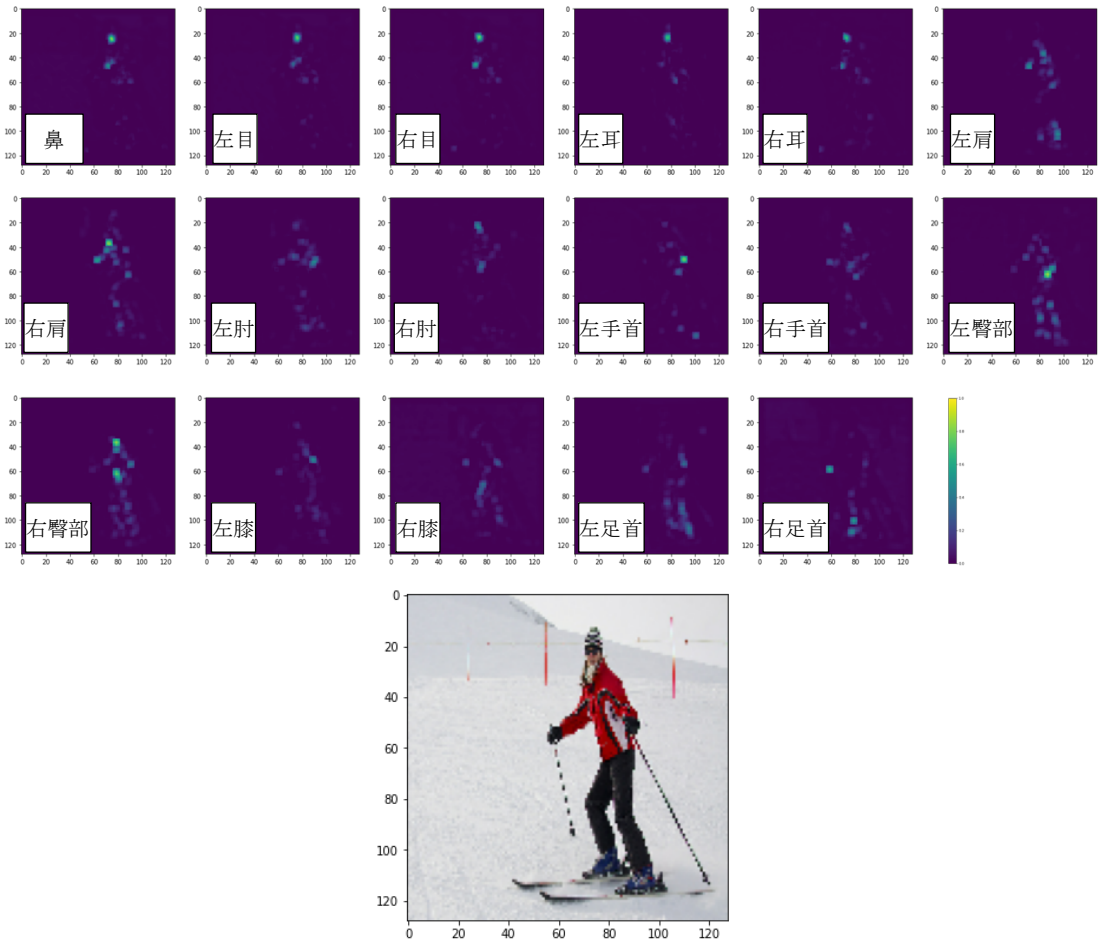

図３　Octave Convolution モデルを用いた画像 B のキーポイントのヒートマップ

<引用文献>

① 文部科学省、小学校学習指導要領（平成 29 年告示）

② 大宮明子、幼児期からの論理的思考の発達過程に関する研究、風間書房、2013

③ Y. Chen, H. Fang, B. Xu, Z. Yan, Y. Kalantidis, M. Rohrbach, S. Yan, J. Feng. Drop an Octave: Reducing Spatial Redundancy in Convolutional Neural Networks with Octave Convolution. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) ,2019

④ H. Andrew, S. Mark, C. Grace, C. Liang-Chieh, C. Bo, T. Mingxing, W. Weijun, Z. Yukun, P. Ruoming, V. Vijay, L. Q. V. , A. Hartwig, “Searching for MobileNetV3,” The IEEE International Conference on Computer Vision (ICCV), pp. 1314-1324, 2019

５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 土村 貫太，小松 貴大
２．発表標題 機械学習による自動採譜システムの開発
３．学会等名 第３ブロック専攻科研究フォーラム
４．発表年 2022年

１．発表者名 兵田 憲信，小松 貴大
２．発表標題 Human Body Pose Estimation Modelfor Smart Devices Using Deep Learning
３．学会等名 第３ブロック専攻科研究フォーラム
４．発表年 2022年

１．発表者名 土村貫太，小松貴大
２．発表標題 ディープラーニングによる自動採譜システムの開発
３．学会等名 情報処理学会第82回全国大会
４．発表年 2020年

１．発表者名 兵田憲信，小松貴大
２．発表標題 姿勢推定を用いた子ども向けプログラミング学習教材
３．学会等名 情報処理学会第82回全国大会
４．発表年 2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関