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　　元 年　　５ 月 ３１ 日現在

機関番号：１２６０８
研究種目：基盤研究(B)（一般）
研究期間：2016～2018
課題番号：１６Ｈ０２８４５

研究課題名（和文）深層学習によるマルチモーダル時系列データ認識基盤の構築

研究課題名（英文）Multimodal time-sequence data recognition platform based on deep learning

研究代表者
篠田　浩一（Shinoda, Koichi）

東京工業大学・情報理工学院・教授

研究者番号：１０３４３０９７

交付決定額（研究期間全体）：（直接経費）　　12,300,000円

研究成果の概要（和文）：本研究では深層学習を用いてマルチモーダル時系列信号を高精度に認識することを目的とした。深層学習におけるEnd-to-End学習方式、少ないデータ量でも動作する深層モデル、マルチタスク学習、耐ノイズ認識などの手法を開発した。特に、音源分離と音声認識の同時学習、音声からの認知症診断、口唇画像を用いたマルチモーダル認識、耐雑音音声認識、の4つのテーマについてこれらの技術を適用し、各々の応用において、識別性能、検出性能を改善することができた。

研究成果の学術的意義や社会的意義
深層学習はこの十年ほど画像認識や音声認識の標準的な技術となった。しかしながら、人間のもつ事前知識の活用、周囲環境の違いや話者の違いなどによる性能の劣化、学習のための大量のデータが得られない応用への適用、などの点においてまだ課題が多い。本研究では、これらの問題を解決する鍵となる、End-to-End学習、少ないデータからの効率的なモデル学習、マルチタスク学習、耐ノイズ認識の方式を提案し、一定の成果を得ることができた。これらの成果は実社会における様々な問題に対して容易に適用可能である。

研究成果の概要（英文）：This research aims to accurately recognize multi-modal time-sequence signals using deep learning. We applied various deep learning techniques such as End-to-end training, deep net which is trainable with a small amount of data, multi-task learning, and noise-robust recognition. Particularly, we improved the recognition and detection performance in simultaneous training for source separation and speech recognition, dementia detection from speech, multi-modal speech recognition using lip reading, noise-robust speech recognition.

研究分野：統計的パターン認識

キーワード：知覚情報処理　音声情報処理　動画情報処理　深層学習

様　式　C－１９、F－１９－１、Z－１９、CK－１９（共通）

１．研究開始当初の背景

音声や画像などのマルチメディアの認識において、多層のニューラルネットワーク(NN) を用いた深層学習(Deep Learning) の有効性が確認され、多く活用されている。深層学習は、計算機技術の進歩により、より大規模な NN の学習とそれによる認識が容易になったことにより、急速に進歩した。深層学習では入力信号(始端) から出力ラベル(終端) を直接学習する(End-to-End 学習)。個々の応用に固有の知識をあまり必要としないという利点がある。特にマルチメディア分野では、データ収集が比較的容易であることもあり、精力的に研究が進められている。

一方、深層学習は入力信号とラベルが直接結びつく比較的単純なタスクでは有効であるが、内部構造が複雑な多様な現象のモデル化には不向きである。これはその多様な現象を網羅するだけの大量データとそのラベルを用意できないからである。また、音声や画像は多くのデータが比較的容易に入手可能であるが、他の応用ではデータの入手が困難な場合が多い。例えば医療データはプライバシーの観点から利用が厳しく制限されている。そこでは、人間のもつ何らかの事前知識を活用したり、モデルの構造を工夫したりすることで、より少ないデータ量で安定に学習する仕組みの構築が求められている。

また、映像処理の応用では、画像信号と音声信号のマルチモーダルな時系列信号が入力となる。そこでは信号間の相関をどのようにモデル化するか、という課題がある。従来の機械学習では多くの研究があるが深層学習ではまだ十分に検討されていない。

さらに、音声認識においては周囲雑音の大きい時にその性能が劣化する。深層学習ではこのような学習データに含まれない外乱要因は考慮されていない。もちろん様々な周囲雑音下で音声を用いて学習すればある程度問題は解決するが、必要なデータ量が増加する。そこでこのような外的要因に対して頑健な学習手法が必要である。

２．研究の目的

上記の背景から、本課題では、音声・映像などの時系列信号を入力とした深層学習による認識の研究において、特に、深層学習の End-to-End 学習における事前知識の利用、データ不足に対して頑健な認識モデル、マルチモーダル認識、周囲雑音に頑健な音声認識の 4 つの課題に焦点を絞り、研究を行った。それぞれにおいて、新しい手法を提案し、既存手法よりも高い性能を得ることを目的とした。

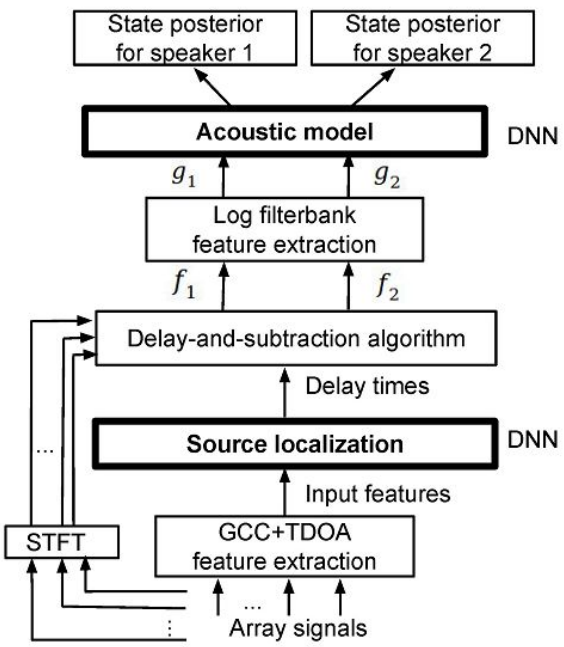
３．研究の方法

３．１ 音源分離と音声認識の End-to-End 学習

会議などで複数の話者が同時に発声している場合に高い認識率を得ることを目的とする。従来、まず音源分離の技術を適用して入力音における個々の音声を分離し、次に分離された個々の音声を認識するという二段階の方法が採られてきた。しかしながら、音源分離と音声認識とは密接に関係しており、その両者を同時に学習する End-to-End 学習の枠組みを用いることにより、音声認識性能が向上することが期待できる。ここでは、一般化相互相関(Generalized cross-correlation, GCC)と到着時刻差(Time difference of arrival, TDOA)の 2 つの特徴を入力とした音源分離用の深層ネットワーク(Deep neural network, DNN)と、音声認識のための DNN を直列に接続して１つの DNN を構築し、それに対して End-to-End 学習を行う。(下図)

３．２ 認知症診断のためのデータ不足に対して頑健な認識モデル

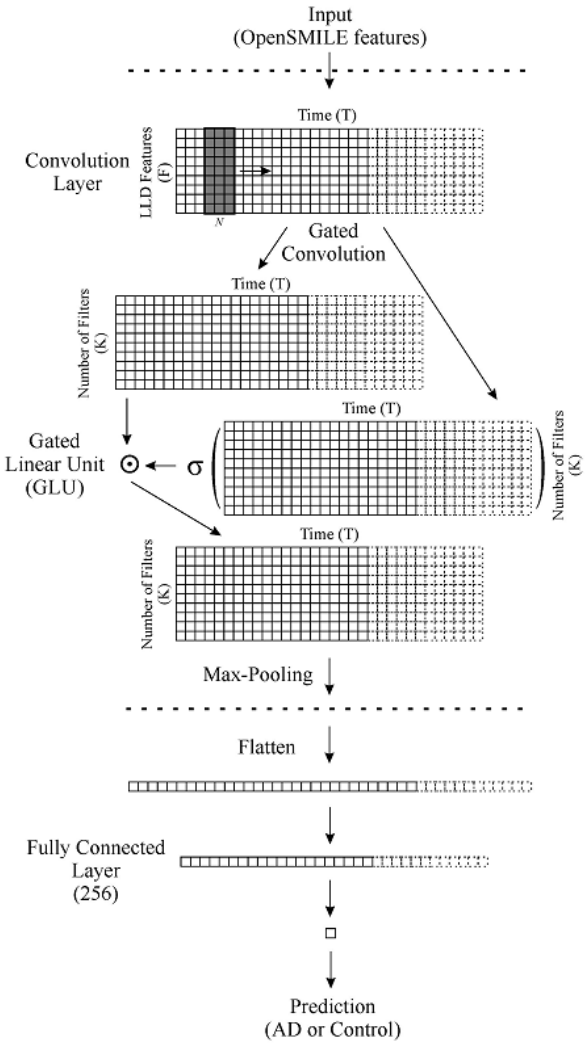
認知症は従来 MRI などの機器を用いた測定や医師による問診が用いられてきたが、ともにコストが高い。そこで、認知症患者を音声における音響的特徴のみを用いて検出することを目的として研究を行った。医療データなので多くのデータ量が得られず、少ないデータ量で高い性能を得ることが課題である。ここでは、ゲート付き畳み込みニューラルネットワーク(Gated convolutional neural network, GCNN)を用いることによりその解決を図った。GCNN は通常の CNN の出力にゲートを設けてその調節を行うことにより、より効率的に音声事象をモデル化する。(次ページ上図)



３．３ 音声と口唇深度画像を利用したマルチモーダル音声認識

雑音下では一般に音声認識の性能が劣化する。そこでは口唇(くちびる)の動きを認識する読唇(リップリーディング)を併用するこ

とにより認識性能の劣化が軽減できることが知られている。ここでは、口唇の深度情報(顔画像の垂直方向の動き)を深度カメラで観測し、その情報も入力に加える。正面画像からでは観測が困難な、口の開閉度合いや舌の位置を推定することで、より高い認識性能を得られると期待できる。音声、口唇画像、口唇深度情報の間の相関をモデル化するため、この 3 モードを入出力とした、深層自己復号化器(Deep autoencoder)を構築し、それを音声認識器と組み合わせた。(下図)



３．４ マルチタスク自己符号化器を用いた耐雑音音声認識

雑音下では音声認識の性能が著しく劣化する。その劣化を軽減するために、近年、深層学習を用いた雑音除去自己復号化器(Deep denoising autoencoder, DDAE)が提案され、効果があることが確認されている。DDAE では、雑音下の音声を入力とし、そこから雑音を取り除いた音声を出力とする自己復号化器を学習する。ここでは、雑音と音声の特徴の違いをより明示的に扱うために、同じ入力に対し、音声を取り除いた雑音を出力とする自己復号化器を構築し、それを DDAE と同時に学習するマルチタスク学習の枠組みを提案した。(次ページ図)

４．研究成果

４．１ 音声分離と音声認識の End-to-End 学習

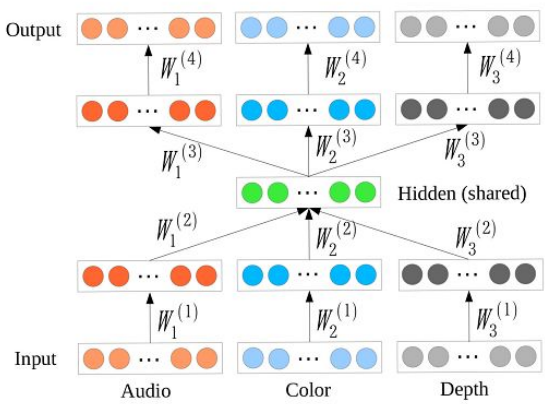
従来の深層学習を用いない手法や、End-to-End 学習をしない手法(2 つの DNN を別々に学習して単に接続する場合)に比べ優位に高い認識性能を得ることができた。また、音源方法が未知の場合においても十分に性能が高くなることを確認した。国際会議 APSIPA2017 にてこの成果を発表した。

４．２ 認知症診断のためのデータ不足に対して頑健な認識モデル

従来の深層学習を用いない手法に比べ顕著な性能向上を得た。また従来の CNN をそのまま用いた場合に比べても性能が向上し、GCNN の効果が確認できた。国際会議 Interspeech2018 にてこの成果を発表した。また、その後、慶応大学医学部との共同研究において、その所有するデータを用いて評価を行ったところ、8 割以上の性能を獲得し、その成果を国際論文誌に投稿予定である。

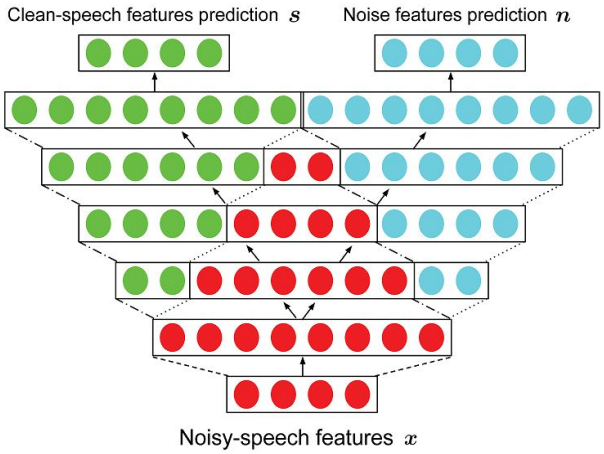
４．３ 音声と口唇深度画像を利用したマルチモーダル音声認識

従来の深層学習を用いない方法に比べ顕著な性能改善を得ることができた。また、口唇の深度情報が音声認識性能の向上に寄与することも確認できた。しかしながら、画像との併用では更なる性能向上が確認できなかった。深度カメラの解像度がこの用途向けにはまだ不十分であることが原因であると考えられる。国際会議 APSIPA2017 にてこの成果を発表した。



４．４ マルチタスク自己符号化器を用いた耐雑音音声認識

雑音を考慮しない手法や DDAE のみに比べ、顕著に耐雑音性能が向上した。今後、音声認識との End-to-End 学習を行うことにより更に性能が向上する見込みがある。国際会議 Interspeech2018 にてこの成果を発表した。



５．主な発表論文等

〔雑誌論文〕（計 3 件）

① Nakamasa Inoue, Koichi Shinoda, “[Invited Paper] Semantic Indexing for Large-Scale Video Retrieval”, ITE Transactions on Media Technology and Applications, 査読有, Vol.4, pp.209-217, 2016.
DOI: 10.3169/mta.4.209

② Ryan Price, Ken-ichi Iso and Koichi Shinoda, “Wise teachers train better DNN acoustic models”, EURASIP Journal on Audio, Speech, and Music Processing, 査読有, pp.1-19, 2016.
DOI: 10.1186/s13636-016-0088-7

③ 篠田 浩一, “音声言語処理における深層学習：総説”, 日本音響学会誌, 査読有, Vol.73, pp. 25-30, 2017.

〔学会発表〕（計 37 件）

① Tran Hai Dang, Nakamasa Inoue, Koichi Shinoda, “Concept Elimination for Zero-Shot Event Detection”, The 22nd Symposium on Sensing via Image Information(SSII), 2016 年 06 月 08 日～ 2016 年 06 月 10 日, パシフィコ横浜アネックス（神奈川県横浜市）.

② Koichi Shinoda, “Deep Learning for Speech, Image, and Video”, International Conference on Computer, Control, Informatics, and Its Applications (IC3INA) (招待講演) (国際学会), 2016 年 10 月 03 日, Indonesia Convention Exhibition (ICE)(Tangerang, Indonesia).

③ 篠田 浩一, “東工大 TSUBAME の活用事例：マルチメディア認識のための深層学習”, GTC Japan 2016 (招待講演), 2016 年 10 月 05 日, ヒルトン東京お台場(東京都港区).

④ Nakamasa Inoue, Koichi Shinoda, “Adaptation of Word Vectors using Tree Structure for Visual Semantics”, ACM Multimedia 2016 (国際学会), 2016 年 10 月 15 日～ 2016 年 10 月 19 日, Theater Tuschinski,(オランダ　アムステルダム).

⑤ Nakamasa Inoue, Ryosuke Yamamoto, Na Rong, Koichi Shinoda, “TokyoTech at TRECVID 2016”, NIST TRECVID workshop (招待講演) (国際学会), 2016 年 11 月 14 日～ 2016 年 11 月 16 日, NIST, Gaithersburg (MA, USA).

⑥ Koichi Shinoda, “Video Semantic Indexing and Localization”, 5th Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan (招待講演) (国際学会), 2016年11月28日～ 2016年12月02日, Hilton Hawaiian Village (Honolulu, USA).

⑦ Conggui Liu, Nakamasa Inoue, Koichi Shinoda, “Speaker Separation in Multi-Channel Environment Using Deep Learning”, 情報処理学会音声言語情報処理研究会, 2017 年 02 月 17 日～ 2017 年 02 月 18 日, 琴平グランドホテル桜の抄(香川県琴平町).

⑧ Koichi Shinoda, “Video Information Retrieval”, The 2017 IEEE SPS Summer School on Visual Image Search and Visual Analytics (VISVA2017) (招待講演), Jul. 5, 2017.

⑨ 安井勇樹, 岩野 公司, 井上 中順, 篠田 浩一, “口唇の深度画像を用いたディープオートエンコーダによるマルチモーダル音声認識”, 情報処理学会研究報告 SLP, Jul. 27, 2017.

⑩ Conggui Liu, Nakamasa Inoue, Koichi Shinoda, “Joint training of speaker separation and speech recognition based on deep learning”, ASJ 2017 Autumn Meeting, pp. 63-64, Sep. 25, 2017

⑪ 安井 勇樹, 岩野 公司, 井上 中順, 篠田 浩一, “口唇深度画像を利用したディープオートエンコーダに基づくマルチモーダル音声認識”, 日本音響学会 2017 年秋季研究発表会講演論文集, pp. 117-118, Sep. 25, 2017.

⑫ 篠田 浩一, “深層学習の音声認識への応用”, 情報処理学会連続セミナー2017 第 4 回デ

ィープラーニングの活用と基盤(招待講演), Oct. 17, 2017

⑬ Mengxi Lin, Nakamasa Inoue, Koichi Shinoda, “CTC Network with Statistical Language Modeling for Action Sequence Recognition in Videos”, ACM Multimedia Thematic Workshop(国際学会), pp. 393-401, Oct. 23, 2017.

⑭ Nakamasa Inoue, Ryosuke Yamamoto , Na Rong , Satoshi Kanai , Junsuke Masada , Chihiro Shiraishi , Shi-wook Lee , Koichi Shinoda, “TokyoTech-AIST at TRECVID 2017: Multimedia Event Detection Using Deep CNNs and Zero-Shot Classifiers”, TRECVID Workshop(国際学会), pp. 1-6, Nov. 13, 2017.

⑮ Yuki Yasui, Nakamasa Inoue, Koji Iwano, Koichi Shinoda, “Multimodal Speech Recognition Using Mouth Images from Depth Camera”, APSIPA（国際学会）, pp. 1233-1236, Dec. 11, 2017.

⑯ Conggui Liu, Nakamasa Inoue, Koichi Shinoda, “A Unified Network for Multi-Speaker Speech Recognition with Multi-Channel Recordings”, APSIPA（国際学会）, pp. 1304-1307, Dec. 11, 2017.

⑰ 篠田 浩一, “高速かつ省資源な深層学習の実現に向けて”, JST・NSF 国際連携シンポジウム(招待講演), Dec. 20, 2017.

⑱ Lin Mengxi , Nakamasa Inoue, Koichi Shinoda, “Action Sequence Recognition in Videos by Combining a CTC Network with a Statistical Language Model”, Technical Reports of IEICE PRMU, vol. 117, no. 362, pp. 1-6, Dec. 16, 2017.

⑲ 生田目 敬弘, 亀岡 弘和, 篠田 浩一, “全層ゲート付き 2 次元畳み込みネットワークによる多重音信号の音高認識”, 研究報告音声言語情報処理（SLP）,. vol. 120, no. 12, pp. 1-7, Feb. 13, 2018

⑳ Haoyi Zhang, Conggui Liu, Nakamasa Inoue, Koichi Shinoda, “Multi-Task Autoencoder for Noise-Robust Speech Recognition”, ICASSP(国際学会),. pp. 5599-5603, Apr. 15, 2018

㉑ 篠田 浩一, “深層学習のための Co-Design”, 電子情報通信学会技術研究報告 SP/PRMU（招待講演）, vol. 118, no. 112, pp. 65, Jun. 29, 2018.

㉒ 金井怜, 井上 中順, 李時旭, 篠田 浩一, “単語分散表現を用いた動画からのイベント検出”, 第 21 回 画像の認識・理解シンポジウム (MIRU), Aug. 7, 2018.

㉓ Yan Long , Nakamasa Inoue, Koichi Shinoda, Yoichi Yatsu, Ryosuke Itoh, Nobuyuki Kawai, “Astronomical Image Subtraction for Transient Detection Using CNN”, 第 21 回 画像の認識・理解シンポジウム, Aug. 7, 2018.

㉔ Tifani Warnita, Nakamasa Inoue, Koichi Shinoda, “Alzheimer's Disease Prediction Using Audio Gated Convolutional Neural Network”, ASJ 2018 Autumn Meeting, pp. 1223-1224, Aug. 29, 2018.

㉕ Jiacen Zhang, Nakamasa Inoue, Koichi Shinoda, “Generative Adversarial Network Based i-Vector Transformation for Short Utterance Speaker Verification”, ASJ 2018 Autumn Meeting, pp. 1345-1346, Aug. 29, 2018.

㉖ Thao Minh Le, Nakamasa Inoue, Koichi Shinoda, “A Fine-to-Coarse Convolutional Neural Network for 3D Human Action Recognition”, British Machine Vision Conference (国際会議), Sep. 3, 2018.

㉗ Tifani Warnita, Nakamasa Inoue, Koichi Shinoda, “Detecting Alzheimer's Disease Using Gated Convolutional Neural Network from Audio Data”, Interspeech (国際会議), pp. 1706-1710, Sep. 4, 2018.

㉘ Jiacen Zhang, Nakamasa Inoue, Koichi Shinoda, “I-vector Transformation Using Conditional Generative Adversarial Networks for Short Utterance Speaker Verification”, Interspeech（国際学会）, pp. 3613-3617, Sep. 4, 2018.

㉙ Nakamasa Inoue, Koichi Shinoda, “Few-Shot Adaptation for Multimedia Semantic Indexing”, ACM Multimedia（国際学会）, pp. 1110-1118, Oct. 23, 2018.

㉚ Nakamasa Inoue, Chihiro Shiraishi, Aleksandr Drozd, Koichi Shinoda, Shi-wook Lee, Alex Chichung Kot, “VANT at TRECVID 2018”, TRECVID workshop（国際学会）, Nov. 13, 2018.

㉛ Thao Minh Le, Nakamasa Inoue, Koichi Shinoda, “Skeleton-based Human Action Recognition with Fine-to-Coarse Convolutional Neural Network”, Technical Reports of IEICE PRMU, vol. 118, no. 362, pp. 61-64, Dec. 13, 2018.

㉜ K. A. Lee, H. Yamamoto, K. Okabe, Q. Wang, L. Guo, T. Koshinaka, J. Zhang, K. Shinoda, “The NEC-TT Speaker Verification System for SRE’18”, NIST 2018 Speaker Recognition Evaluation（国際学会）, Dec. 2018.

㉝ 篠田 浩一, “情報理工学の現状と将来”, 第 40 回蔵前科学技術セミナー（招待講演）, Feb. 23, 2019.

㉞ Dongxiao Wang, Hirokazu Kameoka, Koichi Shinoda, “A robust algorithm of phase recovery for speech enhancement”, 電子情報通信学会 音声研究会 SP, vol. 118, no. 497, pp. 137-142, Mar. 14, 2019.

㉟ Dongxiao Wang, Hirokazu Kameoka, Koichi Shinoda, “Improving the robustness of multiple input spectrogram inversion”, **日本音響学会 2019 年春季研究発表会講演論文集**, pp. 1307-1308, Mar. 7, 2019.
㊱ Raden Mu’az Mun’im, Nakamasa Inoue, Koichi Shinoda, “SEQUENCE-LEVEL KNOWLEDGE DISTILLATION FOR MODEL COMPRESSION OF ATTENTION-BASED SEQUENCE-TO-SEQUENCE SPEECH RECOGNTION”, ICASSP 2019**(国際会議)**, 2019.

〔図書〕（計１件）
① 篠田 浩一, **講談社**, **音声認識(機械学習プロフェッショナルシリーズ)**, 2017, 165.

〔産業財産権〕
なし

〔その他〕
ホームページ：http://www.ks.cs.titech.ac.jp/japanese/index.html

６．研究組織

(1)研究分担者
研究分担者氏名：井上　中順
ローマ字氏名：INOUE NAKAMASA
所属研究機関名：東京工業大学
部局名：情報理工学院
職名：助教
研究者番号（8 桁）：10733397

研究分担者氏名：岩野　公司
ローマ字氏名：IWANO KOJI
所属研究機関名：東京都市大学
部局名：メディア情報学部
職名：教授
研究者番号（8 桁）：90323823